Morphing Isolated Quasi-Harmonic Acoustic Musical Instrument Sounds Guided by Perceptually Motivated Features

Marcelo Caetano

Submitted in partial fulfillment of the requirements for the degree of Docteur de l'Université Paris VI - Pierre et Marie Curie (UPMC) with major in signal processing from the École Doctorale Informatique, Télécommunications et Électronique (EDITE)

Reviewers	Xavier Serra - MTG, Universitat Pompeu Fabra
	Laurent Pottier - Université Jean-Monet Saint Étienne
Examiners	Philippe Depalle - CIRMMT, McGill University
	Bertrand David - Télécom ParisTech
	Jean-Dominique Polack - UPMC
	Axel Röbel - IRCAM
	Xavier Rodet - IRCAM

This work was supervised by Xavier Rodet and conducted at the Institut de Recherche et Coordination Acoustique/Musique (Ircam)

Ircam - CNRS-UMR9912-STMS
Sound Analysis/Synthesis Team
1, place Igor Stravinsky
75004 Paris, FRANCE

November 29, 2011

Final version

Abstract

This thesis is the result of an investigation of musical instrument sound morphing guided by perceptually motivated features. Sound morphing encompasses a set of models and techniques whose goal is to obtain gradual transformations between sounds. Sound morphing has been used in music compositions, in sound synthesizers, and even in psychoacoustic experiments, notably to study timbre spaces. When morphing musical instrument sounds, the focus is on timbral features and how to control them in a perceptually relevant way. The aim of this work is to morph musical instrument sounds across timbre dimensions to create the auditory illusion of hybrid musical instruments.

Sound morphing is an inherently perceptual problem, so the morphing transformation is usually required to produce perceptually intermediate results. A very challenging aspect of this problem is to control the transformation with a single parameter, called morphing or interpolation factor. Ideally, the morphing factor should control perceptually related features of the transformation. In this thesis, the ultimate goal is to obtain a perceptually linear morph when the morphing factor varies linearly. Most morphing techniques proposed in the literature use the interpolation principle, which consists in interpolating the parameters of the model used to represent the sounds regardless of the perceptual impact. The basic idea behind the interpolation principle is that we should obtain a seamsless transition between sounds by interpolating between these parameters. However, most morphing techniques found in the literature tend to produce nonlinear transitions, so this thesis aimed at developing a method to obtain more perceptually linear morphs with the aid of perceptually motivated features.

There seems to be no consensus in the literature about what sound morphing is, or equivalently, what transformations can be considered morphing. This thesis approaches the question from a theoretical and technical perspectives, discussing the requirements of morphing and the difference between morphing and other hybridization processes. This work reviews thoroughly the transformations usually referred to as morphing in the literature, and proposes a system to classify the different types of morph. In this work, the investigation revolves around the morphing transformation that produces several instances of the departure sound that become progressively closer to a target sound, called cyclostationary morph. The cyclostationary morph figures prominently as a very challenging morphing transformation because it requires accurate control of temporal and spectral characteristics of the morph to obtain a perceptually linear result.

This thesis developed a source-filter (SF) model for musical instrument sounds that gives independent control of the spectral envelope and frequency of the partials to perform the transformations. The sounds are decomposed into a sinusoidal and a residual parts, which are represented independently with the SF model. The sinusoidal component comprises a time-varying spectral envelope model (filter) and the frequencies of the partials (source), while the residual component is modeled as white noise (source) shaped by a time-varying spectral envelope (filter). The SF representation was validated with a listening test. Participants were presented the original and SF representation of sounds and asked to assess their perceptual similarity.

This thesis formalized the concept of morphing sounds, proposed a general algorithm, and a

framework to objectively evaluate morphing using three criteria, namely correspondence, intermediateness, and smoothness. Most works about morphing in the literature skip the evaluation of the results, usually considered too difficult and subjective. The evaluation is considered a crucial part of this research work, responsible for the validation of the results. In this work, the evaluation consisted in verifying the linearity of the morph using objective measures and subjective tests. This thesis proposes to use perceptually related features to objectively evaluate the linearity of the morph. The features used are acoustic correlates of salient timbre dimensions derived from perceptual studies. The focus is on spectral envelope morphing, an important part of the SF model that is related to the perception of the timbral subset of attributes called sound color. A major part of the evaluation lay in the comparison of the linearity across several spectral envelope morphing techniques found in the literature together with others prosed in the scope of this thesis. The SF representation of musical instrument sounds was compared to the popular sinusoidal model in a listening test on the perceptual linearity of the cyclostationary morph.

Keywords: *sound morphing, musical instrument, source-filter model, timbre, sound features, acoustic correlates of timbre dimensions*

Résumé

Cette thèse concerne le “morphing” de sons d’instruments de musique guidé par descripteurs audio fondés sur la perception. Le “morphing” sonore inclut un ensemble de modèles et techniques qui ont pour but d’obtenir des transformations graduelles entre sons. Le “morphing” sonore est déjà utilisé en compositions musicales, en synthèse sonore et en expériences psychoacoustiques, notamment pour étudier des espaces de timbre. Pour le “morphing” entre sons d’instruments de musique, l’intérêt porte sur les aspects du timbre et comment les contrôler perceptivement. La spécificité de cette thèse a été de combiner des sons d’instruments de musique à travers différentes dimensions du timbre pour créer l’illusion auditive d’instruments musicaux hybrides.

Le “morphing” sonore est une thématique inhérente à la perception utilisé pour créer de sons perceptivement intermédiaires. Un aspect difficile est contrôler le “morphing” avec un seul paramètre, le facteur de “morphing”, qui dans l’idéal doit contrôler des aspects liés à la perception. Dans cette thèse, le but est d’obtenir une transformation linéaire sur le plan perceptif en variant le facteur de “morphing” linéairement. Les techniques de “morphing” dans la litterature appliquent le principe de l’interpolation, qui consiste à interpoler les paramèteres du modèle utilisé pour représenter les sons sans tenir en compte l’impact perceptif. Le principe de l’interpolation suppose qu’on obtient des transformations graduelles entre sons en interpolant les paramètres de ses représentations. Pourtant, la plupart de techniques de “morphing” décrites dans la litterature présente une tendence à produire des transformations non linéaires sur le plan perceptif. Cette étude a pour but l’obtention de transformations plus linéaires en utilisant les descripteurs audio fondés sur la perception.

Il n’y pas de consensus dans la litterature sur une définition du terme “morphing” ou quelles transformations sonores y correspondent. Cette thèse s’intéresse à cette question d’un point de vue théorique et pratique, et décrit les conditions nécessaires pour qu’une transformation soit considérée comme “morphing.” Ce travail présente une rèvision approfondie du “morphing” dans la litterature et propose un système de classification. Ce travail étudie une transformation en particulier, appelée “morphing cyclostationaire” dont le but est d’avoir plusieurs versions du son de départ qui s’approchent petit à petit du son d’arrivée. Le “morphing cyclostationaire” figure parmi les transformations très difficiles parce qu’il faut contrôler simultanément des aspects temporels et spectraux pour obtenir une transformation linéaire.

Cette thèse a developé une implémentation du modèle source-filtre (SF) pour les sons d’instruments de musique qui permet de transformer indépendamment l’enveloppe spectrale et la fréquence des partiels. Les sons sont décomposés en une partie sinusoïdale et une autre partie bruitée, representées indépendamment avec le modèle SF. La composante sinusoïdale contient un modèle d’enveloppe spectrale variable dans le temps (le filtre) et les fréquences des partiels (la source). La composante bruitée est modélisée comme du bruit blanc filtré par une enveloppe spectrale variable dans le temps. La répresentation SF a été validée lors d’un test d’écoute. Les participants ont jugé la similarité perceptuelle entre l’enregistrement original et sa répresentation SF.

Cette thèse formalise le concept de “morphing sonore” et propose un algorithme avec un cadre d’évaluation objective qui utilise trois critères, correspondence, intermédiarité et “smoothness.” La

plupart des travaux sur le “morphing” sonore dans la litterature ne présente pas d'évaluation des résultats, considerée comme une tâche trop complexe. L'évaluation est une partie essentielle de ce travail, nécessaire à la validation des résultats. L'évaluation consiste à verifier si le “morphing cyclostationaire” est linéaire en utilisant des mesures objectives et un test d'écoute. Ce travail propose l'utilisation des descripteurs sonores fondés sur la perception pour évaluer la transformation. Ces descripteurs sont des corrélats acoustiques des dimensions d'espaces de timbre dérivés d'études de perception. Ce travail met en avance l'enveloppe spectrale, une partie importante du modèle SF qui correspond à la qualité sonore appelée “couleur.” L'évaluation s'est concentrée sur l'enveloppe spectrale et de son effect dans la transformation. La répresentation SF a été comparée avec le modèle sinusoïdal dans un test découte qui a eu pour but de décider quelles méthode donne des transformations plus linéaires sur le plan perceptif.

Mots-clés : *morphing sonore, instrument de musique, modèle source-filtre, timbre, descripteurs sonores, corrélats acoustiques des dimensions du timbre*

To Katerina Pejoska, my inspiration, who gave me strength when I needed and believed in me

Acknowledgements

First of all I'd like to thank Xavier Rodet, my advisor, for giving me the opportunity to do my PhD at Ircam and go through all the life-changing experiences involved. Today, I am most definitely not the same person who arrived here almost 5 years ago now.

I would like to thank all the members of my committee, as well as the people who accepted to be members of my committee but could not make it, Aki Härmä, Richard Kronland-Martinet, and Mark Sandler;

I would also like to thank the following people, without whom this work would certainly have had a different outcome:

All the members of the Analysis/Synthesis team, past and present, who I interacted with, Axel Röbel, Christophe Veaux, Gilles Degottex, Pierre Lanchantin, Marco Liuni, Nicolas Obin, Christophe Charbuillet, Geoffroy Peeters, Maël Derio, Frédéric Cornu, Thomas Hélie, Mathieu Ramona, Fernando Villavicencio, Chunghsin Yeh, Lise Régnier, Alessandro Saccoia, Mathieu Lagrange, Paul Ladyman, Juan José Burred, Carmine Cella, Diemo Schwartz, Niels Bogaards, Snorre Farner, Henrik Hahn, Grégory Beller, Michael Sweeton, Leigh Smith, Joel Ross, Stéphane Hubert, Helori Lanos, Niels Bogaards, Snorre Farner, Julien Bloit, Damien Tardieu, Jean-Philippe Lambert, and Rémi Mignot;

The members of other research groups who I interacted with, Markus Noistering, Gérard Assayag, Patrik Susini, Carlos Agon, Norbert Schnell, Olivier Warusfel, René Caussé, Kurijn Buys, Baptiste Caramiaux, Sarah Fili, Aurélie Frère, Tommaso Bianco, Bruno Zamborlin, Fivos Maniatakos, Pauline Eveno, Arshia Cont, Arnaud Dessein, Grégoire Carpentier, Vincent Freour, Marie Tahon, and Jean Bresson;

Xavier Rodet, whose suggestions improved the quality of the text, and all the people who kindly reviewed my draft chapters. Silvie Benoit (Résumé); Nicolas Obin (Résumé and Abstract); Mathieu Ramona, Christophe Charbuillet, and Henrik Hahn (Introduction); Marco Liuni (Morphing); Jean Bresson (Morphing Sounds); Christophe Veaux (State of the Art); Mathieu Lagrange (The Source-Filter Model); Carmine Cella (Temporal Evolution); Alessandro Saccoia (Temporal Envelope Estimation); Gilles Degottex (Overview of the Method); Nils Peters (Temporal Alignment); Juan José Burred (Spectral Envelope Morphing);

Michael McNabb, for kindly sending me his piece "Dreamsong" and letting me use it in my presentations;

Stephen McAdams, for the insightful conversations, the inspiring interaction, and especially for the support;

Jean-Claude Risset, for the brief but exciting encounter, which inspired me a lot;

Naotoshi Osaka, for the fruitful exchange of ideas and the productive collaboration that became a friendship;

Victor Lazzarini and Joe Timoney, for the invitation to give a talk at the National University of Ireland and the interest in my work;

Katerina Pejoska, for the invitation to give a guest lecture at the Utrecht School of Music and Technology, for having faith in me, and for her immense support;

Aki Härmä, for arranging a guest talk at Philips Research;

Takuro Mizuta Lippit, for inviting me as a guest speaker at STEIM;

Arshia Cont, for his support whenever I needed it;

Anssi Klapuri, Mark Sandler, James W. Beauchamp, Marcelo Wanderley, Sílvio Ferraz, Eduardo Reck Miranda, and Philippe Depalle, for the support and interest in my work;

Malcolm Slaney, for the brief but stimulating interaction and for the interest in my work;

Axel Röbel, for his keen mind and sharp observations;

Geoffroy Peeters, for his helpful comments and suggestions;

Jônatas Manzolli, for his unconditional support and friendship from the very beginning of my trajectory;

Fernando José Von Zuben, for everything he taught me and for his never-ending willingness to help;

Joel Ross, for the all the scripts he wrote for the listening tests;

Sylvain Le Groux, Esteban Maestre, Tae Hong Park, Marcel Wierckx, Jason Brand, and all the people who made all the conferences I attended during my PhD a lot of fun rather than plain professional opportunities;

Sylvie Benoit, for being there whenever I needed;

Gilles Degottex, for being not only a cool roommate but especially a great friend, and for eventually revealing his secret fondue recipe ;)

Frédéric Cornu, for his support and friendship, and all the conversations;

Bruno Belfiore, for his friendship and sense of humor, and for all the "real" lunches we had together;

Sophie Besnard, for all the splendid dinner parties;

Stephen Barrass, for the great conversations;

Leigh Smith, for his friendship, all the awesome conversations, and for sharing his infinite factual knowledge;

Deborah Lopatin, for her friendship, the countless coffees and the beans;

Mika Kuuskankare, for his friendship and for bearing with my rusted Finnish linguistic skills;

Sara Adhitya, Helori Lanos, Niels Bogaards, Snorre Farner, Thomas Goepfer, Delphine Oster, Marina Hinkens, and Vasso Zachari, for the friendship;

Hiroko Terasawa, for her precious help in the beginning, when I most needed it;

Sandra El-Fakhouri, for all the books, articles, and music she provided;

Didier Perrini, Silvie Benoit et Martine Grospiron for their administrative help whenever I needed it;

Jean Bresson and Hélène Caujolle, for giving me shelter near the end;

All the anonymous people who gave me a little bit of their time and did my listening tests (helping me finish), I am grateful for your interest and your help;

Last but not least, I'd like to thank all the friends that I made during the last 5 years who did not make a separate entry above but who will nevertheless always be a part of me, and my family who has always been by my side.

Preface

This work is about sounds, not sound waves or waveforms. Sound waves are the pattern of pressure waves that travels in a compressible physical medium, and, as such, sound waves are the subject of acoustics, a brach of physics that studies mechanical waves and phenomena associated with them. When the sound wave reaches our ears, we hear sounds. Sounds are cognitive representations of sound waves resulting from the subjective experience of sound perception. Psychoacoustics is the branch of science that studies the psychological and physiological responses associated with sound. But even though this work makes extensive use of psychoacoustics, it is not about psychoacoustics. Rather, it is a work on signal processing. Signal processing deals with operations or analysis of signals, such as images, sensor data, biological data, and waveforms. Waveforms are mere graphic representations of sound waves, and even though they are extremely useful in the present work, they are not the central object of study here because they do not represent all possible aspects of sound perception. The aim of this work is to transform waveforms in order to attain a desired perceptual effect. The question to be answered is then "What operations should we perform on what representation of the sound wave in order to achieve a desired perceptual result?"

When I arrived at IRCAM, Xavier Rodet, my advisor, spent a couple of weeks with me asking around, looking for an interesting problem for me to work on. We asked composers and other researchers. Finally, the problem he came up with seemed to me simple to understand but very hard to solve. The basic idea was to perform perceptually linear sound transformations to navigate through a user-defined space. The user would input a sequence of sounds that defines a direction of transformation in an abstract space. My task was to find the direction of transformation defined by the sounds input by the user as a straight line that represents perceptually salient features of the sequence of sounds and allow the user to perform perceptually linear transformations along the axis of transformation using a single parameter that controls the displacement. The idea is illustrated below in figure 1, where we see the sequence of sounds input by the user as "x" in an abstract space defined by perceptually related features of the sounds. Given the sequence of sounds (notice that the sounds are ordered) and a predefined set of features (that define the dimensions of the space), the task is to find the dashed straight line that follows the direction of transformation.

Even though the problem is very interesting, unfortunately it is also ill posed. Nothing guarantees that the sounds input by the user will define a direction of transformation. They might be better fitted by a closed-loop curve (such as a triangle or a circle) rather than a straight line. In principle it is possible to perform cyclic transformations following user-input sounds that close a loop. However, the concept of linearity must be carefully reconsidered. When the sounds input by the user follow no clear pattern, zig-zagging wildly all over the space, the corresponding transformation might sound rather perceptually discontinuous when following the predefined sequence. In any case, obtaining smooth transitions between each pair of sounds was already very challenging. One thing I knew for sure, in order to transform smoothly between any pair, I would need to be able to obtain intermediate sounds. But the real challenge was that I would also need to be able to have independent control of the features to propose a perceptually relevant solution. So I knew that the solution involved some sort of hybridization of sonic features. It had sound morphing

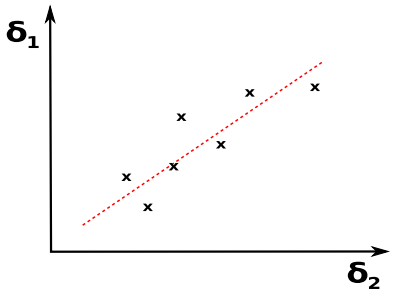

Figure 1: Depiction of user-defined abstract sound space. The figure illustrates the sequence of sounds input by the user as "x" in an abstract space defined by perceptually related features of the sounds.

written all over it.

When I started working on it, trying to find a solution to the sound hybridization problem, I realized that, even though everybody seems to be very excited about sound morphing for its creative potential, not many people have actually worked on it. My first attempts using pure sinusoidal analysis did not give satisfactory results. IRCAM's superVP didn't do a good job either. Most of the audio processing techniques were originally developed for speech and they are relatively unknown to the music technology oriented researchers, especially those working on musical instrument sounds. So I realized that this is I was supposed to do, apply the techniques developed for speech to musical instrument sounds whenever beneficial and develop new improved ones whenever necessary. This constitutes the main contribution of this work. The result is the work that I will describe in the next pages. I hope you will enjoy reading it as much as enjoyed doing it and writing about it.

Contents

1 Introduction **27**
1.1 Background . . . 27
1.1.1 Sound Morphing in a Nutshell . . . 29
1.2 Scope of the Thesis . . . 33
1.2.1 The Approach Adopted . . . 34
1.2.2 Contributions . . . 39
1.3 Overview of the Thesis . . . 40

I On Sound Morphing **43**

2 Morphing **45**
2.1 What is Morphing? . . . 45
2.1.1 Natural Hybridization . . . 47
2.1.2 Artificial Hybridization . . . 49
2.2 Theoretical Considerations . . . 50
2.2.1 Subjectiveness . . . 50
2.2.2 Correspondence . . . 51
2.2.3 Intermediateness . . . 52
2.2.4 Smoothness . . . 55
2.2.5 Conceptual Distance . . . 58
2.2.6 Morphing Algorithm . . . 59
2.2.7 Morphing Guided by Features . . . 60
2.2.8 Intuitive Control of Parameters . . . 61
2.2.9 Types of Transformation . . . 62

3 Morphing Sounds **65**
3.1 From Image Morphing to Sound Morphing: Conceptual Considerations . . . 65
3.1.1 Sound Object . . . 65
3.1.2 Sound Shape . . . 66
3.1.3 Sound Morphing . . . 66
3.2 The Image Morphing Analogy Revisited . . . 67
3.2.1 An Even Better Analogy: Movie Morphing . . . 67
3.3 Formalization . . . 68
3.3.1 Terminology . . . 68
3.3.2 Tentative Definitions . . . 69
3.3.3 What Sound Morphing is Not . . . 69
3.3.4 Sound Transformations that can be Described as Morphing . . . 71

3.3.5 Practical Aspects . . . 74

4 State of the Art **77**
4.1 Sound Signal Models . . . 77
4.1.1 Additive Synthesis . . . 78
4.1.2 Discrete Fourier Transform . . . 80
4.1.3 Short-Time Fourier Transform . . . 80
4.1.4 Phase Vocoder . . . 81
4.1.5 Classical Sinusoidal Modeling . . . 84
4.1.6 Spectral Modeling Synthesis . . . 87
4.2 Interpolation of Sinusoidal Models . . . 88
4.2.1 Interpolation Procedure . . . 89
4.3 Other Approaches . . . 90
4.3.1 Magnitude Spectrograms . . . 90
4.3.2 Physical Models . . . 92
4.3.3 Gaussian Mixture Models . . . 92
4.3.4 Wigner Distribution Analysis . . . 93
4.3.5 Neural Networks . . . 93
4.3.6 Wavelet Analysis . . . 94

5 Hybrid Musical Instruments **95**
5.1 Timbre Perception . . . 96
5.1.1 Timbre Revisited . . . 98
5.1.2 Timbre Spaces . . . 99
5.2 Features . . . 101
5.2.1 Temporal Features . . . 102
5.2.2 Spectral Features . . . 103
5.2.3 Calculation of Features . . . 105

II Estimation and Representation of the Source-Filter Model Parameters **109**

6 The Source-Filter Model **111**
6.1 The Source-Filter Model for Speech . . . 112
6.2 The Source-Filter Model for Acoustic Musical Instrument Sounds . . . 113
6.2.1 Acoustic Musical Instruments and Strong Coupling . . . 114
6.3 Resonances and the Spectral Envelope . . . 116
6.4 The Source-Filter Model from a Temporal Perspective . . . 118
6.5 Mathematical Modeling of Source and Filter . . . 118
6.5.1 Signal Processing Modeling of Source and Filter . . . 120
6.5.2 Estimation of Source and Filter . . . 121
6.5.3 Filter Modifications . . . 122

7 Spectral Envelope Estimation **123**
7.1 Formalization of Spectral Envelopes . . . 124
7.1.1 Estimation of the Spectral Envelope . . . 124
7.2 Linear Prediction . . . 126
7.2.1 Parameter Estimation . . . 127
7.3 Discrete All-Pole Model . . . 134

7.3.1 Limitations of Linear Prediction . . . 135
7.4 Cepstral Smoothing . . . 136
7.4.1 Homomorphic Systems . . . 137
7.4.2 Cepstrum . . . 139
7.4.3 Cepstral Smoothing . . . 144
7.5 Discrete Cepstrum . . . 144
7.6 True Envelope . . . 147
7.6.1 True Envelope Estimation . . . 148
7.6.2 Optimal Order Selection . . . 148
7.7 Alternative Spectral Envelope Representations . . . 149
7.7.1 Line Spectral Frequencies . . . 149
7.7.2 Conversion from Linear Prediction to Cepstral Based Representations . . . 155
7.7.3 Quantization Properties . . . 159

8 Temporal Evolution 161
8.1 Different Regions . . . 162
8.1.1 Attack . . . 162
8.1.2 Sustain . . . 163
8.1.3 Decay . . . 163
8.1.4 Release . . . 163
8.2 The Helmholtz Model . . . 163
8.3 The Classical Attack-Decay-Sustain-Release (ADSR) Model . . . 165
8.3.1 The Attack-Decay-and-Sustain-Release (AD&SR) Model . . . 165
8.3.2 The Attack-Rest (AR) Model for Percussive Sounds . . . 168
8.4 The Amplitude/Centroid Trajectory (ACT) Model . . . 169

9 Temporal Envelope Estimation 171
9.1 Early Temporal Envelope Estimation Techniques . . . 172
9.2 Low-Pass Filtering . . . 172
9.3 Root-Mean Square . . . 172
9.4 Analytic Expression . . . 173
9.4.1 Analytic Signal . . . 174
9.4.2 Hilbert Transform . . . 175
9.4.3 Temporal Envelope Power . . . 176
9.5 Frequency-Domain Linear Prediction . . . 176
9.5.1 Discrete Cosine Transform (DCT) . . . 177
9.6 True Amplitude Envelope (TAE) . . . 177
9.7 Comparison of Amplitude Envelope Estimation Techniques . . . 178
9.8 Model Conversion . . . 178

III Morphing Musical Instrument Sounds Guided by Sonic Features 181

10 Vienna Sound Database 183
10.1 Vienna Sound Database . . . 183
10.1.1 Woodwinds . . . 186
10.1.2 Brass . . . 187
10.1.3 Strings . . . 188

11 Overview of the Method **189**
11.1 Sound Morphing Algorithm . . . 189
11.2 Temporal Processing . . . 190
11.2.1 Temporal Segmentation . . . 190
11.2.2 Temporal Alignment . . . 190
11.2.3 Temporal Envelope Estimation . . . 191
11.3 Spectral Processing . . . 191
11.3.1 Sinusoidal plus Residual Decomposition . . . 192
11.3.2 Spectral Modeling . . . 192
11.4 Morphing Procedure . . . 195
11.4.1 Spectral Morphing . . . 196
11.4.2 Temporal Envelope Morphing . . . 197

12 Temporal Alignment **203**
12.1 Temporal Segmentation . . . 204
12.1.1 Automatic Segmentation . . . 204
12.1.2 Automatic Detection of Boundaries . . . 205
12.1.3 Examples of Automatic Segmentation . . . 206
12.2 Temporal Alignment . . . 207
12.2.1 Interpolation of Lengths . . . 208
12.2.2 Calculate Time-Stretch Factors . . . 208
12.2.3 Temporal Alignment . . . 208

13 Spectral Envelope Morphing **217**
13.1 Morphing Spectral Envelopes . . . 218
13.1.1 Spectral Peak Shifting . . . 219
13.1.2 Spectral Peak Rise and Wane . . . 220
13.1.3 No Spectral Peak Correspondence . . . 220
13.1.4 Spectral Envelope Morphing . . . 221
13.1.5 Balance of Spectral Energy . . . 224
13.2 Target Feature Values . . . 225
13.3 The Problem of Moments . . . 226
13.3.1 Statistical Moments . . . 226
13.3.2 Standardized Moments . . . 227
13.3.3 Characteristic Function . . . 227
13.3.4 Analytical Formulation . . . 228
13.4 Manipulation of Spectral Envelope Representations . . . 229
13.4.1 Interpolation of Spectral Envelope Representations . . . 230
13.4.2 Spectral Envelope Model Conversion . . . 230

14 Evaluation **235**
14.1 Evaluation of the Source-Filter Model . . . 236
14.2 Evaluation of the Transformation . . . 237
14.2.1 Objective Evaluation . . . 238
14.2.2 Subjective Perceptual Evaluation . . . 244
14.2.3 Temporal Envelope Morphing . . . 245

15 Conclusions and Future Perspectives **261**
15.1 Comments and Remarks . . . 262
15.1.1 Perceptual Similarity . . . 262
15.1.2 Perceptual Linearity . . . 264
15.2 Type of Transformation . . . 266
15.3 Sinusoidal plus Residual Decomposition . . . 266
15.4 Source-Filter Model . . . 266
15.5 Temporal Processing . . . 267
15.5.1 Automatic Segmentation . . . 267
15.5.2 Temporal Alignment . . . 268
15.5.3 Temporal Envelope Estimation . . . 269
15.5.4 Temporal Envelope Morphing . . . 269
15.6 Spectral Morphing . . . 269
15.6.1 Spectral Envelope Estimation . . . 269
15.6.2 Spectral Envelope Morphing . . . 270
15.7 Sound Material . . . 270
15.8 Formalization . . . 271
15.8.1 Correspondence . . . 271
15.8.2 Intermediateness . . . 272
15.8.3 Smoothness . . . 272
15.9 Memory Effect . . . 273
15.10 Categorical perception . . . 273
15.11 Conceptual Distance . . . 273
15.12 Timbre Spaces and Morphing . . . 274
15.12.1 Is Perception of Morphing Metric? . . . 274

IV Appendices **289**

A Linear Prediction **291**
A.0.2 Filter Stability . . . 291
A.1 Frequency Domain Formulations . . . 292
A.1.1 Stationary Case . . . 293
A.1.2 Nonstationary Case . . . 293
A.1.3 Linear Predictive Spectral Matching . . . 294
A.1.4 Modeling Discrete Spectra . . . 296
A.2 Error Analysis . . . 296
A.2.1 The Minimum Error . . . 297
A.2.2 Spectral Matching Properties . . . 297
A.2.3 The Normalized Error . . . 299
A.2.4 A Measure of Ill-Conditioning . . . 300

B Discrete All-Pole Model **303**
B.1 Properties of The Error Measure . . . 303
B.1.1 Error Minimization . . . 304
B.1.2 Minimum Error . . . 305
B.1.3 The Solution and its Uniqueness . . . 305

C Cepstral Smoothing **307**
C.1 The Phase Cepstrum . . . 307
C.2 Linear Phase Components . . . 308
C.3 Spectrum Notching . . . 309
C.4 Aliasing . . . 309
C.5 Oversampling . . . 310
C.6 Zero-Padding . . . 310
C.7 Effects of Windowing . . . 311
C.7.1 The Effect of Windowing the Log Spectrum . . . 312
C.7.2 The Effect of Windowing the Complex Cepstrum . . . 312

D Discrete Cepstrum **315**
D.1 Regularized Estimation of the Discrete Cepstrum . . . 315

E Line Spectral Frequencies **319**
E.1 The Fundamental Theorem of Palindromic Polynomials . . . 319

F Perceptual Similarity for Musical Instrument Sound **321**
F.1 Sound Representations . . . 321
F.2 The Test . . . 321
F.3 Framework . . . 321
F.4 Perceptual Similarity . . . 322
F.5 Your Task . . . 322
F.6 Recommendations . . . 322
F.7 Listening conditions . . . 322

G Evaluation of the Spectral Smoothness of Sound Morphing Algorithms **323**
G.1 Sound Morphing . . . 323
G.2 The Test . . . 323
G.3 Framework . . . 323
G.4 Smoothness . . . 324
G.5 Your Task . . . 324
G.6 Recommendations . . . 324

List of Figures

1 Depiction of user-defined abstract sound space 12

1.1 Schematic view of computer sound transformations 28
1.2 Example of multidimensional MDS timbre space 30
1.3 Temporal difference between morphing transformations 31
1.4 Depiction of the classical morphing scheme using the interpolation principle 33
1.5 Depiction of the source-filter model 35
1.6 Schematic representation of the temporal alignment 37
1.7 Frame to frame correspondence after temporal alignment 38
1.8 Spectral correspondence 38

2.1 Depiction of object morphing 46
2.2 Depiction of the problem of interpolation of shape 47
2.3 Depiction of two possible applications of morphing 48
2.4 Geometrical representation of morphing 49
2.5 Depiction of face hybridization 50
2.6 Depiction of face morphing as a special case of face hybridization 51
2.7 Depiction of the fundamental steps of morphing 52
2.8 Depiction of image morphing 53
2.9 Depiction of the concept of intermediatness 54
2.10 Convex combination 56
2.11 Continuous vs Categorical perception 57
2.12 Visible spectrum of light 57
2.13 Conceptual distance between objects 58
2.14 Another example of face morphing 60
2.15 Morphing by feature interpolation principle 61
2.16 Illustration of perceptually relevant transformations 62

3.1 Depiction of movies 68
3.2 Depiction of Warped Dynamic Morphing 72
3.3 Depiction of static or stationary morphing 73
3.4 Depiction of dynamic morphing 74
3.5 Depiction of cyclostationary morphing 74

4.1 Filter-bank vs. Fourier transform interpretation 81
4.2 Depiction of the peak matching algorithm 86
4.3 Optimal partner search problem 90
4.4 The three stages of audio morphing 91

5.1 Example of multidimensional timbre spaces . . . 96
5.2 Timbre, tone color, and sound quality . . . 97
5.3 Hierarchical model of the auditory image . . . 98
5.4 Orthogonal metric space . . . 100
5.5 Descriptor extraction flowchart . . . 102
5.6 Skewness . . . 105
5.7 Kurtosis . . . 105
5.8 Perceptual Spectral Shape Descriptors . . . 106

6.1 Schematized diagram of the vocal apparatus . . . 112
6.2 Linear model of speech production . . . 113
6.3 Changing pitch under weak and strong coupling . . . 115
6.4 Strongly and weakly coupled components of the source-filter model . . . 116
6.5 The construction of the violin . . . 117
6.6 Temporal evolution . . . 119
6.7 Spectral representation of partials . . . 121
6.8 Spectro-temporal illustration of the source-filter model . . . 122

7.1 Canonic form for system for homomorphic deconvolution . . . 138
7.2 Frequency domain representation of a homomorphic system for convolution . . . 139
7.3 Cepstral Smoothing . . . 145
7.5 An example LPC spectrum overlaid with the corresponding vertical LSP frequencies 150
7.6 The altered set of LSPs . . . 155
7.4 "True Envelope" estimation . . . 160

8.1 The Helmholtz model of acoustic musical instrument sounds . . . 164
8.2 ADSR model . . . 166
8.3 Attack Decay and Sustain Release model . . . 167
8.4 Attack Rest model of temporal evolution . . . 168
8.5 Method of efforts . . . 169
8.6 The amplitude/centroid trajectory (ACT) model . . . 170

9.1 True amplitude envelope estimation technique . . . 178
9.2 Amplitude envelope estimation techniques . . . 179

10.1 Pitch range for musical instruments . . . 185
10.2 Woodwind instruments . . . 186
10.3 Brass instrument . . . 187
10.4 String instruments . . . 188

11.1 Musical instrument sound morphing algorithm . . . 190
11.2 Spectral modeling . . . 192
11.3 Sinusoidal plus residual decomposition . . . 198
11.4 Spectral view of the source-filter model . . . 199
11.5 Comparison between sinusoidal and source-filter representations . . . 200
11.6 Temporal variation of spectral shape features . . . 201

12.1 Temporal alignment . . . 204
12.2 Temporal segmentation . . . 205
12.3 Automatic segmentation for the woodwinds I . . . 209
12.4 Automatic segmentation for the woodwinds II . . . 210

12.5 Automatic segmentation for brass I 211
12.6 Automatic segmentation for brass II 212
12.7 Automatic segmentation for brass III 213
12.8 Automatic segmentation for strings I 214
12.9 Automatic segmentation for strings II 215

13.1 Hybrid spectral envelopes 219
13.2 No spectral peak correspondence 221
13.3 Original spectral envelope curves 223
13.4 Interpolation of line spectral frequency representation 224
13.5 Interpolation of spectral envelope parameters 225
13.6 Different statistical distributions 226
13.7 Spectral envelope morphing 231
13.8 Accuracy of spectral envelope representation 233

14.1 Perceptual similarity 239
14.2 Behavior of spectral peaks I 246
14.3 Behavior of spectral peaks II 247
14.4 Behavior of spectral peaks III 248
14.5 Behavior of spectral peaks IV 249
14.6 Behavior of spectral peaks V 250
14.7 Behavior of spectral peaks VI 251
14.8 Behavior of spectral peaks VII 252
14.9 Behavior of spectral peaks VIII 253
14.10Error calculation 254
14.11Error calculation for each spectral frame 254
14.12Error analysis 255
14.13Perceptual linearity 256
14.14Morphing the temporal envelope curve 257
14.15Morphing the temporal envelope cepstral coefficient representation 258
14.16Error analysis for temporal envelope morphing 259

15.1 Nonmetric representation of conceptual objects 275
15.2 Example of multidimensional timbre spaces 275

List of Tables

7.1 Values of m 157
7.2 Decomposition of n into positive integers 157

10.1 Vienna sound database 184

14.1 Numerical scale for similarity test 237
14.2 Sounds used in the listening test 238
14.3 Behavior of spectral envelope peaks 241

Prologue

"Once upon a time, in a very lonely place, there lived a man endowed by Nature with extraordinary curiosity and a very penetrating mind. For a pastime he raised birds, whose songs he much enjoyed; and he observed with great admiration the happy contrivance by which they could transform at will the very air they breathed into a variety of sweet songs. One night this man chanced to hear a delicate song close to his house, and being unable to connect it with anything but some small bird he set out to capture it. When he arrived at a road he found a shepherd boy who was blowing into a kind of hollow stick while moving his fingers about on the wood, thus drawing from it a variety of notes similar to those of a bird, though by quite a different method. Puzzled, but impelled by his natural curiosity, he gave the boy a calf in exchange for his flute and returned to solitude. But realizing that if he had not chanced to meet the boy he would never have learned of the existence of a new method of forming musical notes and the sweetest songs, he decided to travel to distant places in the hope of meeting with some new adventure. As the man roved, he encountered songs made by a bow sawing upon some fibers stretched over a hollowed piece of wood, by the hinges of a temple gate, by a man running his fingertip around the rim of a goblet, and by the beating wings of wasps. As his wonder grew, his conviction proportionately diminished that he knew how sounds were produced; nor would all his previous experiences have sufficed to teach him or even allow him to believe that crickets derive their sweet and sonorous shrilling by scraping their wings together, particularly as they cannot fly at all. Well, after this man had come to believe that no more ways of forming tones could possibly exist...he suddenly found himself once more plunged deeper into ignorance and bafflement than ever. For having captured in his hands a cicada, he failed to diminish its strident noise either by closing its mouth or stopping its wings, yet he could not see it move the scales that covered its body, or any other thing. At last he lifted up the armor of its chest and there he saw some thin hard ligaments beneath; thinking the sound might come from their vibration, he decided to break them in order to silence it. But nothing happened until his needle drove too deep, and transfixing the creature he took away its life with its voice, so that he was still unable to determine whether the song had originated in those ligaments. And by this experience his knowledge was reduced to diffidence, so that when asked how sounds were created he used to answer tolerantly that although he knew a few ways, he was sure that many more existed which were not only unknown but unimaginable."

Galileo Galilei, from Il Saggitore (The Assayer) 1623

Chapter 1

Introduction

This thesis is about morphing musical instrument sounds. Thus the focus is on timbre and how to control spectral and temporal characteristics of the sounds produced by quasi-harmonic acoustic musical instruments to obtain a gradual transformation. In essence, sound morphing is intrinsically a perceptual problem. Consequently, the morphing transformation is usually required to produce perceptually intermediate results that would be associated with hybrid musical instruments. The aim of this thesis is the development of a method that allows intuitive control of the morphing transformation guided by perceptually motivated features.

Sound transformations weave the background of sound morphing techniques, which can be considered to be a specific subset of transformations with specific characteristics. In this work, the main purpose is to study morphing from a technical point of view. Therefore, theoretical, conceptual, and aesthetic aspects of the problem are approached from a practical perspective. The central questions investigated concern the requirements for a sound transformation to be considered morph, which transformations respect these conditions, how to obtain a perceptually linear morph, and how to objectively evaluate the results.

Sound transformations, morphing among them, have been used in different contexts, artistic, technical and scientific. There are different possible types of morph, depending on the intended application and desired effect. Technically, some of the morphing transformations are more difficult to achieve than others. This thesis studies a very challenging case, dubbed cyclostationary morph, which produces several instances of the departure sound that become progressively closer to a target sound. The result of the cyclostationary morph is evaluated using three criteria, namely correspondence, intermediateness, and smoothness.

An important contribution of this work is the development of an implementation of the source-filter (SF) model based on a sinusoidal plus residual decomposition of acoustic musical instrument sounds. The SF model is very appropriate to independently manipulate spectral and temporal characteristics of the sounds using features as guides. The approach adopted in this work impels the investigation of perceptual and conceptual consequences of sound morphing, such as whether the mental representation of musical instrument timbre is inherently categorical or metric in nature.

1.1 Background

The 20th century witnessed a paradigm shift in music. According to Trevor Wishart [Wishart, 1996], the main responsible for this important change in our vision of what constitutes music is closely related to the invention of sound recording, and later sound manipulation procedures, usually intimately connected with the medium used to store the sounds. Sound record-

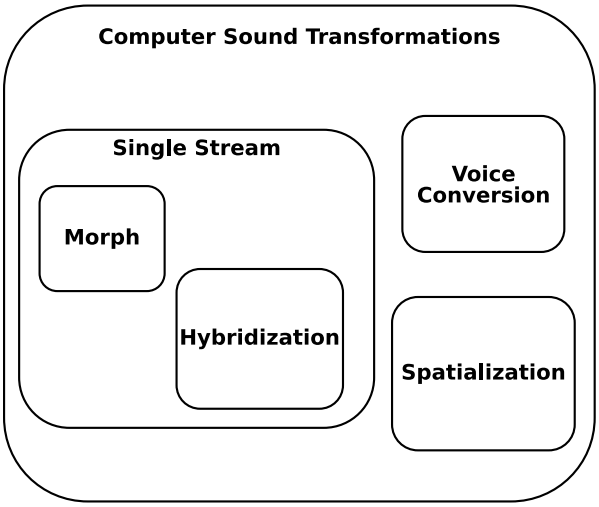

Figure 1.1: Schematic view of computer sound transformations.

ing and especially manipulation have broadened our knowledge about the nature of sounds and how we perceive them. Sound manipulation techniques whose aim is to change the way sounds are perceived, called sound transformations, have been the focus of interest from both compositional [Landy, 2011, Wishart, 2011] and technical [Amatriain et al., 2003, Amatriain et al., 2002] standpoints.

The advent of the digital computer is at the core of the revolution in the representation and manipulation of sounds. There are no theoretical limitations to the use of the computer in sound representation and manipulation, and even as a source of new sounds that were before unimaginable. Computer sound transformations are a very challenging class of sound transformations that cannot be performed by analog means. Notable examples are sound transformations that require that the result be perceived as a single auditory stream, such as sound hybridization (also called cross-synthesis) and morphing, as shown in the schematic view in figure 1.1. Therefore, there is a burgeoning interest in the search for computational techniques that allow the user to obtain perceptually relevant sound transformations and seamless transitions because computer sound transformations are widespread in a myriad of activities, such as music composition, sound design, and sound synthesis, among others.

An important requirement of any sound transformation technique from the user point of view is to have intuitive control of the results. Manipulation of the values of the parameters that control the transformation should lead to predictable changes. An example of counterintuitive parameter control can be found in frequency modulation techniques, where changing the value of one parameter sometimes can lead to rather unpredictable changes in the perceptual aspects they affect. A desireable aspect of sound transformation techniques is to have parameters independently control perceptually relevant characteristics of the results. Increasing the value of a parameter should increase a perceptual characteristic of the result without changing the others.

Among the many different possible sound transformations [Serra and Bonada, 1998], the manipulation of characteristics of sounds that are related to timbre perception stand out as the most exciting to date. Sound morphing figures prominently as one of the most interesting timbre manipulation techniques due to its enormous creative potential, opening up an exciting new world of possibilities.

Many sound transformations are referred to as morphing in the literature, ranging from artistic contexts, such as music compositions [Wishart, 1996, McNabb, 1981, Harvey, 1981], to more technical applications, such as the design of sound synthesizers [Tellman et al., 1995], and even in psychoacoustic experiments, notably to study timbre spaces [McAdams et al., 2006, Caclin et al., 2005, Krimphoff et al., 1994, McAdams et al., 2005, Grey, 1975, Grey and Gordon, 1978].

1.1.1 Sound Morphing in a Nutshell

In music compositions, sound morphing allows the exploration of the sonic continuum, as theoretically proposed by Wishart [Wishart, 1996]. There exist notable examples of the exploration of the sonic continuum in music composition using sound morphing techniques. Jonathan Harvey's "Mortuos Plango, Vivos Voco" [Harvey, 1981] morphs seamlessly from a vowel sung by a boy to the complex and rich spectrum of a bell consisting of many partials. Another example is Trevor Wishart's "Red Bird" where the word 'listen' gradually morphs into birdsong, among other unusual morphs [Wishart, 1996]. Wishart himself mentions [Wishart, 1997] Michael McNabb's "Dreamsong" [McNabb, 1981] and its particularly striking opening and closing morphs. Most of these examples were achieved by hand, either using studio techniques or with the aid of the computer. Ideally, we would like to have an automatic morphing technique that takes as input the sounds we want to morph between and how we want to do the transformation and automatically outputs the result.

The composition and manipulation of audio to create a certain effect or have a specific perceptual impact is one of the most challenging tasks sound designers face. Sound morphing figures prominently among the techniques that can be used to achieve such impact. Some of the software available for sound design (be it commercial, shareware or open source) include sound transformation techniques that are usually referred to as morphing. A remarkable example is Pete Johnston's Bantu webpage `http://www.bantusound.com/SoundMorphing/SoundMorphingPage.html`, which includes examples online designed using Kyma workstation. Kyma Sound Design Workstation `http://www.symbolicsound.com/cgi-bin/bin/view/Company/WebHome` is a commercial sound design environment. Kyma's webpage also includes some sound morphing examples. Loris `http://www.cerlsoundgroup.org/Loris/` (the defunct Lemur `http://www.cerlsoundgroup.org/Lemur/`) is a shareware Macintosh application for sound analysis, transformation, and synthesis based on sinusoidal analysis. Loris' webpage includes interesting tutorials on sound morphing using sinusoidal models. Finally, the Composer's Desktop Project (CDP) is an international network of composers and programmers that provides a software-only music system designed specifically to transform existing sound samples for musical purposes `http://www.composersdesktop.com/`. Morphing is one of the transformations.

Tellman et al. [Tellman et al., 1995] propose a sound morphing technique based on sinusoidal modeling that is intended to improve the performance of a sample-based synthesizer. They morph between sounds of the same instrument to obtain intermediate pitches, dynamics, and other effects. The most interesting results of sound morphing, though, are obtained when we morph between different musical instruments to obtain sounds that would correspond to hybrid instruments.

Sound morphing can notably be used in psychoacoustic experiments to study timbre perception. Figure 1.2 shows a three-dimensional timbre space obtained with multidimensional scaling (MDS) by Grey [Grey and Gordon, 1977] with acoustic musical instruments occupying specific spots. The MDS algorithm maps the subjective distances into an orthogonal metric space which has the number of dimensions specified by the investigator [Grey and Gordon, 1977]. The distances between pairs of instruments represent the perceptual (dis)similarity between them when the sounds have the same pitch, duration and dynamics. Notice that timbre spaces obtained from acoustic musical instrument sounds are essentially sparse in nature, which means that the space is mostly void and each musical instrument corresponds to a small portion that does not overlap with others. This is

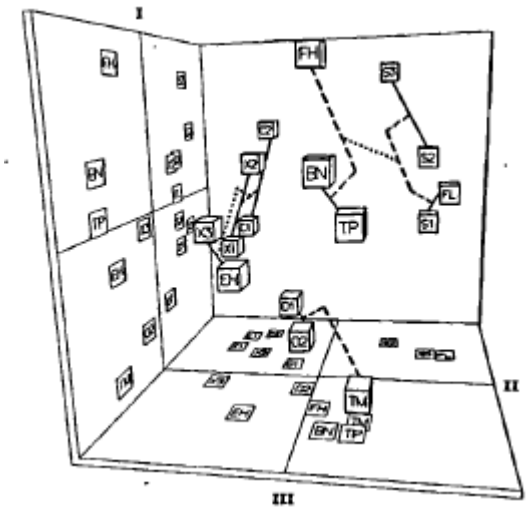

Figure 1.2: Example of multidimensional timbre spaces. After Grey [Grey and Gordon, 1977]

where sound morphing plays its fundamental role. When the morphed sound represents a hybrid between two musical instrument sounds, it would be placed between them in the underlying timbre space. Sound morphing would fill the voids and permit exploration of the sonic continuum.

However, not all of the transformations called sound morphing in the literature are conceptually similar. Figure 1.3 illustrates the two main types of transformation that are usually called sound morphing, but which differ notably in the temporal dimension. In figure 1.3 the two sounds being morphed, shown at the top, are represented by different shapes. In this simplified example, each shape represents all the temporal and spectral characteristics of sounds. When the transformation occurs during the course of the sound, we call it dynamic morphing and represent it in figure 1.3 as a transformation from the shape used to represent sound 1 to that used to represent sound 2. Notice that the shapes change dynamically while the morphed sound is heard. On the other hand, when the morphed sound is perceived as equally intermediate during all its duration, we call the transformation stationary morphing and represent it as the same intermediate shape along the whole duration of the morphed sound. Depending on the temporal nature of the transformation, stationary or dynamic, the morphed sound could either correspond to a point in timbre space or to a trajectory.

This lack of uniformity in nomenclature calls for the formalization of a theory of morphing, clearly defining which type of transformation can be considered morphing and allowing us to categorize them according to certain criteria. Defining precisely and formally the requirements of the desired morphing process has proved to be a difficult problem [Caetano and Rodet, 2010b]. On the other hand, describing the desired result is not any easier (partly so because more than one type of transformation is commonly called morphing). For now, suffice it to say that simply playing the instruments simultaneously is not enough to achieve the desired auditory result. The ear is capable of telling multiple auditory streams apart (otherwise listening to polyphonic music would be an entirely different experience). So, a requirement of the sound transformation usually called morphing is that the result should fuse into a single percept that somehow resembles both instruments at the same time. This requirement rules out mixing the sounds to try to achieve stationary morphs and cross-fading to try to perform dynamic morphs. As we will see later, from

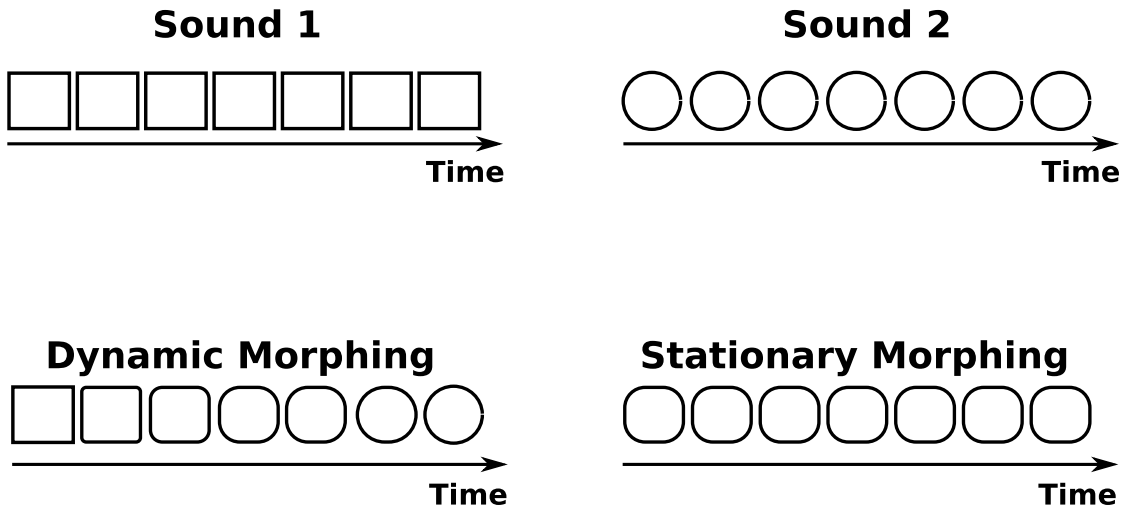

Figure 1.3: Temporal difference between the two main types of transformation usually called morphing.

a technical point of view, stationary morphing is a much more challenging problem than dynamic morphing because we need to transform both temporal and spectral characteristics of sounds.

This work investigates musical instrument sound morphing across dimensions of sound perception usually associated with timbre. This is sometimes called timbre morphing [Tellman et al., 1995] or timbre interpolation [Osaka, 1995] in the literature. The focus of interest gravitates around this type of question: "What is the auditory result of the morph between a violin sound and a french horn sound?" As a result, most of the techniques described in this document were developed specifically to deal with the aspects of the problem from the point of view of a specific set of sonic features that somehow capture relevant information related to the perception of musical instrument sounds.

In this work, the term "parameter" refers to coefficients that can be used to resynthesize sounds, while "feature" refers to coefficients used to describe or identify a particular aspect of a sound. Usually, we cannot resynthesize sounds directly from feature values. Waveforms are useful representations of sounds because they can be played (and therefore heard). More sophisticated models are useful to allow the investigation of different aspects of sounds that are not clear from the waveform representation. Whenever we want to hear sounds, we need to invert the parameters of the model back to its waveform representation. This process, usually called resynthesis, cannot be done from just any model representation. Some models represent information that is not invertible. Sound manipulation, on the other hand, is usually performed by changing the values of the parameters of a given sound representation to change the way the sound is perceived.

One very challenging aspect of morphing is that we want to be able to control the transformation with a single parameter α, called morphing or interpolation factor. Ideally, the morphing factor α should control perceptual aspects of the transformation, such that the morphed sound should be perceived as halfway between source and target when $\alpha = 0.5$, for example. Notice that, following this requirement, successive application of the transformation procedure with different values of the morphing factor α ranging from 0 to 1 would produce a sequence of sounds that starts perceptually very close to one of the sounds being morphed, and gradually shifts closer and closer to the other one. This is the ultimate goal of the morphing procedure described here and the results will be

evaluated as to whether this task is successfully accomplished with the implementation of the source-filter model.

One important and difficult requirement we face when morphing sounds is that it is a perceptually related problem. That is, when we define the desired result, we usually do so by means of psychological dimensions of sound perception such as pitch and loudness; or dimensions related to timbre perception, such as sharpness of attack, brightness, onset asynchrony, spectral flux, roughness, among others. On the other hand, the classical approach to sound morphing is to interpolate the parameters of a model. The idea behind the interpolation principle illustrated in figure 1.4 is that if we can represent two different sounds by simply adjusting the parameters of a model used to describe them, we obtain a somewhat smooth transition between the sounds by interpolating between these parameters. However, when blindly following the interpolation principle, the perceptual impact of the interpolation of parameters depends largely on what type of parameters we chose to represent the sounds with. If the parameters of the model we used represent information that is not directly related to how we perceive sounds, the result of the interpolation of these parameters will very likely have very little perceptual significance.

The adoption of a sound model whose parameters are more closely or directly related to sound perception can greatly improve the results when morphing musical instrument sounds. Morphing musical instrument sounds guided by perceptually related features requires a sound model whose parameters are closely related to perceptual aspects. This is a difficult problem because it involves the representation and manipulation of both temporal and spectral features of musical instrument sounds, which are often interdependent. The independent representation and manipulation of such features is a key aspect of the problem and, as such, in this work it led naturally to the adaptation of the source-filter model of speech production [Rabiner and Schafer, 1978] to the problem at hand. It is generally accepted that the source is associated with prosody for speech and expressivity for music, and the filter carries information about speaker [Stylianou, 2008] or musical instrument [Peeters, 2003] identification. I studied techniques for the estimation and representation of each part of the model aiming at the specific characteristics and requirements of stationary morphing of musical instrument sounds. In this work, the filter is represented as the spectral envelope (estimated with true envelope and manipulated in the line spectral frequency domain), while a very flexible representation of the source is sinusoidal models for the quasi harmonic component and white noise for the noisy residual. Finally, the parameters of the source-filter model allow the direct manipulation of signal-level counterparts (sonic features) of dimensions of musical instrument sound perception. One example is the fundamental frequency, which can be extracted from the sound signal and is correlated to pitch perception. Chapter 5 develops in depth the correlation between musical instrument timbre perception and the features adopted to evaluate the morphing results.

Although there are a few exceptions [Osaka, 1998], most authors pose the problem of morphing sounds using perceptual requirements, but hardly ever perform perceptual evaluations of their results mainly because perceptual evaluations are cumbersome and costly, and there are no standard evaluation criteria established for sound morphing. Actually, most works about morphing sounds [Ahmad et al., 2009, Boccardi and Drioli, 2001, Ezzat et al., 2005, Fitz et al., 2003, Hatch, 2004, Osaka, 2005, Osaka, 1995, Röbel, 1998, Tellman et al., 1995, Williams and Brookes, 2007, Williams and Brookes, 2009] include the description of the type of transformation referred to as morphing (due to the lack of consensus in the literature), the sound model used (which can vary greatly depending on the type of sound material being morphed), the approach to apply the interpolation principle and then some examples, usually spectrograms of morphed sounds using the method. Very seldom do the authors of these works present any evaluation of their results.

One very important contribution of this work lies in the evaluation procedure adopted, following the three criteria originally proposed by Osaka [Osaka, 1998] to evaluate sound morphing

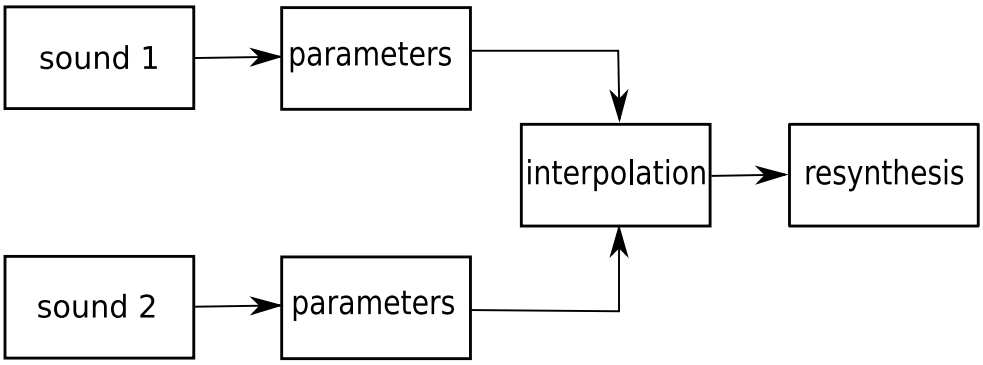

Figure 1.4: Depiction of the classical morphing scheme using the interpolation principle.

algorithms, namely correspondence, intermediateness, and smoothness, appropriately adapted to the main goal of this work described above. The objective evaluation uses an error measure in a quantitative evaluation of the results. The proposal is to use temporal and spectral features of sounds that can be estimated directly from the parameters of the source-filter model, such as log attack time and spectral centroid, as objective measure. The values of the features are used in the quantitative evaluation as a means to objectively estimate the perceptual impact of the transformation under the three evaluation criteria adopted.

The evaluation investigates how accurately the morphing factor α controls the morph. There is an emphasis on the spectral envelope morphing procedure. The interpolation properties of several spectral envelope representations are investigated according to two criteria, behavior of the spectral peaks and variation of the spectral shape features (spectral centroid, spread, skewness, and kurtosis). Finally, the results of listening tests done to compare and validate the results will be discussed. There was a perceptual similarity test to validate the source-filter (SF) model developed, and a "smoothness" test, comparing morphs obtained with the SF and the traditional sinusoidal model.

The bulk of the work described here is devoted to the techniques I developed to estimate and manipulate the parameters of the source-filter model in order to produce the variation in the values of the features that would reflect the desired transformation on the perceptual level. The temporal representation involves the estimation of the temporal envelope and the automatic segmentation of the sounds into perceptually relevant regions, such as attack, steady state, and release. The spectral model is estimated for every frame of a spectro temporal representation of the sound. For each frame we estimate the spectral envelope to represent the filter and we use a sinusoidal model to represent the source part of the model. After the estimation step, we can perform the proper manipulation of the parameters of the model according to the requirements of the problem of sound morphing.

1.2 Scope of the Thesis

The main objective of this work is to morph isolated musical instrument sounds across timbre dimensions, so as to produce the perception of hybrid musical instruments. The problem of morphing sounds is essentially perceptual because most formulations use perceptual requirements to define their objectives. However, this would require a perceptual evaluation of the results as well, which can be cumbersome and very long.

In this work, I propose to approach the problem of sound morphing from a more technical point of view, which allows for the adoption of objective evaluation methods. I propose to adapt the evaluation criteria originally proposed by Osaka [Osaka, 1998] and use a set of sonic features correlated with musical instrument timbre perception to compare the results of the method I proposed with a standard sinusoidal modeling approach.

The musical instruments used in the experiments reported in this work were selected from the Vienna database, which contains recordings of sounds from musical instruments commonly found in orchestras. This means that the sounds available are quasi-harmonic and most are sustained (rather than percussive).

1.2.1 The Approach Adopted

When we think of the applicability of morphing, the number of events contained in the source material we use greatly influences the type of transformation that can be performed. When we have sounds that are perceived as continuous isolated events, such as a long trumpet note or a scream, dynamic transformations usually produce impressive results with considerably little effort because we only need to perform the spectral transformation. However, when we consider sounds that can be segmented into several discrete events, such as a melody played by a musical instrument or a sentence uttered by a speaker, only performing spectral transformation is not enough. In this case we need to match the events (correspondence) and transform between each pair taking temporal and spectral aspects into account, such as attack and duration, among many others. One interesting application of this type of transformation would be morphing between two instruments playing the same melody, for example. If we have two recordings of two different musical instruments playing the same melody, we could imagine a transformation where the melody starts being played by one of the instruments that gradually morphs into the other, playing the same melody. In this case, each note of the melody is a discrete event that has to be morphed separately, and temporal features such as attack and duration need to be transformed as well. This type of gradual transformation between different sonic events, which is the main applicability of the work described here, requires techniques to morph temporal and spectral characteristics of each event, as well as an appropriate sound model. The type of sound morphing studied, called cyclostationary morph in this text and described in more detail in chapter 3, was chosen to satisfy the requirements of challenging transformations such as morphing many instances of discrete sonic events. As a brief overview of the work described in this text, I will describe in general lines the aspects of the sound model adopted, and then the morphing procedure.

1.2.1.1 Source-Filter Modeling

The source-filter model was originally proposed to explain speech production [Rabiner and Schafer, 1978, Rabiner, 1993]. According to this model, speech is viewed as the result of passing a glottal excitation signal (source) through a time-varying linear filter that models the resonant characteristics of the vocal tract. The most well known source-filter system is based on linear prediction (LP) of speech [Markel and Gray, 1976, Makhoul, 1975]. In its simplest form, a time-varying filter modeled as an autoregressive filter is excited by either quasiperiodic pulses (during voiced speech), or noise (during unvoiced speech). A more compact and at the same time flexible representation of the excitation signal has been proposed from a family of signal representations referred to as sinusoidal models [McAulay and Quatieri, 1986]. For musical instrument sounds, the filter is associated with the resonant cavity of the instrument, and the source with the excitation method. Figure 1.5 illustrates the use of the source-filter model from a spectral perspective in part a), and part b) shows the spectro-temporal representation of sounds, which corresponds to a temporal succession of the representation in part a). The source-filter

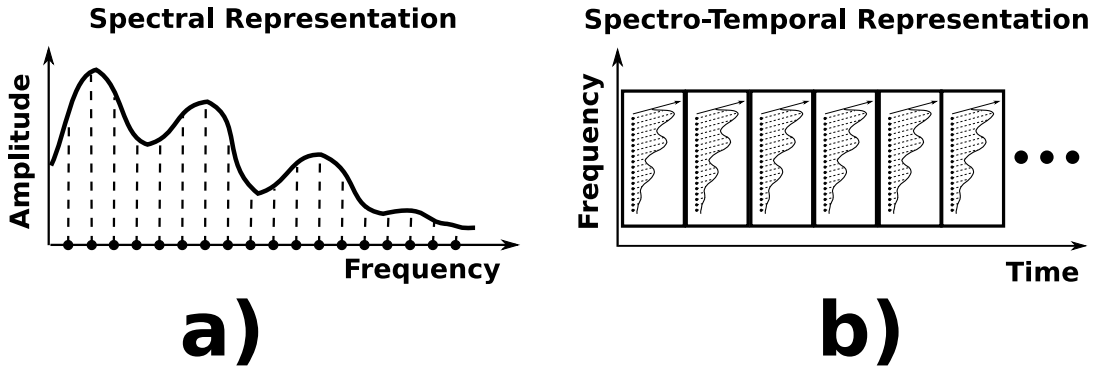

Figure 1.5: Depiction of the source-filter model. The figure shows the spectral representation of the source and the filter in part a) and the spectro-temporal representation, which corresponds to the temporal succession of the representation in part a)

model of musical instrument sounds developed in this work models the filter as the spectral envelope and the source as either sinusoids or white noise, after decomposing the signal into a sinusoidal (quasi-harmonic component) and a noisy residual. The theoretical and mathematical source-filter model developed will be presented in Chapter 6.

1.2.1.2 Sinusoidal plus Residual Decomposition

The musical instrument sound $y(t)$ is separated into a sinusoidal component $y_s(t)$ plus a residual component $y_r(t)$ as follows

$$y(t) = y_s(t) + y_r(t) \tag{1.1}$$

where $y_r(t)$ is obtained by subtraction of the purely sinusoidal component $y_s(t)$ from the original sound $y(t)$ as follows $y_r(t) = y(t) - y_s(t)$. Both the sinusoidal component $y_s(t)$ and the residual component $y_r(t)$ are modeled as source and filter. The filter component of both is modeled via spectral envelope estimation, while the sources are modeled separately. The source part of the sinusoidal component is modeled as sinusoids using sinusoidal analysis [McAulay and Quatieri, 1986, Serra and Smith, 1990], and the source part of the residual component is modeled as white noise. Chapter 7 presents spectral envelope estimation, which corresponds to the filter modeling. Sinusoidal modeling is presented earlier in the text, in chapter 4, as part of the review of the models commonly used in the literature of sound morphing.

1.2.1.3 Signal-Level Features and Perception

However controversial and notoriously avoided for being ill-defined, the concept of timbre is intrinsically intertwined with musical instruments. In this work, the focus on morphing sounds from different musical instruments demands a specific attention to aspects of sound perception normally associated with timbre. Therefore, sonic features that are correlated with timbre perception are extremely relevant in the context of this work. Among many possible choices (tristimulus, zero crossing rate, etc), the sonic features that are correlated with dimensions of timbre spaces studied in psychoacoustics, such as spectral centroid and log attack time, are very convenient because they can be directly calculated from the source-filter model representation adopted, and this representation can be used to retrieve the parameters of a sinusoidal plus residual model that are used for

resynthesis. The temporal sonic features considered are log attack time, temporal centroid and duration of transition, steady state and release. The spectral sonic features are spectral centroid, spread, skewness and kurtosis. The temporal sonic features can be estimated from the temporal representation of the sounds, and are mainly related to the perceptual dimension of percussiveness [Skowronek and McKinney, 2006], even though they are not totally independent of perceptual features usually associated with spectral information. For example, Hartmann [Hartmann, 1978] showed that there is a relation between the temporal envelope of sinusoids and their pitch. The spectral sonic features, on the other hand, are calculated on each frame of a spectro-temporal representation of the source-filter model. We estimate the parameters of the model (i.e, the spectral representation of source and filter) for each temporal frame. Then the spectral sonic features (spectral centroid, spread, skewness and kurtosis) are calculated on each frame of the source-filter representation. The spectral sonic features are a measure of the spectral shape, and as such, they are related to dimensions of timbre perception that depend on the distribution of spectral energy. Slawson [Slawson, 1985] refers to these dimensions as sound color and associates them with the filter representation.

Chapter 5 explains the relation between the temporal and spectral features adopted in this work (to evaluate the results) and musical instrument sound perception. We will pay special attention to the association of hybrid musical instruments and dimensions of timbre spaces unveiled in psychacoustic (perceptual) experiments.

1.2.1.4 General Description of the Morphing Algorithm Developed

Here I will briefly present how to manipulate the parameters of the model presented above to achieve perceptually meaningful transformations of isolated musical instrument sounds across timbre dimensions. Stationary morphing of musical instrument sounds requires attention to both temporal and spectral features of sounds. Due to the dynamic nature of musical instrument sounds, the temporal aspects can be global or local. Global aspects, such as sharpness of attack or overall duration, only need one value to describe them. Local aspects, on the other hand, are usually related to spectro-temporal changes, so we need a succession of values to capture the rate of change along the duration of the sound. One simple example is a piano note heard backwards. Even though the spectral characteristics of the sound are the same, the perception is different because they are heard in a different order.

Any stationary morphing algorithm should deal with the temporal and spectral characteristics of sounds. In this work, there are two main procedures applied to achieve a perceptually meaningful morphed sound, namely, temporal alignment followed by spectral morphing.

1.2.1.5 Temporal Alignment

The temporal alignment procedure uses global temporal information to align the local spectral variations of both sounds being morphed. Figure 1.6 illustrates the concept of temporal alignment, supposing that the sounds being morphed can be segmented into three perceptually different regions, the attack, the sustain and the release. Only a very naive morphing algorithm would combine the original sounds shown in figure 1.6 regardless of duration because one of them is longer than the other, as highlighted in the figure. One possible solution to this problem would be to time-stretch the shorter sound (or equivalently compress the longer one) only taking the total duration into account. The result would be somewhat similar to the simple temporal alignment shown in the figure, where we combine similar regions together, but also cross regions as highlighted in the figure. The result of cross combinations is a less perceptually convincing or less meaningful morph. Ideally, when stationary-morphing isolated musical instrument sounds, the desired temporal alignment is the one shown in figure 1.6, which allows the spectral combination of similar regions together,

resulting in a more seamless morph. Naturally, the temporal alignment procedure relies on the previous segmentation of the sounds. Chapter 8 presents an automatic temporal segmentation algorithm developed specially for this work [Caetano and Rodet, 2010a]. Chapter 12 deals with the consequent temporal alignment of isolated acoustic musical instrument sounds using the automatic segmentation method.

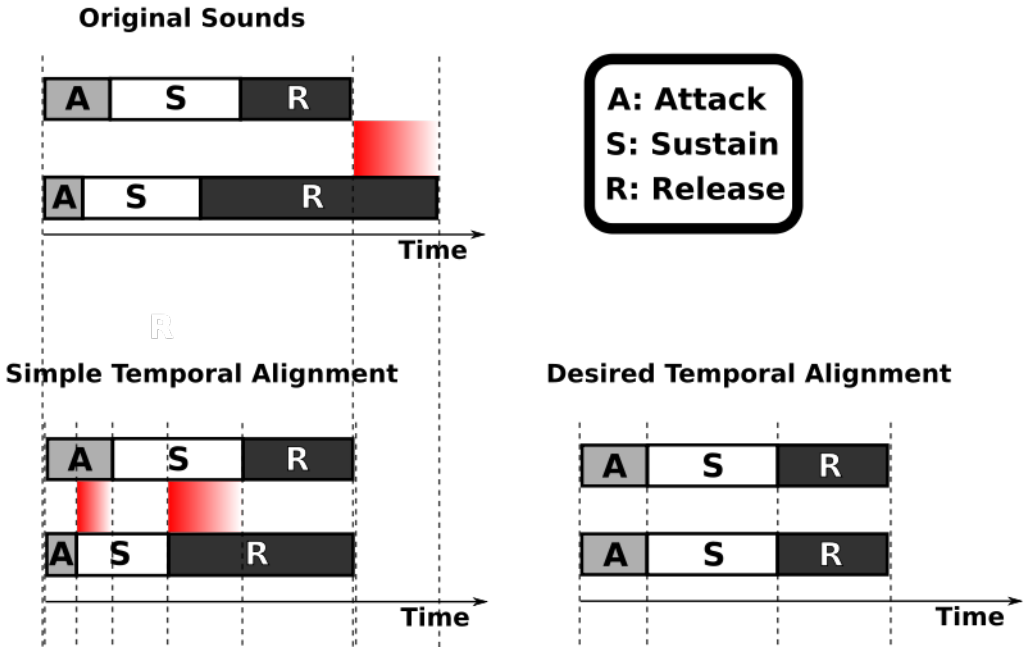

Figure 1.6: Schematic representation of the temporal alignment.

1.2.1.6 Spectral Morphing

After the sounds are properly aligned in time, we proceed with the spectral morphing procedure. In the morphing algorithm developed in this work, there is perfect correspondence between frames after the temporal alignment, as shown in figure 1.7. Correspondence is one of the three main requirements in the evaluation criteria adopted. Naturally, the spectral morphing procedure also requires correspondence. Figure 1.8 illustrates the lack of spectral correspondence by showing two spectra that have a different number of characteristics. The spectra shown in figure 1.8 have a different number of spectral peaks and a different number of partials, such that we cannot easily establish correspondence between them either in terms of spectral peaks or partials. The spectral correspondence depends on the parametric description of the spectra. That is, we can choose to describe the spectra in terms of spectral peaks, partials, spectral envelope curves, among other possibilities. Each parametric representation will present advantages and disadvantages depending on the problem at hand.

As we will see in chapter 3, when the spectral representation we use is the classical sinusoidal model, correspondence between partials becomes easily an issue. This is partially the reason why the source-filter model is used in this work instead. I will show in chapter 13 that representing source and filter separately brings back the necessary spectral correspondence, and also allows

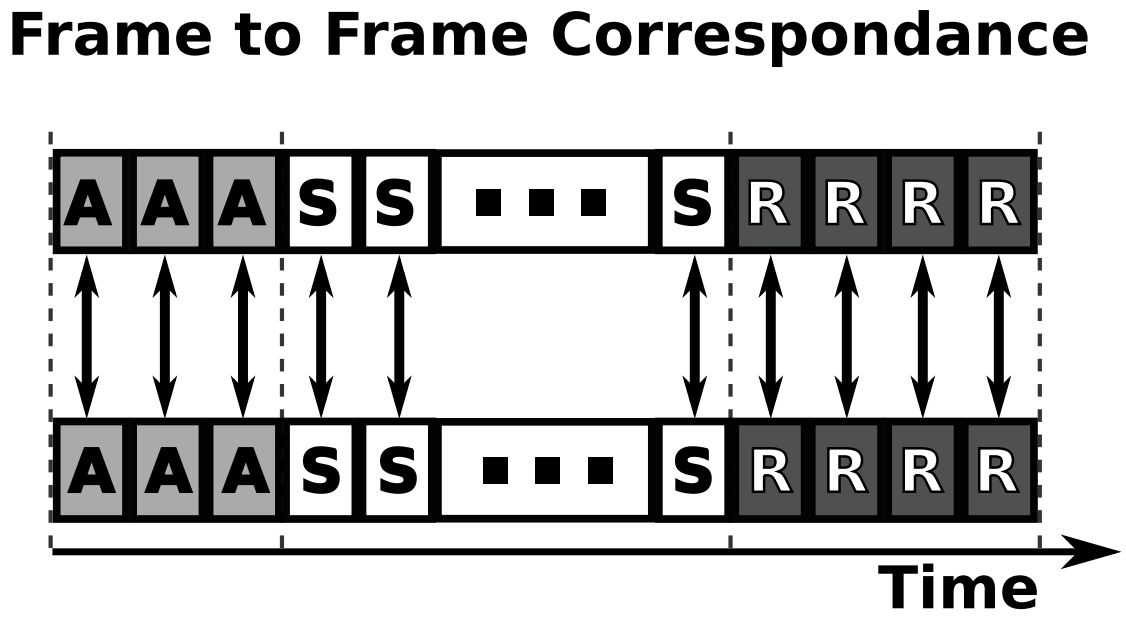

Figure 1.7: Frame to frame correspondence after temporal alignment.

us to tackle some other perceptually related issues that come up when morphing the spectral representation of sounds. The spectral envelope morphing techniques represent a major part of the work developed, and as such will be the focus of the evaluation procedure. The use of the spectral shape features as objective measure led to the adoption of a minimum quadratic error deviation to choose the most appropriate spectral envelope representation when morphing musical instrument sounds.

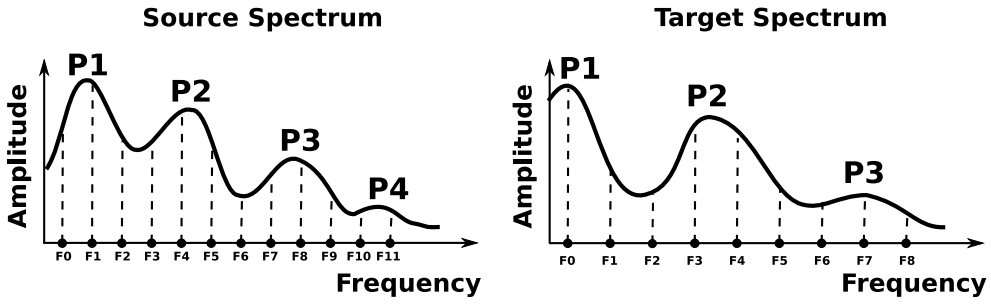

Figure 1.8: Spectral correspondence. The figure shows two spectra without correspondence because they have different numbers of spectral peaks (represented by P) and different numbers of partials (represented by F).

1.2.1.7 Temporal Envelope Morphing

Finally, the temporal envelope is an important aspect in musical instrument sound perception, and as such, should be treated and morphed separately. In the morphing algorithm presented in this work, the temporal envelope, which modulates the frames of the source-filter representation of the sounds, plays a fundamental role in the morphed result. Chapter 9 presents a temporal envelope estimation technique based on true envelope cepstral smoothing developed in the context of this work [Caetano and Rodet, 2011a] and compares it to traditional temporal envelope estimation techniques.

1.2.2 Contributions

One of the main contributions of this work is the development of a model specifically devoted to morph musical instrument sounds. Perceptually relevant features of musical instrument sounds are modeled independently such that we can manipulate only one of them without changing the others.

Another major contribution of this work is the development of specific techniques that allow us to more intuitively control the features associated with each part of the model when morphing both the temporal and spectral representation of sounds. The features are used as objective measures to evaluate the results according to three evaluation criteria adopted for morphed sounds, correspondence, intermediateness, and smoothness. In order to present our model and justify the need to use the features and the decisions I made during the development of the model, I will present the theoretical aspects of morphing and its applications in image morphing and sound morphing. This is another important contribution of this work, the formalization of concepts related to morphing and their application to the specific problem of sound morphing. The music signal processing community can largely benefit from a solid theoretical background and formalization of sound morphing because nowadays the concepts are fuzzy and knowledge is scattered. For example, the terminology is not consistent and the definitions are vague or imprecise. Yet another important aspect of this document is the extensive bibliographical research presented.

The specific problem I set to tackle raised a long list of theoretical and technical issues about sound morphing, which I will address accordingly in this text. Most of the questions are strictly related to the lack of formalism in the literature. One of the main goals of the next chapter is therefore to present a theoretical and mathematical formalization of morphing that will form the basis of the work described and that offers the possibility to propose solutions to the many problems that arise in the context of morphing isolated quasi-harmonic acoustic musical instrument sounds.

The main contributions of this work are

- the formalization of morphing objects in general, from a theoretical, mathematical and algorithmic point of view;
- the formalization of sound morphing, explicitly listing the known sound transformations that can be considered morphing and those that cannot according to the theoretical formalization above;
- an extensive review of the literature on sound morphing, describing the different techniques and goals. An important side-effect of this review is the proposal of a homogeneous nomenclature;
- a signal-processing formalization of source-filter model for musical instruments. The sinusoidal source is modeled as a set of time-varying sinusoidal partials and the filter as the short-time spectral envelope;

- the development of a temporal envelope estimation technique;
- the proposal of a perceptually inspired automatic temporal segmentation technique, which lead to the temporal alignment algorithm used in this work;
- the adoption of evaluation criteria for morphing and the consequent development of objective measures for morphed musical instrument sounds;
- the study of the perceptual impact of several popular spectral envelope morphing techniques as measured by the spectral shape features;
- the derivation of an analytic correspondence between the cepstral coefficients and the spectral shape features;

1.3 Overview of the Thesis

The text is divided in three parts. Part I deals with the theoretical and practical problems we encounter when morphing, sounds or otherwise. In part II we will examine the sound model developed in the context of this work to morph musical instrument sounds. Finally, part III explains the morphing techniques applied to different parts of the model, together with the evaluation of the results. The document finishes with the conclusions and future perspectives of this work.

The next chapter introduces the general problem of morphing objects from a theoretical point of view with the aim of presenting all the conceptual issues raised in the problem of morphing objects theoretically. Chapter 2 makes extensive use of the image morphing analogy to exemplify the many conceptual and theoretical questions we are faced with when morphing. Chapter 3 after that narrows it down to sound morphing, focusing mainly on the specific problems we encounter when morphing sounds, more specifically musical instrument sounds. In chapter 3 we will formalize the requirements of morphing, the different types of transformation that meet these requirements, and we will also adopt criteria to evaluate the result of a sound morphing technique from a technical point of view. Chapter 4 presents a thorough review of the literature of sound morphing, scrutinizing each model and algorithm applied in the problem. One of the most popular sound models used to morph sounds, sinusoidal modeling, is introduced here in detail, as well as the general morphing technique when we choose to use it. To conclude part I, chapter 5 focuses on morphing musical instrument sounds across timbre dimensions. In chapter 5 we will review timbre perception with a focus on timbre spaces and their acoustic correlates, that is, features that can be calculated from representations of sounds that are correlated with sound perception. These features will be used later to evaluate the results from an objective point of view.

Part II begins with the source-filter model from a rather theoretical point of view. In chapter 6 we will see how the source-filter model can be applied to musical instruments, and we will also explore the relationship between parameters of the model and musical instrument sound production as well as perception. The next chapters in this part present theoretical aspects of the estimations of different parts of the source-filter model., Chapter 7 is dedicated to the estimation of spectral envelopes, reviewing the major spectral envelope estimation techniques in the literature, those based on linear prediction and cepstral smoothing. The aim of this chapter is to subsidize the theoretical foundations necessary to justify the application of different techniques in the estimation and representation of spectral envelopes in this work. Next, chapter 8 presents different models used to segment musical instrument sounds into salient perceptual regions, such as attack and steady state. This is the basis of the automatic temporal segmentation technique developed in the context of this work. Chapter 9 reviews the major temporal envelope estimation techniques in the literature, and then presents the temporal envelope estimation method proposed in this work.

Part III starts with chapter 10, in which the Vienna sound database is briefly presented, and the musical instrument sounds used throughout the rest of the text are introduced. Then, an extensive overview of the morphing method is covered in chapter 11, where the implementation of the source-filter model is explained, together with the result of the estimation techniques applied to each part of the model. Chapter 11 also explains the morphing algorithm developed in this work in detail, showing the temporal and spectral morphing techniques step by step. This is followed by a description of the morphing techniques used in this work. Chapter 12 explains the temporal alignment of perceptually salient regions of musical instrument sounds, followed by chapter 13, where the spectral morphing techniques are presented. Chapter 13 is very important in the context of this work because of the strong focus of spectral techniques when morphing musical instrument sounds across timbre dimensions. The evaluation presented in chapter 14 is intrinsically connected with chapter 13. Chapter 14 proposes to evaluate the results using the values of the features as an objective measure, and presents a listening test to cross validate the results of the objective evaluation.

Finally, chapter 15 concludes the document and presents some remarks for people interested in pursuing the extremely difficult but highly rewarding task of morphing musical instrument sounds.

Part I

On Sound Morphing

Chapter 2

Morphing

This chapter is entirely dedicated to the concept of morphing objects in general. In this chapter, these conceptual objects can be geometrical, graphic or auditory in nature. The main idea is to shed light on the pitfalls we face when morphing sounds from a simpler viewpoint. The first important question that will be addressed is about what morphing is. The relationship between hybridization and morphing will be discussed. Then, some theoretical considerations on morphing will be presented, focusing on how they affect the procedure and the results of morphing.

The requirements of morphing, namely, correspondence, intermediateness, and smoothness, will be discussed. Among these, intermediateness is at the core of the standard approach to morphing objects, called the interpolation principle, which uses a convex combination to perform the morph, Smoothness, on the other hand, is what we expect from a gradual transition. The nature of the perception of the objects being morphed, continuous or categorical, determines the smoothness of the results and it is an intrinsically perceptual issue.

The way the subjectiveness in evaluation and the conceptual distance between the objects affect the result will be addressed. The criteria used when evaluating the results of morphing usually depend on the purpose or application. A general purpose morphing algorithm will be presented in this chapter, along with different types of transformation that respect the requirements of morphing discussed here.

An interesting analogy, image morphing, will be used to explore the concepts and problems attached to morphing. At the very beginning of this chapter, the focus will be on morphing objects and how their shapes are transformed. More complex scenarios will be gradually introduced along the chapter to pave the way to the central question addressed in this chapter, morphing by feature interpolation. The next chapter will apply the ideas developed here to morphing sounds.

2.1 What is Morphing?

The word morph comes from the Greek μορφ / morphe, which means form, and is associated with different scientific domains, ranging from biology to linguistics. The Merriam-Webster Dictionary lists three entries for 'morph', a noun, a verb, and an abbreviation.

- Main Entry: 1 'morph', Function: noun, Etymology: back-formation from morpheme Date: 1947. 1 a : allomorph b : a distinctive collocation of phones (as a portmanteau form) that serves as the realization of more than one morpheme in a context (as the French du for the sequence of de and le) 2 a : a local population of a species that consists of interbreeding organisms and is distinguishable from other populations by morphology or behavior though capable of interbreeding with them b : a phenotypic variant of a species

- Main Entry: 2 'morph', Function: verb, Etymology: short for metamorphose Date: 1982 transitive verb : to change the form or character of : transform. intransitive verb : to undergo transformation; especially : to undergo transformation from an image of one object into that of another especially by means of computer-generated animation.
- Main Entry: 3 'morph', Function: abbreviation morphology.

We are interested in the second entry, which has the general meaning of to transform or undergo transformation. It is interesting to notice that this definition makes no reference to sound at all. Also, according to the dictionary, it is especially used to refer to the transformation of images of objects, revealing that morph usually refers to a transformation of shape, as illustrated in figure 2.1. Finally, we should notice that the definition emphasizes the use of the digital computer to perform the transformation.

Figure 2.1: Depiction of object morphing. The figure shows two objects with intermediate shapes between the cube and the pyramid.

First of all, let us introduce the nomenclature adopted in this document. The objects being combined to produce the morph will be referred to as base objects. The result of the morphing will be dubbed morphed object and we say it occupies an intermediate point in space between the base objects or, alternatively, it has an intermediate shape. We should notice that sometimes there is more than one possible way of defining intermediate (or morphed) shapes between two. Figure 2.2 illustrates this important concept. What is the intermediate shape between the circle and the square? When we imagine the gradual transformation of shape between these two geometrical forms, the polygon with more and more sides seems to be as good an intermediate shape as the square with rounded corners. The "best" or "most appropriate" hybrid shape is application-dependent when we use objective criteria to evaluate the transformation or user-dependent when we use subjective criteria, that is, a user's personal taste or aesthetics.

Figure 2.3 depicts a flowchart with the two major applications of morphing, followed by the evaluation type and criteria proposed in this work, and possible measures used for them. When the purpose of morphing is artistic, the evaluation of the results is usually subjective. This means that each individual will evaluate the results differently, according to their own aesthetic criteria. In practice, we use the opinion of one single expert to judge the quality of the results instead of measuring the mean of several evaluations by different people because in this case the standard deviation would be very large. When we want to obtain morphed results for technical applications, such as image or sound synthesizers or to study perceptual aspects of morphing (such as continuous spaces), we need to establish an objective way of measuring the quality of the results. In figure 2.3 we see three criteria originally proposed by Osaka [Osaka, 1998]. Each criterion will be explored in this chapter as a motivation to the challenges we face when morphing. Let us consider other important issues that affect the quality of the final result.

When we need to specify the direction of transformation, instead of base objects we will talk about source and target objects. This is usually the case when the transformation happens dynamically, changing in time from the source to the target object. In this case a trajectory specifies the course of the morph. Dynamic transformations between static objects, such as images, give rise to morphed objects that possess an added temporal dimension, such as movies, where each frame corresponds to a different intermediate position in the trajectory or equivalently a different static

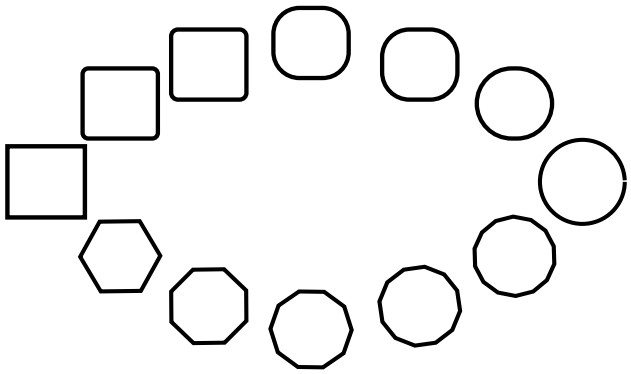

Figure 2.2: Depiction of the problem of interpolation of shape. The figure illustrates two different possible paths that contain shapes that could be considered intermediate between the square and the circle.

morphed object. The result of different transformations between different types of objects will be examined in detail from a theoretical point of view in this chapter, and later in chapter 3 for the specific case of morphing sounds. Naturally, for sound objects we need to specify what we mean when we talk about sound shape. Section 3.1.2 briefly presents Dennis Smalley's formalization of sound shape [Smalley, 1997], which is heavily influenced by Pierre Schaffer's concept of sound object [Schaeffer, 1966].

Another important aspect to be taken into account is the features that influence the transformation. When we are only considering shape, both transformations shown in figure 2.2 satisfy intuitively the requirement of a gradual transformation. But we can always consider simultaneously more than one such feature of the objects being transformed. For the objects depicted in figure 2.1, instead of only considering the shape, we could also include color, for instance. Now we have two independent dimensions, such that we have more than one possible transformation. Consider, for example, objects that inherit the color from one of the base objects, and the shape from the other. But inheriting features such as color and shape from either "parent" or base object is not the only possible hybridization process. We can also imagine morphed objects whose color and texture are combinations of the corresponding features of the base (original) objects being combined.

According to the above, morphing can be viewed as a transformation that involves hybridization of form or even other features. The term hybridization is applied in many areas generally to refer to a process that involves the combination of two (or more) objects, individuals, varieties, etc, depending on what it is being considered. A first important aspect of the problem of hybridization is to understand that there are several possible ways of combining two things. We are going to consider two simplified hybridization processes, one commonly found in nature and the other one usually only accessible by artificial means, in order to specify the transformation we refer to when we use the term morphing in this text.

2.1.1 Natural Hybridization

The hybridization process called sexual repreduction is ubiquitous in nature. In general terms, when a couple has children, we can usually easily recognize who the parents are because of physiological similarities. In other words, the kids take after their parents, and we usually say that

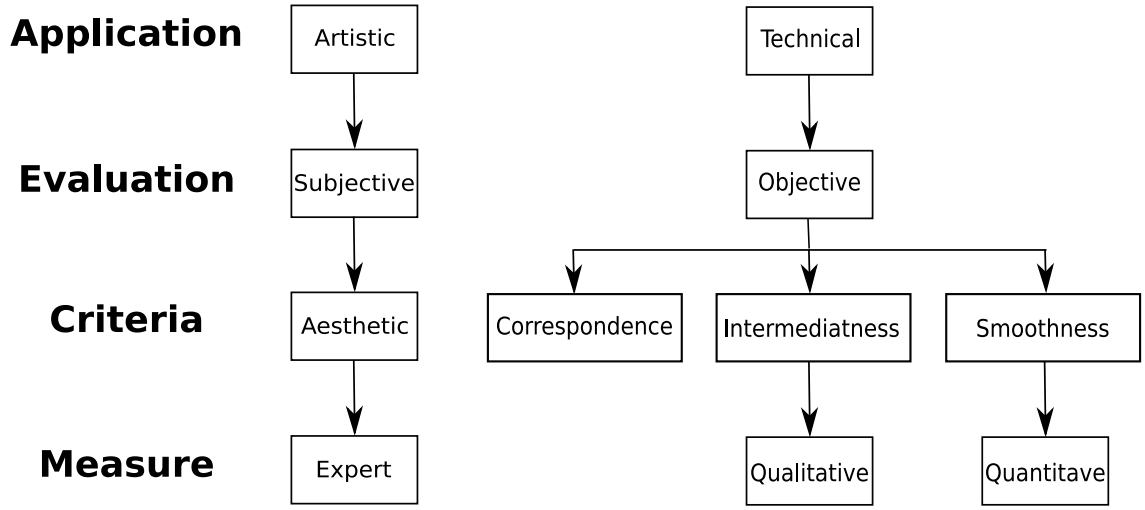

Figure 2.3: Depiction of two possible applications of morphing and the corresponding evaluation criteria used to evaluate the results.

they might have their mother's nose, the father's eyes, etc. This is a specific case of hybridization where the hybrid individual (the child) consists of a combination of parts from either parent (and eventually parts that are a combination of the corresponding original versions from both parents, but we will not consider this case at this point for clarity's sake.) We could identify how close to one of the parents the child is by a simple ratio r , defined as $r_M = \frac{\Pi_M}{N}$, where Π_M represents the number of parts that resemble the mother's (hence subscript M, the father would be referred to as F) and N represents the total number of parts. Notice that this ratio depends on which parent we use as reference, such that $r_M = 0.7$ means that the child resembles more the mother because subscript M stands for parts that the child "inherited" from the mother. Another way of calculating the same information would be to express P and Π as the inner product of vectors. Now $N = \alpha P = [1, 1, \cdots, 1]\,[p_1, p_2, \cdots, p_N]^T$, where$N$ is the total number of parts, α is a membership vector, and P is a vector listing all the parts taken into consideration. We express Π_M or Π_F by means of vector P and a membership vector $\alpha_{M,F}$ that contains zeros and ones, following the notation used in classical set theory, where one means that the element belongs to a group and zero means it does not. Hence, we can express the number of parts inherited from one parent, say the mother, as $\Pi_M = \alpha_M P = [1, 0, \cdots, 1]\,[p_1, p_2, \cdots, p_N]^T$. Notice that vector P defines the hybridization process by specifying which specific part we are considering, but the hybridization ratio r depends uniquely on the membership vectors α. We can redefine the hybridization rate as $r_{M,F} = \frac{\sum \alpha_{M,F}}{\sum \alpha} = \frac{\sum \alpha_{M,F}}{N}$.

For a geometrical interpretation of the natural hybridization process, we go back to figure 2.4. Now each part can be seen as a feature represented as an independent dimension in an abstract feature space. Suppose that the original parents (base objects) are the red square and blue circle. The natural hybridization process can only lead to the red circle (B2) and blue square (B1), apart from the original base objects.

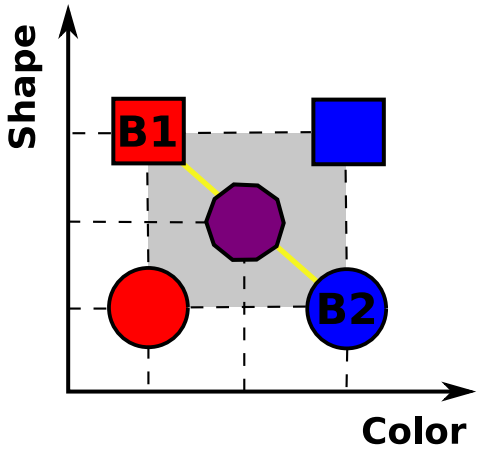

Figure 2.4: Geometrical representation of morphing. The figure illustrates two objects in a feature space where the axes represent features. The position of the objects in the space depends on the values of each feature individually. A hybrid object in this space should have values of features that are a combination of the original base objects, marked B1 and B2.

2.1.2 Artificial Hybridization

Following our analogy with the shapes and colors of objects shown in Figure 2.4, another possible hybrid much more interesting for us inherits body parts that are a combination of the corresponding parts from both parents. In other words, the eyes are a combination of father's and mother's, the mouth, etc. This special hybridization process is illustrated in Figure 2.6 and is usually only possible to implement via artificial means such as the digital computer. Here we should notice that the notion of ratio of parts to describe the hybrid does not apply anymore because each part comes neither from the mother nor from the father. Actually, the parts found in the hybrid individual did not exist previously. New parts are created (by means of a special hybridization process we will later define as morphing) combining the corresponding parts from the parents.

Now we have to redefine a way of measuring the degree of similarity between the offspring and either parent, hence we introduce the concept of degree of membership from fuzzy set theory. Now the membership vector α becomes a real-valued membership function in the interval $[0, 1]$ to express the gradual assessment of elements in a set. In the context of hybridization, α expresses the degree of combination for each part. For example, $\alpha_M = [0.25, 1, 0.5, 0.75]$ means that the first part, say the mouth, is the result of a combination of 25% the mother's mouth and 75% the father's, etc. Since we can specify a different degree of combination for each part separately, we can say that the number of parts is the number of degrees of freedom. Notice that, even though the global appearance of the hybrid individual can be fairly complex because of the number of degrees of freedom, we expect the appearance of each individual part to respect the degree of combination specified in α.

A geometrical interpretation of the artificial hybridization process can help us gain insight into this type of transformation. If we consider figure 2.4, this procedure generates hybrids inside the shaded region.

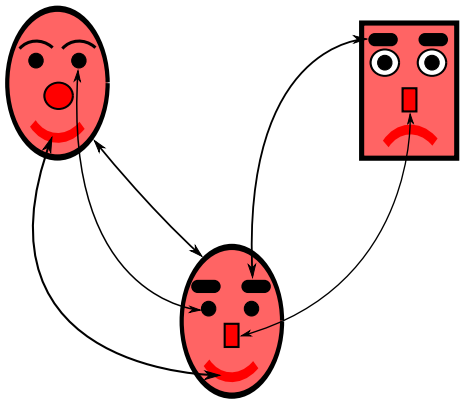

Figure 2.5: Depiction of face hybridization. The figure illustrates the hybridization case when each attribute is inherited either from the mother or from the father.

2.1.2.1 Morphing

An interesting special case of the artificial hybridization process occurs when all elements of the membership function α are numerically equal, reducing it to a scalar constant. In this case, one scalar value and one reference individual/object are enough to describe this specific hybridization process, referred to as morphing throughout this text. This procedure will be formalized later, but we can always gain some insight now by looking back at figure 2.4 and realizing that this type of combination can only lead to hybrids along the solid line.

Next, we will see some theoretical considerations about the general problem of hybridization, the specific procedure that leads to the transformation called morphing in this thesis, and how to evaluate the results from a technical point of view. The evaluation uses the three conditions that should be respected to produce a successful morph, inspired by the formalization introduced by Osaka [Osaka, 1998].

2.2 Theoretical Considerations

Motivated by the geometrical interpretation of morphing, we will see some theoretical considerations using image morphing as examples to illustrate the myriad possibilities and potential pitfalls of sound morphing. One drawback of this approach is that images are static and as such they fail to capture the intrinsic temporal nature of sounds. A closer analogy would be movie morphing. But we have a lot of ground to cover until we get there, so we will use the simpler image morphing analogy for now to make the more elementary concepts easier to grasp.

2.2.1 Subjectiveness

Conceptually, the main difficulty in morphing is probably the fact that we are usually looking for a result that only exists as an abstraction. In other words, even though we can compose known objects as combinations of simpler base objects, like illustrated in figure 2.1 following the mathematical formalization by Alexa [Alexa and Müller, 1999], the most interesting use of morphing is to obtain new objects that were previously intangible, only accessible to our imagination. When morphing basic forms, such as those shown in figure 2.1, it is difficult to imagine other hybrid

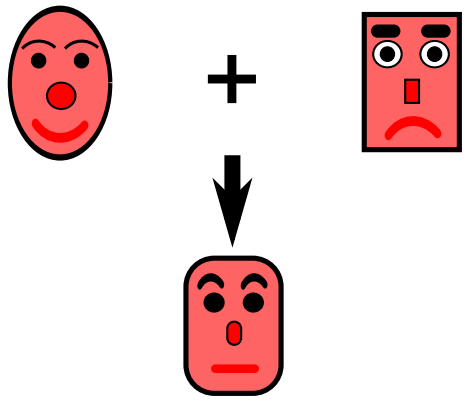

Figure 2.6: Depiction of face morphing as a special case of face hybridization. The figure illustrates a face whose constituent parts (nose, mouth, etc) are a combination of the corresponding parts from both parents.

objects with intermediate shapes between the cube and the pyramid. But even for basic shapes we might encounter cases where there is more than one alternative transformation between two given shapes. A very simple example can be seen in figure 2.2, what is the "best" intermediate shape between the square and the circle?

As we go up the complexity of forms and shapes, adding texture, colors, and other attributes, a world of possibilities opens up before us and the true creative potential of morphing becomes clear. The creative potential of morphing will be illustrated with a few graphic examples in the next paragraphs, notably face morphs. Still, when we want to explore the morphing possibilities between two objects with complex features (such as shape, texture, etc), we will eventually be faced with questions such as: "What is the result of the morph between a tiger and a car?"

But we are truly interested in morphing sounds, which can be even more complex and posses very abstract qualities, so the result of sound morphing can very easily be different depending on who imagines it. For example, we could try to imagine what the result of a morph between a dog bark and a trumpet note would sound like, but since this abstraction is often subjective, it becomes difficult to objectively evaluate the results. In other words, each individual has their own idea about how such transformation should be performed and especially about what the result should be like. This is fine when we are morphing for artistic purposes, when the artist creates the morph using their own aesthetic criteria.

However, in a more technical context, this raises the question of how to objectively evaluate the morphed sounds. What are the qualities that we expect to find in a good morph? Listeners are likely to be disappointed and give a low score when assessing morphed sounds simply because they do not meet their expectations, independent of the quality of the transformation. Therefore, this thesis proposes an objective evaluation procedure based on the three independent criteria shown in figure 2.3 along with the formalization of morphing. Let us examine each objective evaluation criterion in turn.

2.2.2 Correspondence

The first obstacle we face when morphing between any two objects is correspondence. In its most general formulation, morphing from one object to the other (be they graphic or sound objects)

requires two basic steps: a description of the objects followed by the process of establishing correspondences between these descriptions, as depicted in figure 2.7. According to this viewpoint, the morphed objects correspond to descriptions whose elements are intermediate between the objects being combined.

If there are elements in one of the objects that do not have a correspondence in the other(s), the transformation is not straightforward. One example for faces can be seen in figure 2.5. If one of the faces has a feature that does not have a corresponding one on the other (for example one of the faces has a mole), we will have to decide how we are going to represent intermediate versions of it. One of the consequences of the lack of correspondence betwen the objects being morphed is that we can have multiple possible transformations depending on how we decide to deal with the free feature. Looking back at figure 2.2, now we can interpret the existence of more than one possible transformation as a consequence of the lack of one to one correspondence between the objects.

Usually, we perform the transformation using a map between the objects. In algebra, structure-preserving maps are called homomorphisms. When the map admits an inverse (such that we obtain identity when applying the map and its inverse in succession), it is called an isomorphism. Informally, an isomorphism shows a relationship between two properties or operations. If there exists an isomorphism between two structures, we call the two structures isomorphic. In a certain sense, isomorphic structures are structurally similar. Therefore, when morphing between two objects, isomorphism should be one of the requirements. This is exactly what Tellman [Tellman et al., 1995] means by "equal number of features" in his seminal article about sound morphing. We may establish correspondence between parameters or features associated with the objects.

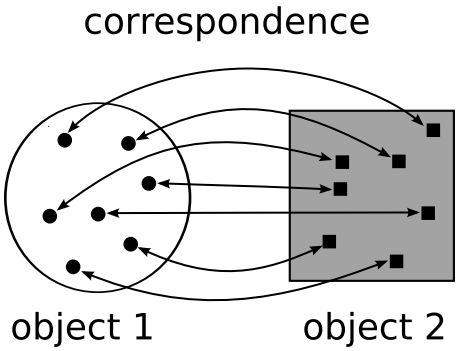

Figure 2.7: Depiction of the fundamental steps of morphing. The figure shows two sets of elements extracted from the objects we want to morph between and the correspondence between these elements. The mapping with which the transformation is performed depends on the correspondence.

2.2.3 Intermediateness

The morphed objects should be perceived as intermediate between the base objects used. For example, when transforming between a square and a circle we want to avoid transforming the square into another recognizable shape first (say, a triangle) that is not perceived as intermediate between the square and the circle and then finishing the transformation from this shape into the circle. Intuitively, when transforming between a child's and a man's face like shown in figure 2.8, we expect all all the hybrids to be human faces because it would be counterintuitive otherwise.

Formally, the principle of intermediateness states that the morph must change along dimensions that represent differences, preserving the position in all the others. Thus, conceptually, the transformation between a child's and a man's face should be the face of a person getting older. In other words, the transformation should happen along the conceptual dimension of age only, because man and child are conceptually apart only regarding age.

At this point, we should notice that it is always possible to perform the transformation via another object. For example, we could gradually transform both the child's and the man's face into, say, a soccer ball. Even though artistically this remains a valid choice, this kind of transformation is not considered to be proper morphing in this work because it does not respect the principle of intermediateness we adopted.

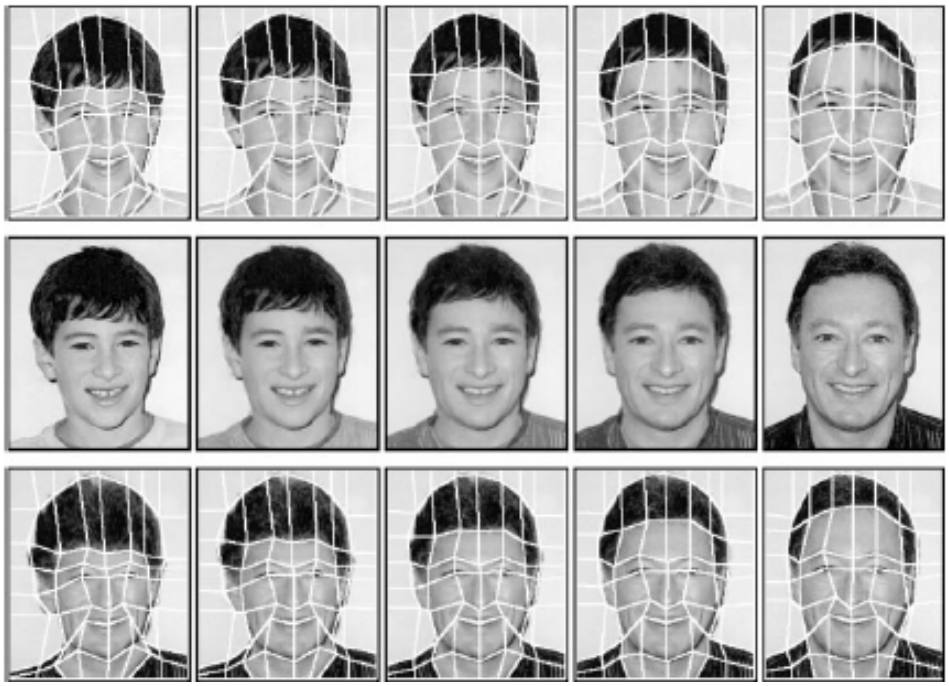

Figure 2.8: Depiction of image morphing to illustrate the correspondence between the two objects that facilitates the warping from one form to the other. After Wolberg [Wolberg, 1998]

When we represent the objects to be morphed as points in a Hilbert space, the requirement of intermediateness is equivalent to a straight line connecting the points, not another trajectory figure 2.9 illustrates this concept. When we go from point A (child's face) to point C (man's face) we want to pass through point B (teenager's face), not D (soccer ball).

So we see that having the same distance from both base objects is necessary but not sufficient to satisfy intermediateness. The condition of intermediateness requires that the points lie along the segment $\overline{AC}$, such that the distances from A to B plus the distance from B to C be equal to the distance from A to C. Notice that in figure 2.9 point D lies at the same distance from points A and C, yet, it is not intermediate between them. This is equivalent of saying that the points must fail the triangle inequality to preserve intermediateness.

In practice, when morphing simple objects, images or even sounds, we make use of the interpolation principle to respect the principle of intermediateness. The interpolation principle used in morphing is actually a convex combination of the parameters of a description of objects (that

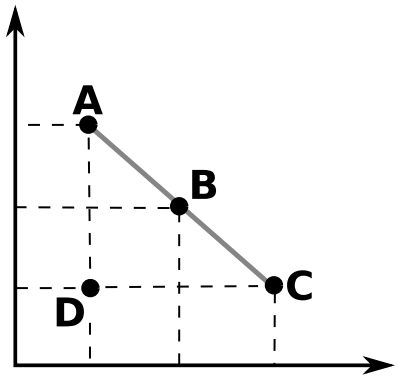

Figure 2.9: Depiction of the concept of intermediatness. The figure shows point B, intermediate to points A and C, and point D, which is not intermediate. The idea is that the points must be in the same line in Hilbert spaces to guarantee intermediateness.

usually comes from a model). A convex combination leads to intermediate representations in the space of parameters. Next we will expand the interpolation principle to include the morphing by feature interpolation concept.

2.2.3.1 The Interpolation Principle

The interpolation principle uses combinations of the descriptions of the base objects to obtain morphed results. The interpolation principle relies on the correspondence between the descriptions of the objects to be morphed, like shown in figure 2.7. In abstract terms, each element we obtain from the description of the base objects can be associated with a dimension of a mathematical space. Then, the number of elements determines the dimension of this space, and each base object occupies a point in this space, determined by the specific elements used to describe it. It follows naturally from this interpretation that, if we can describe two (or more) different objects as distinct points in this multidimensional space, we can always obtain new objects via a simple linear combination of the original objects. The positions the new objects occupy in space depend on the coefficients of the linear combination. We usually call morphing the result of convex combinations, which will be defined next. The term interpolation, in this case, refers to the operation of obtaining an intermediate object from two (or more) base objects.

2.2.3.2 Convex Combination

A convex combination is a linear combination of points (which can be vectors, scalars, or more generally points in an affine space) where all coefficients are non-negative and sum up to 1. All possible convex combinations will be within the convex hull of the given points, as illustrated in part a) of figure 2.10. Part a) of Figure 2.10 illustrates that point A can be described as the result of the convex combination of points $\boldsymbol{\sigma}_p, \boldsymbol{\sigma}_q, \boldsymbol{\sigma}_r$, while point B cannot. In fact, the collection of all such convex combinations of points in the set constitutes the set's convex hull, highlighted in part a) of Figure 2.10. More formally, given a finite number of points $\boldsymbol{\sigma}_p, \boldsymbol{\sigma}_q, \cdots, \boldsymbol{\sigma}_n$, in a real vector space, a convex combination of these points is a point of the form

$$\alpha_1 \boldsymbol{\sigma}_p + \alpha_2 \boldsymbol{\sigma}_q + ... + \alpha_n \boldsymbol{\sigma}_n \tag{2.1}$$

where the real numbers α_i satisfy $\forall\alpha_i \geq 0$ and $\alpha_1 + \alpha_2 + ... + \alpha_n = 1$. As a particular example, every convex combination of two points lies on the line segment between the points, as illustrated by the solid line in part b) of Figure 2.10. Alexa et al. [Alexa and Müller, 1999] extend the concept of morphing for more than two objects, interpreting morphing as a mechanism to describe an object in terms of a composite (convex combination) of other objects.

When morphing two objects, equation 2.1 reduces to

$$\alpha_1\boldsymbol{\sigma}_p + \alpha_2\boldsymbol{\sigma}_q = \alpha\boldsymbol{\sigma}_p + (1-\alpha)\,\boldsymbol{\sigma}_q = \boldsymbol{\sigma}_{p,q} \tag{2.2}$$

where α is a scalar called interpolation or morphing factor, and $\boldsymbol{\sigma}_n$ represents a vector of parameters. Notice that the condition $\alpha_1 + \alpha_2 = 1$ allows us to write the combination as a function of a single parameter α. Here we should make clear that the vectors $\boldsymbol{\sigma}_n$ belong to a vector space Σ that has dimension N, and can be described in terms of its vector components σ_n, which are scalar quantities, as follows

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \vdots \\ \sigma_N \end{bmatrix} \tag{2.3}$$

By the usual definition of the vector operations of vector addition and scalar multiplication, equation 2.2 can be rewritten as

$$\alpha\overline{\sigma}_p + (1-\alpha)\,\overline{\sigma}_q = \alpha \begin{bmatrix} \sigma_{p1} \\ \sigma_{p2} \\ \sigma_{p3} \\ \vdots \\ \sigma_{pN} \end{bmatrix} + (1-\alpha) \begin{bmatrix} \sigma_{q1} \\ \sigma_{q2} \\ \sigma_{q3} \\ \vdots \\ \sigma_{qN} \end{bmatrix} = \begin{bmatrix} \alpha\sigma_{p1} \\ \alpha\sigma_{p2} \\ \alpha\sigma_{p3} \\ \vdots \\ \alpha\sigma_{pN} \end{bmatrix} + \begin{bmatrix} (1-\alpha)\,\sigma_{q1} \\ (1-\alpha)\,\sigma_{q2} \\ (1-\alpha)\,\sigma_{q3} \\ \vdots \\ (1-\alpha)\,\sigma_{qN} \end{bmatrix} = \overline{\sigma}_{p,q} \tag{2.4}$$

At this point it should be clear that, in order to obtain a point in any of the paths represented by dashed lines in part b) of Figure 2.10, the coefficients α_j must also be vectors, represented as

$$\bar{\alpha}_j = \begin{matrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_N \end{matrix} .$$

The morphing factor α varies from 0 to 1, and defines a direction of transformation because, according to equation 2.2, $\alpha = 1$ gives $\boldsymbol{\sigma}_p$ and $\alpha = 0$ gives $\boldsymbol{\sigma}_q$. The value of the morphing factor α can be interpreted as the relative distance from the base objects, e.g. $\alpha = 0.5$ means exactly halfway between them.

2.2.4 Smoothness

Finally, we need to consider the smoothness of the transformation procedure, which is intrinsically tied to continuity. Ideally, we want to obtain morphed objects that change gradually from source to target. Since we control the transformation with the morphing factor α, continuously varying α should lead to a gradual transformation.

Adding the constraint of smoothness to the simple shape transformation depicted in figure 2.2 may help us solve the dilemma. Now we can decide which possibility produces a smoother,

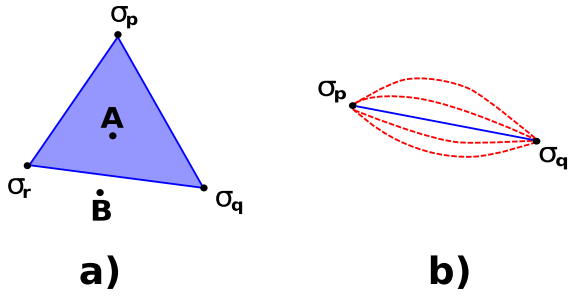

Figure 2.10: Convex combination. In part a) we see that, given three vectors $\bar{\sigma}_p$, $\bar{\sigma}_q$, $\bar{\sigma}_r$ in a plane, vector A is a convex combination of $\bar{\sigma}_p$, $\bar{\sigma}_q$, $\bar{\sigma}_r$, while B is not. Part b) shows that the convex combinations of points σ_p andσ_p define the solid line connecting them, while any other path depicted as a dashed line between points σ_p andσ_p can be obtained as linear combinations of their components that are not convex combinations.

more gradual transformation. Even though we should probably apply some quantitative measure, intuitively we can examine the smoothness of the transformations.

When we combine the constraints of intermediateness and smoothness we get linearity. Adding the same factor to the stimulus should increase the perception by the same amount. In the case of morphing, we expect a linear variation of the morphing factor α to produce the perception of a linear, gradual transformation. However, we must take into account the nature of the stimulus under study. Some stimuli are perceived continuously, others categorically due to their cognitive representation. When the perception is categorical, even a continuous variation of the stimulus leads to a discontinuous percept that changes very little inside a category and much more between categories, as will be clearer next.

2.2.4.1 Categorical Perception

The question of smoothness is related to a much more profound one that is intrinsically tied to categorical perception. Categorical perception means that a change in some variable (stimulus) along a continuum is not perceived as gradual, rather as instances of discrete categories, as shown in figure 2.11. In other words, discrimination between stimuli is much more accurate between categories than within them. Ideally, all stimuli in a given category should be perceived as indistinguishable, whereas stimuli from different categories, no matter how close on the continuum, should be perceived as different.

One simple example is color perception. The perception of color stems from the cognitive representation of different wavelengths (or equivalently frequencies) of light. Light, an electromagnetic wave, has a continuous range of frequencies, also called spectrum. Color, on the other hand, is merely a cognitive label associated with certain socially constructed ranges of frequencies, as represented in figure 2.12. Figure 2.12 shows the visible spectrum of light with linear variation of the values of frequency. Our brains interpret this information categorically; that is, even though the frequency varies continuously, our perception of colors is (more or less) separated into stripes labeled red, blue, etc.

The German mathematician David Hilbert (1862-1943) coined the term *spectrum*. In a lecture he delivered at the University of Göttingen in 1905, Hilbert considered linear operators acting on certain infinite-dimensional vector spaces, and it was in this context that Hilbert first used the term spectrum to mean a complete set of eigenvalues. Spectrum is a Latin word meaning "image".

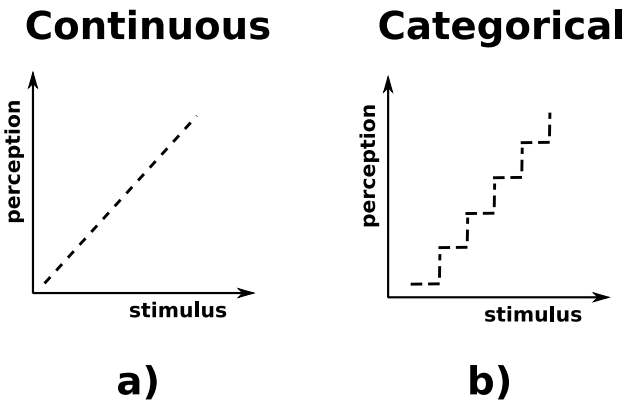

Figure 2.11: Continuous vs Categorical perception. The figure depicts an idealized view of the difference between continuous and categorical perception. In part a) the perception presents an ideally linear variation as the stimulus varies continuously. In part b), the perception presents an ideal categorical variation as the stimulus varies continuously because within certain ranges of values of the stimulus the perception does not change, whereas the perception presents discontinuities for other specific values (representing the transition between categories.)

When atoms vibrate, they emit light. And when light passes through a prism, it spreads out into a spectrum of light that is emitted from the prism. Thus we can literally see the eigenvalues of the atom in its spectrum, and for this reason, it is appropriate that the word spectrum has come to be applied to the set of all eigenvalues of a matrix (or operator). We will explore further the use of the term spectrum for sounds using this interesting analogy with images later in chapter 6.

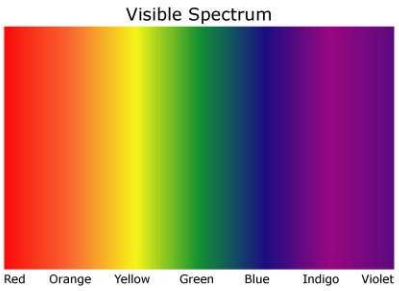

Figure 2.12: Visible spectrum of light. The figure illustrates categorical perception of colors by showing that a continuous variation of frequencies leads to a discrete perception of colors due to the cognitive representation of electromagnetic frequency.

The examples of Figures 2.8 and 2.13 imply that it might be possible to obtain a gradual transition between faces. Naturally, the relevant question in the context of this thesis is whether we can do the same between musical instrument sounds. The work developed and presented here would provide a means to study the categorical perception of musical instrument sounds.

Actually, the glue that holds the formalization of morphing together is the assumption that it

is possible to create objects that do not belong in the same category of either the base objects used in the morph. So a fundamental question this work raises is whether the perception of musical instrument sounds is categorical or continuous. If the answer is categorical that means we are trying to achieve the impossible task of obtaining a sequence of sounds that would be perceived as continuously changing from the sounds we combine. It is out of the scope of this work to examine categorical perception of musical instrument sounds. Nevertheless, chapter 14 presents the result of a listening test that compared the results of two sound morphing techniques aiming at determining which one produces more perceptually linear (or smoother) transitions.

2.2.5 Conceptual Distance

Finally, the conceptual distance between the objects we want to morph also plays an important role. Even if the two objects we want to morph between have the same number of elements, such that it is always possible to find correspondences between them, we are bound to encounter examples that will lead to artificial hybrids simply because the objects are conceptually very far from each other. When we compare the example of figure 2.8 with that of figure 2.13, both from Wolberg [Wolberg, 1998], it becomes clear that even though a cat's face and a man's face have the same elements (two ears, two eyes, a nose, etc), the hybrid images we obtain when morphing between them look less natural than between two human faces because of the conceptual distance between them.

As a general rule of thumb, the naturalness of the hybrid objects is inversely proportional to the conceptual distance between them. The farther apart the objects are in the conceptual space, the more challenging it is to obtain convincing hybrid objects between them. Naturally, when we are aware of this inverse relationship, we can always choose to use it to our advantage, that is, the consequence of the conceptual distance can be an explicit aesthetic choice.

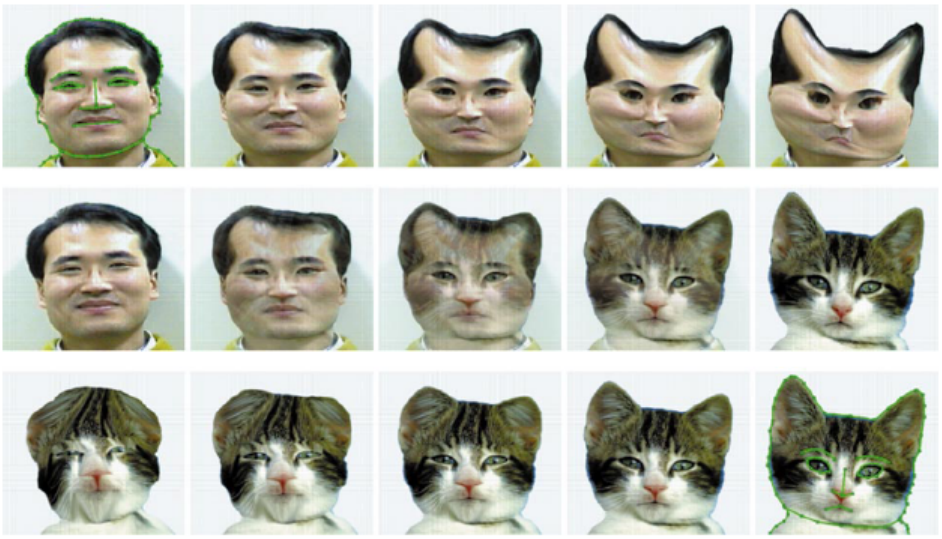

Figure 2.13: Depiction of image morphing to illustrate the effect of the conceptual distance between the objects on the morphed objects. After Wolberg [Wolberg, 1998]

Even though Wolberg's work [Wolberg, 1998] is restricted to morphing graphic objects (images), we can profit from the examples in figures 2.8 and 2.13 to learn more about morphing.

Wolberg states that image morphing involves coupling image warping with color interpolation [Wolberg, 1998]. Let us examine this statement closely with the aid of figures 2.8 and 2.13.

Figure 2.14 shows us that in order to establish the correspondence between the sets of elements extracted from the images, we must first and foremost examine both images to find the elements. Naturally, we should know what we are looking for. For faces, we want to identify salient features such as the eyes and mouth. This is what the grids in figure 2.8 represent. We need a model to identify the elements we want to represent. Again, for the faces a simple model would be a general description of the elements we expect to identify, such as points at the corners of the mouth, the eyes, the ears, etc. Notice that the relative distances between these points vary depending on the face and that provides a unique way of describing individual faces.

After adjusting the parameters of the model to the specific images we want to morph between, the next step is to warp the grids so that the points coincide and finally interpolate the colors. Warping means that points in the image are mapped to different points without changing the colors and it can be done mathematically by any function from (part of) the plane to the plane. Warping is an essential step in the morphing process because it guarantees that we will combine equivalent elements. For example, if we look at the middle column in figure 2.13, we see that both the man's and the cat's face were distorted to align points that correspond to the same elements, such as nose and mouth.

How do we calculate the final position of these points prior to warping? The most straightforward way of doing it is by interpolating the coordinates of the points. This gives rise to the classic morphing technique based on the interpolation principle. The idea behind the interpolation principle is that we should obtain a somewhat smooth transition between the objects if we interpolate (convex combination) the parameters of their representation. There is a huge assumption behind it that there exists a sequence of intermediate images (faces in this case) that will be perceived as a continuous transition between the two rather than categorically, i.e., with a sharp change of perception at the position of the parameter continuum where there is identity change.

2.2.6 Morphing Algorithm

Image morphing provides a graphic example of most of the considerations above. For example, we can easily see how correspondence between elements in the objects we combine affects the quality of the morph. It is easier to morph between two faces than between a face and a hand, for example, because the faces are isomorphic structures. Figure 2.8 illustrates the correspondence between two faces followed by the warping of one into the other. At the top of figure 2.8, we see the boy's face mapped by a grid that plays the role of description of the elements, and at the bottom of the same figure, we see the same for the man. After the description of the elements (that is, the grid) and the establishment of the correspondence between them (associating points and lines in both grids), the process of morphing from the boy's into the man's face uses an isomorphism that warps the grids (to obtain intermediate shapes) and then blends the colors (to obtain intermediate hues).

This thesis discusses the equivalent of this process for sound morphing. It also describes techniques to do the same for acoustic musical instrument sounds both theoretically and technically. The general morphing process can be described in four steps, as explicitly shown below as a general algorithm for morphing.

1. Analysis: analyze objects to be morphed (base objects) according to a model. This step fits the model parameters to describe the base objects;

2. Correspondence: establish correspondence between parameters of the model for the base objects. This should somehow reflect an intrinsic correspondence between elements (or features) of the objects;

3. Convex combination: obtain convex combination of values of corresponding parameters. For two objects this step becomes the interpolation between pairs of corresponding parameters;

4. Resynthesis: resynthesize an object from the values resulting from the convex combination of parameters of the model. The result of the morphing operation is called morphed (or hybrid) object.

2.2.7 Morphing Guided by Features

Inspection of figure 2.13 reveals that the process of warping and color interpolation might still produce artificial hybrid images depending on the source and target. Instead of simple color interpolation to blend the hues, we need to extract features such as skin and hair texture and combine them in an efficient manner. Feature extraction is a crucial step of the morphing process because of the potential impact on the result.

Figure 2.14 shows a striking example of image morphing to illustrate the impact that the hybrid images can have when they are perceptually convincing. Notice that, in this case, not only does the skin color change, but also its texture. The same can be said about the hair and the clothes. Examine how key features of the faces such as the eyes and nose shape change in the intermediate images. Note that even the position of the shoulders and the smile change gradually. Figure 2.14 emphasizes the impact of the model in the final result. A simple model leads to unnatural results. When we use a model that represents perceptually relevant features of the objects we want to morph between, the intermediate representations will correspond to convincingly natural hybrids.

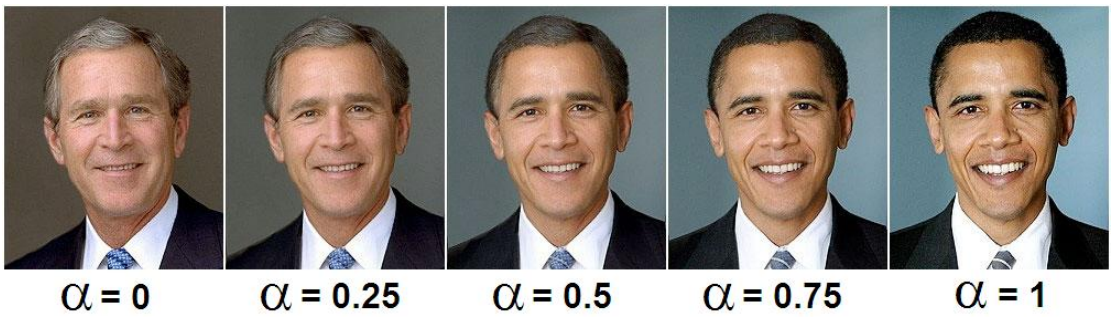

Figure 2.14: Another example of face morphing. In this case, features such as skin texture are morphed. Original Image from `http://i40.tinypic.com/11tqy52.jpg`

How do we attain a morph that resembles figure 2.14? The idea is fairly simple because it is a straightforward application of the morphing by feature interpolation principle. Instead of simply interpolating the parameters of the model we use to represent the base images, we are going to extract features from these parameters, interpolate the feature values, and retrieve the set of parameter values that correspond to the interpolated feature values. Each set of parameters $\boldsymbol{\sigma}$ has a corresponding set of features $\boldsymbol{\delta}$, as seen in figure 2.15.

Supposing that the features represent information that is more perceptually meaningful, such as skin texture or shape of the eyes for the faces, we need to interpolate the feature values in order to obtain hybrid images with intermediate features. There always exists a map φ from the parameter space Σ to the feature space Δ, which is simply how we calculate the features from the parameters. The question is "Is there an inverse map from the feature space Δ back to the parameter space Σ that permits retrieval of parameter values $\boldsymbol{\sigma}_p$ that correspond to certain feature values $\boldsymbol{\delta}_p$?"

When the features capture perceptually meaningful information, the answer is generally 'no'. Still, we are faced with the difficult task of retrieving a set of parameter values that correspond to the desired feature values. This problem is sometimes called feature-based synthesis and is notoriously difficult to solve. When morphing sounds using the morphing by feature interpolation principle, in each step we are faced with choices that might affect the quality of the final result, such as what model we use to extract the parameters, how we are going to represent these parameters, what features we are going to use to extract perceptually relevant information from the parameters, and a crucial step lies naturally in the retrieval of parameter values from feature values for resynthesis. The rest of this document is dedicated to describe a model and associated techniques developed in this thesis to solve this problem.

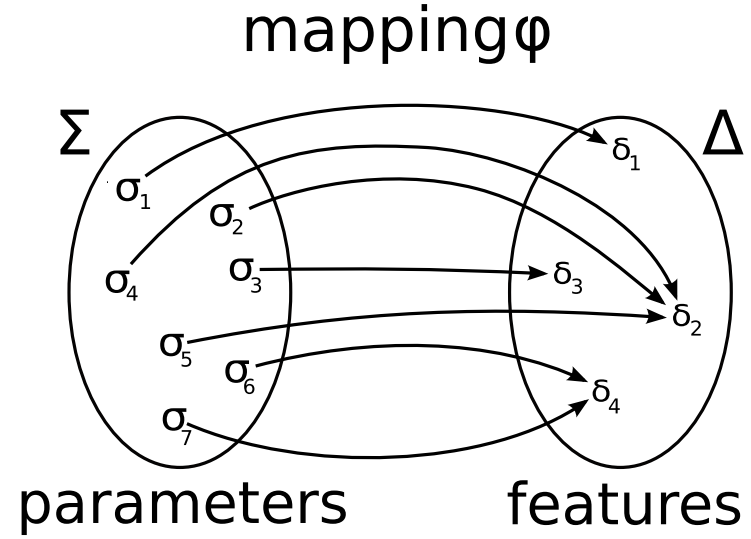

Figure 2.15: Mathematical interpretation of the morphing by feature interpolation principle.

2.2.8 Intuitive Control of Parameters

When the features we extract represent well perceptually relevant characteristics of the objects (i.e., the images or sounds), the hybrids will be perceptually convincing. An example of a simple but powerful feature of faces that has a great impact on the perceived gender of the individual depicted in the image is facial contrast. Figure 2.16 shows a human face obtained by averaging the faces of several Caucasian men and women to obtain an androgynous face. Features such as nose and mouth shape, curvature of the eyebrows, among others that are usually used as cues to identify gender are the same in the two images. The only difference between the two faces shown is the contrast between adjacent areas, much sharper for the face shown on the right of figure 2.16.

By isolating the contrast and showing us that the faces are still perceived as male or female, Russell [Russel, 2009] demonstrates the existence of a sex difference in facial contrast. Ideally we would like to be able to do the same with sounds, that is, to be able to identify perceptually relevant features of sound and manipulate them independently of the others. Let us not forget that usually we manipulate the features indirectly by manipulating the parameters. So the question becomes "What is the most appropriate set of parameters or representation for a given transformation?"

The answer depends on the nature of the objects we want to transform, the transformation we want to perform and the features we wish to manipulate.

Figure 2.16: Illustration of perceptually relevant transformations obtained by manipulation of features that capture symbolic information related to a cognitive representation of the images, in this case, gender.

2.2.9 Types of Transformation

In this section I will present different types of transformation that respect the interpolation principle (therefore can be considered morphing) from the theoretical point of view. We have already discussed how morphing is a transformation that can be applied to either static objects, such as images, or to dynamic objects, such as sounds or movies. We call them dynamic objects because of their intrinsic temporal dimension, that is, they evolve in time.

All the examples we have considered so far use static objects and we have concentrated on obtaining one hybrid object with intermediate form. This leads us to the first type of transformation, the static transformation. But we also know from figure 2.14 that it is possible to obtain a series of different hybrid objects by specification of different values of the morphing factor alone. We can specify a (discrete) trajectory between the base objects, and obtain several hybrid objects, each corresponding to a point in this trajectory. This procedure gives rise to the second possible type of transformation, a dynamic transformation. We finally have four possibilities when we apply the two types of transformation to the two types of object.

- Static transformations between static objects generate static objects;
- Dynamic transformations between static objects generate dynamic objects;
- Static transformations between dynamic objects generate dynamic objects;

In this case we usually interpolate the temporal dimension. That is, the duration of the hybrid object is dictated by the duration of the base objects being combined.

- Dynamic transformations between dynamic objects generate dynamic objects.

In this case we can decide to either respect the temporal dimension or warp it. In other words, we can interpolate the duration or simply make it shorter or longer at will, repeating or omitting hybrids. Finally, it is important to bear in mind that all these possibilities arise uniquely from the unique nature of morphing. We should be able to specify any such transformation by simply choosing the base objects and specifying the trajectory defined by the values of the morphing factor.

However, when we want to morph dynamic objects (such as movies or sounds), we need to take the intrinsically temporal dimension into account. Morphing dynamic objects is much more difficult than than static objects because the temporal dimension leads to many different possibilities. How can we transpose the formalization above to intangible and immaterial things such as sounds, that are mere cognitive representations of patterns of air pressure that reach our eardrums? Chapter 3 addresses this question.

Chapter 3

Morphing Sounds

The aim of morphing between different musical instrument sounds across perceptually salient timbre dimensions is to obtain sounds that would correspond to hybrid instruments. That is, sounds that seem to come from an instrument that contains characteristics that are intermediate to the original sounds'. The morphed sound must be perceived as one single sound and it must not contain characteristics of either sounds used to create it. Rather, it must have its own features derived by combining those of the original sounds. We can identify two major difficulties when morphing between musical instrument sounds. The first is directly related to the relationship between timbre and sound source in the context of musical instrument sounds. The second is more subtle and is related to the perceptual attributes of musical instrument sounds.

This chapter explores how the conceptual and theoretical considerations about morphing in general presented in the previous chapter apply specifically to sounds. Morphing sounds is much more complicated than morphing static objects such as images because of the intrinsic temporal nature of sounds. Now the time dimension makes the image morphing analogy imperfect and calls for more appropriate imagery. This chapter addresses theoretical aspects of sound morphing aiming at formalization of the problem and homogenization of nomenclature. More importantly, this chapter reviews the types of sound transformation that can be considered morphing according to this formalization. Special attention is given to the cyclostationary morph, the aim of this thesis, and how to evaluate it under the criteria presented in the previous chapter, namely, correspondence, intermediateness, and smoothness.

3.1 From Image Morphing to Sound Morphing: Conceptual Considerations

The previous chapter explained how morphing is intrinsically associated with objects, which, in turn, shows that use of the word morph is tied to form in a tangible way. On a more abstract plane, we can extend the concept of form and associate it to sound objects [Schaeffer, 1966] and their characteristics, which Smalley [Smalley, 1997, Smalley, 1986] calls sound shape.

3.1.1 Sound Object

Pierre Schaeffer introduced the concept of sound object in his monumental study entitled "Traité des Objets Musicaux [Schaeffer, 1966]." Much of his text is a philosophical defense of an attitude toward sonic experience that derives from the *musique concrète* tradition. In the core of the study,

Schaeffer describes an ambitious research program whose aim is nothing less than the classification of all sound and a pedagogy to train musicians in the musical use of the classification scheme.

Essential to Schaeffer's approach is the development of the concept of reduced hearing. The common mode of listening, in which we respond to a sound stimulus by identifying its source - the sound "is" an oboe, a jet plane, etc - must be distinguished, according to Schaeffer, from another mode, in which we purposely - perhaps in some sense automatically - divorce what we hear from its source, concentrating instead on the properties of the sound itself. This kind of objectification or reduction of sound is required for a sonic event to be heard as a "sound object." Slawson [Slawson, 1985] remarks that "reduced hearing seems a prerequisite for the abstract mode of listening that hears timbre as a dimension of sound." In other words, the concept of sound object and the reduced hearing process involved challenge the simplistic view of timbre as sound source identification. It fully supports the notion of timbre space, timbre perception as multidimensional, dimensions of timbre perception, correlates of timbre dimensions, sonic continuum, etc. I will delve deeper into the discussion of timbre perception in chapter 5.

3.1.2 Sound Shape

The composer Denis Smalley complements the concept of sound object with that of sound shape, explored in his theory of spectromorphology [Smalley, 1997, Smalley, 1986]. A key element in his theory of spectromorphology is the rupture between the perception of sounds and musical instrument identification so deeply rooted in our listening tradition. Smalley defines the concepts and terminology of spectromorphology as tools for describing and analyzing the listening experience, and states that the two parts of the term refer to the interaction between sound spectra and the ways they change and are shaped through time. According to Smalley, a spectromorphological approach sets out spectral and morphological models and processes, and provides a framework for understanding structural relations and behaviors as experienced in the temporal flux of the music.

The concept of sound shape is directly related to the problem of graphic representation of electroacoustic music. In other words, the traditional score based on the grid criticized by Wishart [Wishart, 1996] needs to be adapted to represent the temporal flow, hierarchical structures, and organization of sonic material in electroacoustic pieces. One possible solution is to use shapes to represent different sonic events according to certain criteria. One problem we are faced with is the lack of homogeneity of notation, which was already a concern when Schaeffer first introduced the concept of sound object. Smalley explains how the concept of sound shape can be applied to traditional instrumental music and the spectral flux resulting from the associated gestures. Concepts such as attack, sustain and release appear throughout, together with the concept of spectral shape and the flux of spectral information in time, which are central in the formalization of sound morphing presented in the next pages.

3.1.3 Sound Morphing

With the previous notions clarified, we are ready to apply the general concept of object morphing to sounds. In very broad terms, the sound object imagery combined with the object morphing analogy lead us to think of sound morphing as a gradual change of shape from one sound object to another. This gradual shape transformation would naturally involve the attack, sustain and release characteristics of sounds, as well as the spectral shape and the spectro-temporal flux intrinsic to sounds.

In the light of the spectromorphological formalization put forward by Smalley, the global temporal and spectral dimensions of musical instrument sound perception will be treated independently. In this thesis, the focus of interest is to morph between musical instrument sounds across dimensions of timbre perception using perceptually motivated features as guides. We will consider

global temporal features of sounds such as attack time, sustain and release separately from features strictly related to spectral shape, such as spectral centroid, spread, skewness and kurtosis. The obvious link between the two is the temporal flux of spectral shapes that constitutes the basis of perception of global temporal features as a succession of events (attack, sustain, release).

The basic proposal of this work is to objectively measure the sound shape of musical instrument sounds with sonic features and use them to guide the morphing process. Conceptually, we could imagine that intermediate values of features imply intermediate shape, and use the same evaluation criteria for general object shape morphing. This approach becomes more relevant to the problem at hand when the features we use to guide the morphing transformation are correlated with musical instrument sound perception, such that intermediate values of features would imply perceptual intermediateness. Intermediateness is one of the three requirements in morphing presented in chapter 2, and it was adopted in the evaluation procedure used in this work. The feature values are the basis of the objective measures used here.

Chapter 5 explains the link between the sonic features we chose to guide the transformation and musical instrument sound perception. However, we still need to clearly visualize how the temporal flow of spectral information influences morphing from a conceptual, aesthetic and practical perspectives.

3.2 The Image Morphing Analogy Revisited

Images are static (or stationary) objects. Due to the intrinsic temporal nature of sounds, we need a better analogy that captures the dynamic evolution of sounds and allows us to better understand the technical requirements of sound morphing.

3.2.1 An Even Better Analogy: Movie Morphing

Due to the intrinsic temporal nature of sounds, a better analogy would be that of movie morphing [Slaney et al., 1996], where the aim must be reviewed to better fit the dynamic nature of the media, depicted in figure 3.1. This is a somewhat trickier problem than image morphing because of the added temporal dimension. In short, we are going to treat sound morphing as if it were movie morphing.

Now our sound morphing analogy has closer correspondences. For example, each movie frame could correspond to an STFT frame resulting from the analysis of the sounds we intend to morph between. Also, we can imagine that each frame's visual features have a corresponding set of sonic features that also evolve in time and that this evolution in time itself carries important information about how we perceive the movie (sound).

Notice that figure 3.1 depicts movies (or sounds, as in the case of this thesis) with different numbers of frames, therefore, different lengths (supposing the same frame rate). In other words, we view each frame of the STFT as a static "image" or snapshot, such that each frame has a set of features associated. But the frames are not independent, they are perceived as a sequence in time, and the evolution of the features is important.

The sound morphing problem starts with finding the correspondence between the frames of the sounds we are going to morph between, addressed theoretically in chapter 8 and in practical terms in chapter 12. But we also have to consider how we are going to morph between frames, because the correspondence problem, which Osaka [Osaka, 2005] calls the matching problem, is essential in morphing. Chapter 13 deals with the spectral correspondence problem, among others concerning morphing the spectral shape guided by perceptually correlated features.

The result depends on several aspects, such as:

- What type of transformation we want to perform;

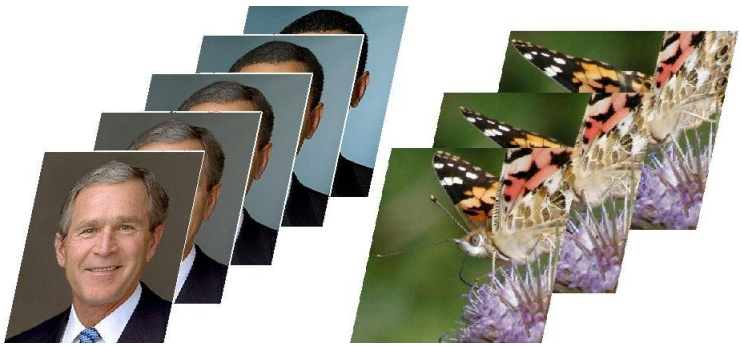

Figure 3.1: Depiction of two movies shown frame by frame

- What features we focus on;
- The sound model or representation that we use to perform the morph;
- The sound material we use as source and target.

Each of these issues will be considered in the next pages. However, before we delve into the technical details of morphing, we need to take care of the formalization of some theoretical aspects of sound morphing, such as terminology and definitions. Indeed, these have become issues due to the lack of formalism in the literature.

3.3 Formalization

This section is devoted primarily to the formalization of the terminology associated with sound morphing, such as tentative definitions that appear in the literature, and a careful revision of whether sound transformations that are commonly referred to as morphing fit the formalization of morphing presented in chapter 2.

3.3.1 Terminology

After a thorough review of the literature on the hybridization of sounds, it appears that there is much confusion in terminology. One of the aims of this work is to clarify a little bit the techniques referred to as morphing and the terminology itself. Apart from sound morphing, some authors refer to this transformation as audio morphing [Slaney et al., 1996], while others prefer timbre morphing [Tellman et al., 1995, Osaka, 1995] or even timbre interpolation [Hikichi, 2001, Osaka, 1995] to refer to similar goals, and some choose to use these terms interchangeably. The result of such transformations has been called hybrid [Fitz et al., 2003, Haken et al., 2006], intermediate [Caetano and Rodet, 2009, Caetano and Rodet, 2010c], interpolated [Hikichi, 2001] or even mongrel sound [Hope and Furlong, 1998].

In this work, we reserve the term sound for the auditory impression or the sensation perceived by the sense of hearing, whereas audio refers more specifically to the signal. Moreover, we make a distinction between interpolation and morphing. Interpolation acts on the parameters of a model, being restricted to the signal level, whereas we reserve morphing for the blending of perceptual qualities. So I propose sound morphing as the most appropriate term to meet the requirements of this work, and and to call "morphed sounds" the intermediate states of a sound morphing process.

3.3.2 Tentative Definitions

There seems to be no widely accepted definition of morphing in the literature. Instead, most authors either attempt to provide a definition of their own or simply explain what the aim of their work was. Some definitions are too system dependent to be useful, Fitz et al. [Fitz and Haken, 1996] define morphing as "the process of combining two or more Lemur files to create a new Lemur file with an intermediate timbre". Others are too general, such as Boccardi's [Boccardi and Drioli, 2001] "modifying the time-varying spectrum of a source sound to match the time-varying spectrum of a given number of target sounds".

Definitions based on the concept of timbre are common [Hikichi, 2001, Tellman et al., 1995, Fitz et al., 2003, Osaka, 2005]. Usually, these authors define timbre morphing as "the process of combining two or more sounds to create a new sound with intermediate timbre" [Tellman et al., 1995] or "to achieve a smooth transition from one timbre to another" [Hikichi, 2001]. We should notice that these refer to different goals. All in all, we prefer to avoid any definition that relies heavily on a concept as loosely defined and misunderstood as timbre, that can encompass many different perceptual dimensions of sounds [Letowski, 1992]. Although these authors usually do not define what they mean by timbre, most seem to refer to timbre as the set of attributes that allow sound source identification. In musical instrument contexts, this usually means that timbre becomes a synonym of musical instrument and thus timbre morphing reduces to hybrid musical instrument sounds. It is possible, though, to morph between sounds from the same instrument (different loudness or even different temporal features) [Tellman et al., 1995].

Slaney et al. [Slaney et al., 1996], on the other hand, prefer to avoid a direct definition altogether and explains the concept by analogy with image morphing instead, where the aim is to gradually change from one image (the source) to the other (the target) producing convincing intermediates (or hybrids) along the way. Other authors have proposed the same analogy [Fitz et al., 2003]. Nonetheless, they rely on the concept of sound object especially because they do not restrict their goal to musical instrument sounds.

Instead, this thesis defines the aim of sound morphing as obtaining a sound that is perceptually intermediate between two (or more). When morphing musical instrument sounds, the focus is on timbral qualities independent from loudness and pitch (LP-timbre, as defined by Letowski [Letowski, 1992]), especially those related to the spectral shape [Caetano and Rodet, 2010b], which Slawson [Slawson, 1985] termed sound color.

3.3.3 What Sound Morphing is Not

Thus far, we are already aware that a morphed object is expected to present intermediate characteristics inherited from the base objects that compose it. For faces, for example, it is not enough to compose a hybrid face from two using elements from either one. That is, if we compose a new face using the eyes from one face, the ears from the other, and so forth, we will indeed end up with a hybrid face, but the result of this operation does not fit into our definition of morphing.

The same goes for sounds. We can think of musical instrument sounds that possess characteristics of two others, such as the xaphoon, whose attack qualities resemble the saxophone's, but the more sustained portion of the sounds have a clarinet quality to them. A classic example is the guitarpschord, a hybrid acoustic musical instrument constructed by replacing the strings of a harpschord with a guitar's, giving the instrument a somewhat hybrid quality.

Nevertheless, the sound quality of the guitarpschord cannot be considered as morphing because some of its qualities come uniquely from the guitar (notably those associated with the material the strings are made of), while others come uniquely from the harpsichord (the hammers hitting the strings). After a while, the listener is able to say that the sounds heard are guitar strings

struck by hammers. [1] In the present document, we are more interested in obtaining sounds whose qualities are intermediate between two (or even more) other sound's, and specially on how to do it. The blending of perceptually relevant/related features of musical instrument sounds and the model developed to do it are the subject of this thesis.

3.3.3.1 Mixing Vs Morphing

The first thing we might be tempted to do when trying to blend perceptual qualities of sounds is to play them together. It does not take a lot of experience to be convinced that playing sounds at the same time (the signal processing counterpart of which is called mixing) will not give us the desired result. In fact, in our everyday experience, we are surrounded by counter examples. We know that the environment is constantly presenting us with a rich sonic experience (sometimes called sonic landscape or soundscape), with sounds coming from different sources, at different locations, and usually we can identify the different sources and spatial location even when the sound waves reach our ears concomitantly (i.e, a bird singing while an airplane flies by).

In fact, an even better example would be most music we hear. Usually there are several musical instruments playing at the same time and our brain is capable of keeping track of all of them separately. That is, the sounds of the instruments playing together do not blend, fusing into an amorphous mass of sound. This is mostly due to the way in which we hear sounds, keeping track of common relationships presented by partials produced by one instrument. For example, the attack of a plucked string imprints a unique quality to all the partials resulting from this sonic event, and our brains use this information to group them together. Spatialization cues resulting mostly from reverberation and affecting specially phase relationships between the partials are also used to group them together under a single sonic event.

In his classic "Computer study of trumpet Sounds [Risset, 1966]", Jean-Claude Risset discovered that the partials of trumpet sounds are slightly mistuned (i.e., they are not perfectly harmonic, or they are quasi-harmonic). He also verified that the partials present onset asynchrony, that is, each partial attacks at slightly different times, with higher partials tending to attack later than lower ones. Finally, he also described how the partials tend to fluctuate about a frequency in an erratic way. Risset postulated that the brain uses these factors as cues to group the partials produced by one instrument together. The same phenomenon was later verified for other acoustic musical instruments as well [Risset and Mathews, 1969].

In conclusion, simply mixing or cross-fading is not enough to obtain a result that can be defined as morphing according to the formal requirements we established earlier. The partials would have to be carefully aligned to fool our ears (and our brains). There are some interesting examples of morphed sounds obtained by studio techniques that include careful mixing and cross-fading, notably Trevor Wishart's "Red Bird" [Wishart, 1996]. However, as a general rule, we need to extract parameters from an analysis of the sounds according to a model, and somehow combine the parameters from both sounds to obtain a hybrid sound that contains intermediate qualities.

3.3.3.2 Cross-Synthesis Vs Morphing

It is very common to find authors who refer to cross-synthesis as sound morphing [Wen and Sandler, 2010]. We should keep in mind that the morphed sound should have characteristics that are perceptually intermediate between those associated with the sounds used in the morph.

The difference between sound morphing and cross-synthesis can be understood in an analogy with the problem of voice conversion as opposed to voice morphing. Stylianou [Stylianou, 2008]

[1] The interested reader can find out more about hybrid musical instruments on http://www.oddmusic.com/gallery/.

remarks that voice morphing is a type of transformation where the same sentence is uttered by two speakers and we want to generate a third speaker by combining the characteristics from the utterances of the original speakers. In voice conversion, on the other hand, the sentence to be converted, uttered by the source speaker, has never been uttered by the target speaker.

However, in voice morphing, there are two speakers that generate a new voice that utters the same sentence as the original speakers. In voice conversion there is only one source utterance and a target speaker. Here the aim is to transform the source utterance so as to imprint the voice characteristics of the target speaker on it, that is, we want to convert the utterance like it was spoken by the target speaker instead of generating a new speaker.

In cross-synthesis we have two sounds, the modulated or carrier sound and the modulator. In other words, a sound obtained by traditional cross-synthesis techniques (imprinting the spectral envelope of one sound onto the other) will produce a sound whose features are either from one or from the other sound. Notably, features associated with the source (the partials) will come from the modulated sound (or carrier), while those associated with the filter (the spectral envelope, responsible for sounds color) will come from the modulating sound.

Still, many different types of transformation are referred to as morphing in the literature. In the next section, we will review them using the movie morphing analogy.

3.3.4 Sound Transformations that can be Described as Morphing

We need to choose what kind of transformation we intend to do. Coming back to figure 3.1 and the movie morphing analogy, we could simply make a movie that contains an intermediate number of frames, but we need to account for important temporal information to make it more convincing. If the first movie shows an explosion at the beginning (similarly to the abrupt attack of a plucked string or a percussive sound) and the other a butterfly gently flapping its wings and then flying away, we might need to align relevant temporal cues to produce an interesting morph.

Moreover, there are a number of possible transitions between the two. Do we want an intermediate movie that contains morphed images of each frame (here called static or stationary morphing because the morphing factor α is constant), or are we going for a movie that starts as the first and dynamically changes into the other (here called dynamic morphing because α varies in time)? We could choose to run the first frames of the first movie until we stop at a selected frame, gradually morph it into another selected frame of the second, and then proceed by showing the rest of it (warped dynamic morphing), choosing to somehow warp the length of the result in order to achieve a given effect.

Finally, another possibility would be to produce several morphed movies in different intermediate points (i.e., different values of α) of the path between source and target (cyclostationary morphing). The different morphing processes are presented in order of increasing complexity from a purely procedural point of view. That is, the first process is the simplest to implement, independently of the impact of the result.

3.3.4.1 Warped Dynamic Morphing

As explained earlier, the transformation dubbed warped dynamic morphing consists of running the frames of the first sound up to a certain point, stopping at a selected frame, morphing smoothly from this selected frame to a selected frame of the second sound, and playing rest of second sound. The process is illustrated with movies in figure 3.2.

This transformation is very simple to perform and usually leads to remarkably impressive results. Notice that in this case we only morph one frame of each sound gradually to achieve the result, such that the hybrid sound preserves most of the original frames from either source or target sounds. We do not even need to concern ourselves with temporal issues like the duration

of the result because the transformation is usually performed using frames from the middle of the sounds (usually called steady-state or sustain, we will come back to this later in chapter 8), more spectrally stable than frames from the beginning (attack) or from the end section (release).

In fact, in this case we do not need to concern ourselves with the temporal evolution of features at all because we just use one isolated frame from each sound to perform the transformation. Temporal features such as duration of the attack and release of both sounds are not meddled with, so they are preserved.

This transformation warps the temporal dimension, which is not explicitly represented in timbre space, such that it is difficult to imagine its graphic representation. All the others have easier visualizations, as will be clearer later on.

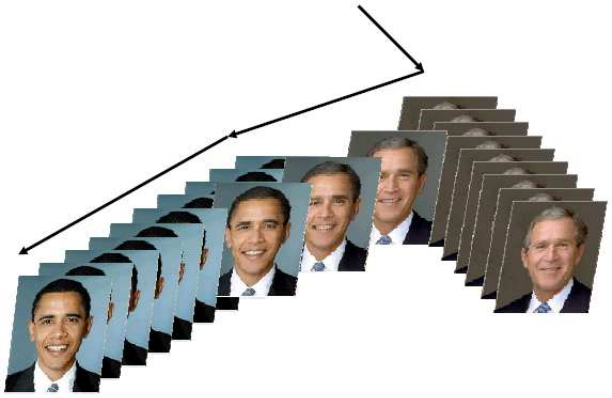

Figure 3.2: Depiction of Warped Dynamic Morphing Using Movie Frames

3.3.4.2 Static or Stationary Morphing

In the case of static or stationary morphing the picture changes radically. This transformation consists in establishing a correspondence between every single frame of both sounds and morphing between them with the same interpolation factor α, such that the hybrid sound has only morphed frames, like shown in figure 3.3.

This is a much more challenging transformation to perform because we need to take care of the temporal aspects as well as spectral aspects. That is, if the sounds do not have the same duration (therefore different number of frames), the the first decision is how we are going to associate frames from the first with frames from the second sound. Naturally, we could simply time-compress the longer sound (or equivalently time-stretch the shorter one) to guarantee that they have the same number of frames or a one-to-one correspondence.

Alternatively, we could compromise and obtain a hybrid sound whose duration corresponds to the interpolation of the durations of the sounds used in the transformation (perhaps obtained with the same interpolation factor α to be more consistent), but time stretching or compressing all the sounds with the same factor is hardly a good strategy because it does not take into account the temporal evolution of the features of the sounds. That is, a sound whose attack has been time-stretched to double the length, for example, is usually considered perceptually different from a sound played by the same instrument that is twice as long. Playing a longer sound in acoustical instruments usually does not affect the attack characteristics. This calls for strategies to align perceptually similar regions in time before trying to make a correspondence between the frames.

Automatically detecting these regions is an important contribution of this work and it is described in chapter 12.

Static or stationary morphing corresponds to a single point between source and target sounds in timbre space. The challenging aspect of stationary morphing is how accurately the morphing factor controls the intermediateness of the result. This is very difficult to evaluate perceptually, thus in this work we will use the values of the features in the evaluation.

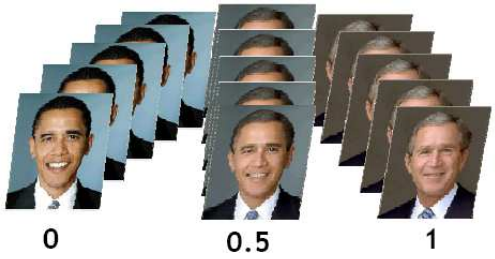

Figure 3.3: Depiction of static or stationary morphing using movie frames

3.3.4.3 Dynamic Morphing

Dynamic morphing is equivalent to static morphing in practical terms, the only conceptual difference being the dynamically changing morphing factor, as shown in figure 3.4.

In other words, we also need to concern ourselves with temporal considerations before establishing the correspondence between the frames, but once this step is behind us, we morph each pair of frames with a different interpolation factor. The result of this transformation is a morphed sound that gradually morphs from source into target sound along the course of the sound.

Naturally the result has almost only morphed frames (except for the first and last frames, maybe), and the same considerations about the duration of the hybrid sound apply here. Notice that in this case it is not obvious to interpolate the duration because we do not have only one value of interpolation factor, rather, we have as many as we have frames.

This transformation corresponds to a (sampled) continuous trajectory going from source to target in a projection of the dimensions of timbre space that do not depend on the attack. Challenging aspects of dynamic morphing are related to the smoothness and intermediateness of the transformation, which are intrinsically intertwined in a dynamic transformation.

This type of dynamic transformation is what most people expect to hear when they imagine the result of the morph between two sounds. However, dynamic morphing can only be effectively applied when the sonic material being morphed consists of a single event. When the sounds we want to morph between contain a sequence of events, cyclostationary morphing is a more interesting and appropriate way of doing it, as we will see next.

3.3.4.4 Cyclostationary Morphing

Finally, the cyclostationary morphing is achieved by simply repeating the static morphing N times with a morphing factor varying from 0 to 1 in $\frac{1}{N-1}$ steps, as shown in figure 3.5. The result of this transformation is N sounds (including source and target) that, when played sequentially, represent a cyclic sequence of morphed sounds, each corresponding to a point in a discrete path going from source to target in timbre space.

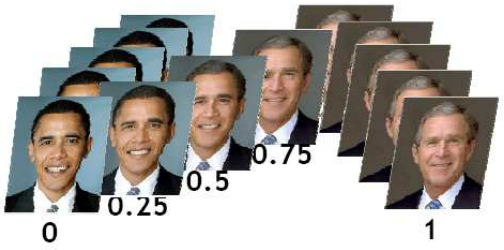

Figure 3.4: Depiction of dynamic morphing using movie frames.

Cyclostationary morphing is by far the most challenging morphing transformation. Just like stationary morphing, cyclostationary morphing also involves morphing temporal and spectral features of musical instrument sounds, so intermediateness is always an issue. But the added difficulty here involves the control of the smoothness of the transformation by the morphing factor α. For cyclostationary morphing, we need to have accurate control of the perceptual distance across steps of the morph to achieve a gradual morph.

In this thesis, the criteria of intermediateness and smoothness will be used to evaluate cyclostationary morphs between musical instrument sounds. Therefore, the next section will address theoretical and practical aspects of this particular choice of morph.

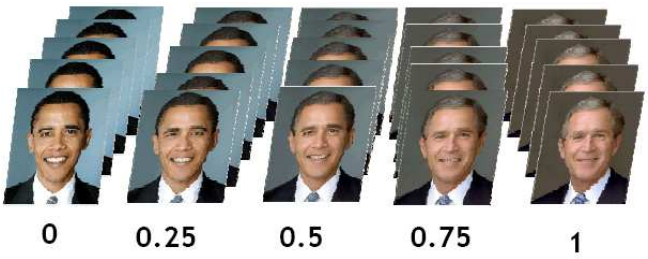

Figure 3.5: Depiction of cyclostationary morphing using movie frames

3.3.5 Practical Aspects

This section analyzes practical aspects of the theoretical consideration raised in section 2.2 when morphing isolated musical instrument sounds.

In this work, we present models and techniques developed to automatically morph between isolated quasi-harmonic acoustic musical instrumental sounds. As will be clear throughout the text, this choice has determined several technical aspects, such as the choice of the sound model used to describe the sounds, among others.

Even though the focus is very specific, this does not mean that the techniques described here are restricted to the class of sounds they were originally developed for. We can easily think of ways to extend most of them because they obey the general principles presented in section 2.2.

3.3.5.1 Correspondence

A direct consequence of the principles of correspondence and conceptual distance presented earlier in chapter 2 is that the choice of sound material has great impact on the quality of the results. Naturally, it would be easier to morph between two quasi-harmonic musical instrument notes than between the singing voice and drums because of the conceptual distance. When the sounds being morphed are quasi-harmonic, both can be described by partials, so the correspondence between the partials could use the partial number.

Another important factor that affects the quality of the morph is the number of events. When each sound is one single event (note), it is easier to find correspondences between them. When one sound has multiple events (beats) and is highly inharmonic (percussion) and the other is one single quasi-harmonic event, it demands more refined techniques to find correspondences between them.

The techniques applied to achieve the results vary according to the sounds we choose to morph between. Also, the features we focus on depend on the sound material. For instance, pitch is a salient attribute of quasi-harmonic musical instrument sounds, such that we need to consider it very carefully when morphing between pitched sounds. Inharmonic environmental sounds would probably require that we focus on different features to obtain perceptually convincing morphs.

3.3.5.2 Hybrid Musical Instruments

In the context of musical instruments, the concept of timbre is intimately linked to sound source identification. The term timbre is sometimes applied to refer to different concepts, and some authors prefer to avoid using the word timbre altogether and propose alternatives, such as sound quality or color. In simplistic terms, we could say that timbre is what allows us to identify the source of the sound we hear, or the instrument that played the sound, even though we know that the same instrument might possess multiple registers corresponding to timbral variations. For example, the clarinet has three distinct registers, each of which has its own characteristics and sounds different from the others. A brassy trumpet sound is very different from a softer one, but we are still able to recognize the source of both as being the same instrument.

We will elaborate on these matters later in chapter 5, but it is important to remember that musical instrument/sound source recognition is part of the musical education and, as such, has to be practiced. This means that musical instrument recognition is a cognitive task that we train our brains to perform, so it might be more intrinsically connected with our musical system than with the sound features themselves. We should not be surprised by the conclusion that the result of using of a brassy trumpet sound or a softer one in a morphing would lead to perceptually different results.

The consequence for sound morphing is that it might not be enough to obtain morphed sounds that are recognized as hybrid instruments, we might have to control how we blend the perceptual features of the original sounds in order to attain a perceptually convincing morph. In the next chapter the state of the art of sound morphing will be thoroughly reviewed. The main objective of this review of the literature is to evaluate whether these sound morphing techniques meet the three evaluation criteria of a good morph: correspondence, intermediateness and smoothness. Chapter 5, right after that, will come back to the question of perception of hybrid musical instruments using timbre spaces as guides.

Chapter 4

State of the Art

This chapter is entirely dedicated to present the state-of-the-art sound morphing techniques. The information presented here is the result of an extensive and thorough review of the literature, and it is organized as follows. Firstly we will succinctly review additive sound models, which led to the development of sinusoidal modeling by Smith and Serra [Smith and Serra, 1985] and McAuley and Quatieri [McAulay and Quatieri, 1986, McAuley, 1984] independently.

Sinusoidal modeling stands out as one of the most popular models used in sound transformations in general [Serra and Bonada, 1998, Amatriain et al., 2003, Amatriain et al., 2002]. Sound morphing figures prominently as one of the most successful transformations attained with sinusoidal models. Thus we will discuss how to obtain morphed sounds with sinusoidal models by following the interpolation principle presented in chapter 1.

Finally, we conclude this chapter by briefly reviewing other sound morphing techniques proposed in the literature. This chapter is intended primarily to position this thesis in the context of the existing sound morphing techniques. Interestingly, we will also find in this chapter the motivation behind the development of new sound morphing techniques, such as those proposed here.

This chapter should be read with the perceptual characteristic of the sound morphing problem in mind. This will be the background against which we will compare the adoption of the source-filter model and the development of the temporal segmentation, temporal alignment, and spectral envelope morphing techniques developed in the course of this thesis and presented in later chapters.

4.1 Sound Signal Models

Sinusoids are the cornerstone of sound signal analysis. On the one hand, Fourier's theorem states that any periodic signal can be decomposed into a sum of harmonically related sinusoids [Hartmann, 2007, Hartmann, 1998]. On the other hand, sinusoids emerge in acoustic models of sound production as the elementary solutions (modes of vibration) to a large variety of oscillating systems [Hartmann, 2007, Fletcher and Rossing, 1998]. As a result, the representation of sound signals by a sum of amplitude-frequency modulated sine waves and analysis/synthesis techniques based on this representation have become essential tools in music and speech sound processing.

The physical generation of music signals is in part similar to the generation of speech signals, and thus it is not surprising that sinusoidal-based processing, useful in one area, is also useful in the other. In certain wind instruments, for example, a vibrating reed excites the instrument's cavity, while in speech the vibrating vocal cords excite the vocal tract. There are common signal classes in the two domains, all of which can be represented to a certain degree by a sum of amplitude and frequency modulated sinusoids.

Quasi-periodic signals, such as steady speech vowels and the sustained region of certain musical instrument sounds, can be accurately represented as a finite sum of harmonically related sinusoids with slowly time-varying amplitudes and frequencies. Noise-like signals, such as speech fricatives and musical instrument turbulence, have no clear harmonic structure and transients, such as speech plosives and musical instrument attacks and decays, may be neither harmonic nor noise-like, consisting of short acoustic events that occur prior, during, or after steady regions. Noise-like and transient sounds can be represented approximately by a sum of inharmonically related sinusoidal partials, generally without coherent phase structure.

A typical sound is often a mixture of these components whose relative weights, timing, and duration can be key to accurate modeling. Early approaches to music analysis relied on a running Fourier transform to measure sine-wave amplitude and frequency trajectories. This technique evolved into a filter bank-based processor and ultimately to signal analysis/synthesis referred to as the phase vocoder [Flanagan and Golden, 1966]. The phase vocoder is at the core of sinusoidal modeling, one of the most popular sound signal models due to the accuracy and flexibility of representation and manipulation of parameters.

4.1.1 Additive Synthesis

Additive synthesis is the oldest, conceptually simplest, and perhaps most widely used sound signal model. It relies on the assumption that any sound may be modeled as the sum of a number of sinusoids with time-varying parameters, also called partials (or harmonics when they are harmonically related).

There are a great number of variations and extensions to this model, all of which maintain in some fashion the basic principle of partial summation. Several of these variations will be discussed later in this chapter.

In general, the series of partials which may be used to represent any sound is given by

$$x(t) = \sum_{k=1}^{K} A_k \cos \phi_k(t) \tag{4.1}$$

where A_k is the instantaneous amplitude of the k^{th} sinusoid, ϕ_k is its phase, and K is the number of partials we include. Equation 4.1 is used to define the value of the time-domain waveform $x(t)$ at time t. Each of the parameters is continually evolving. Successive phase and amplitude values are used to describe the evolution of each sinusoid, the summation of which can create complex wave shapes and rich timbres. The evolution of the phase may be better defined as follows

$$\phi(t) = \int_0^t \omega(\tau)\, d\tau \tag{4.2}$$

where ω is the frequency in radians. The frequency f in Hertz is equal to $\frac{\omega}{2\pi}$ and may be determined from the evolution, or rate of change, of the phase

$$f = \frac{1}{2\pi}\frac{d\phi}{dt} \tag{4.3}$$

There is a close relationship between the concept of additive synthesis and Fourier analysis, the most widely used method for converting a sound into its spectral representation. As we saw in chapter 2, the word spectrum was coined by the German mathematician David Hilbert in reference to the set of all eigenvalues of a linear operator. Sinusoids are eigenfunctions of linear shift-invariant (LSI) systems, and as such form the basis of Fourier analysis. Therefore, understanding the Fourier

transform is a necessary first step to understanding additive synthesis as it allows the parameter values for each partial in an additive synthesis representation to be determined.

The Fourier series is the starting point for all of the representational techniques to be described herein. The Fourier transform and its inverse allows for a lossless transformation of a signal into the spectral domain and back. Fourier's theorem can be expressed mathematically as

$$x(t) = \sum_{k=1}^{K} A_k \cos(2\pi f_k t + \phi_k) \tag{4.4}$$

where $x(t)$ is a periodic signal, A_k and f_k are the amplitude and frequency, respectively, of the k^{th} sinusoidal component at time t, and ϕ_k is its initial phase. Equation 4.4 serves not only as a model to analyze musical instrument sounds and speech, but also as a model for sound synthesis. In synthesis, the control functions A_k and $\omega_k = 2\pi f_k t + \phi_k$ were initially set manually based on knowledge of the sound we wish to synthesize, such as a musical instrument sound.

One of the first attempts to estimate the control functions is perhaps due to Moorer [Moorer, 1977]. Assuming the presence of one periodic sound in a measurement $x(n)$, the length of the signal is set equal to the waveform's pitch period N. The real and imaginary components are then given by

$$c_k(n) = \sum_{r=n}^{n+N-1} x(r) \cos(rk\omega_0) \tag{4.5}$$

$$d_k(n) = \sum_{r=n}^{n+N-1} x(r) \sin(rk\omega_0) \tag{4.6}$$

where $\omega_0 = {}^{2\pi}/_{N}$ and from which we can obtain the estimates of the slowly time-varying amplitude $\hat{a}_k(n)$ and phase $\hat{\theta}_k(n)$ of each harmonic

$$\hat{a}_k(n) = \sqrt{c_k^2(n) + d_k^2(n)} \tag{4.7}$$

$$\hat{\theta}_k(n) = \arctan\left[\frac{d_k(n)}{c_k(n)}\right] \tag{4.8}$$

The frequency of each harmonic is given approximately by the derivative of the unwrapped version of the phase $\hat{\theta}_k(n)$. A limitation of this method is that the pitch period must be known exactly to obtain reliable estimates.

The Fourier series supposes that the signal $x(t)$ being analyzed is periodic. To investigate the frequency components present in any signal, we use the Fourier transform instead, as shown in equation 4.9.

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi f t)\, dt \tag{4.9}$$

where $X(f)$ is the frequency spectrum of the input signal $x(t)$. The Fourier transform assumes that the period of repetition of the signal is infinite, as expressed by the limits of integration from$-\infty$ to ∞ . The continuous-time Fourier transform in equation 4.9 models the signal as a distribution of a continuous variable t, but in practice its implementation is not feasible given the discrete nature of the digital computer. The equation must thus be modified to reflect sampled data such as digital sound signals and discrete spectra. The discrete Fourier transform (DFT) is the standard way of representing the discrete-time signal $x(n)$ as a spectrum $X(k)$ with N discrete frequency bins.

4.1.2 Discrete Fourier Transform

The sequence of N real or complex numbers $x_0 ... x_{N-1}$ is transformed into the sequence of N complex numbers $X_0 ... X_{N-1}$ by the discrete Fourier transform (DFT) according to the formula

$$X(k) \triangleq \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi}{N} kn} \tag{4.10}$$

for $k = 0 ... N-1$. The original sequence can be recovered through the inverse transform, defined as

$$x(n) \triangleq \frac{1}{N} \sum_{n=0}^{N-1} X(k) e^{\frac{j2\pi}{N} kn} \tag{4.11}$$

for $n = 0 ... N-1$.

In practice, we use an algorithm introduced by Cooley and Tukey [Cooley and Tukey, 1965] known as the fast Fourier transform (FFT).

The Fourier transform considers the signal as a whole and does not permit the identification of events in time. One way to get past this drawback is to break the input signal into a series of very small segments evenly distributed in time. This leads us to the short-time Fourier transform (STFT). There are two immediate benefits to this process. First, it allows a time-localized representation of the spectrum. That is, it is possible to see which frequencies are present in the signal at a specific point in time (i.e. over a period of a few milliseconds). Second, it is computationally much more efficient to compute the transform for each segment than for the entire signal.

4.1.3 Short-Time Fourier Transform

In order to extract many short segments from the sound signal, a clever trick is to zero-out data outside of our consideration, leaving only a small segment of data. This technique is known as windowing, as it is akin to viewing only a small window of the data. Really, it is nothing more than a specific type of temporal envelope designed for spectral analysis.

Then, by performing Fourier analysis on each windowed segment, a sequence of measurements that constitute a time-varying spectrum is obtained. Together with the windowing function, this is given by

$$X(k,n) = \sum_{m=-\infty}^{\infty} w(n-m) x(m) \exp\left[-j\left(\frac{2\pi}{N}\right) km\right], \; k = 0, 1, \cdots, N-1 \tag{4.12}$$

where $X(k,n)$ is the amount of spectral activity at the k^{th} frequency bin, as determined by data centered around sample n, and $w(m)$ is the window of length M that selectively determines the portion of $x(m)$ being analyzed.

For a real signal, the STFT yields a sampled complex spectrum with $N/2 + 1$ complex values, where N is the number of samples used in the analysis. There are two equivalent but distinct interpretations to equation 4.12, the Fourier transform and the filter bank formulations. Figure 4.1 illustrates both complementary views of the STFT.

The Fourier transform interpretation views $X(k,n)$ as the Fourier transform of of the modified sequence

$$y_n(m) = x(m) w(n-m) \tag{4.13}$$

.

For this case, we interpret $X(k,n)$ as a function of the frequency index k for a fixed value of the time shift n.

In the Fourier interpretation, the number of filters is simply the number of points in the Fourier transform. Similarly, the equal spacing in frequency of the individual filters can be recognized as a fundamental feature of the Fourier transform. On the other hand, the shape of the filter passbands is determined by the shape of the window function applied before calculating the Fourier transform. Equation 4.13 shows that, for n constant, $y_n(m)$ is a product of x and w. Thus the Fourier transform of y_n is the complex convolution of the Fourier transforms of x and w. As such, the details of the resulting short-time Fourier transform are greatly influenced by the choice of the window w. Thus it is important to use a window that features the desired time and frequency resolution of the STFT.

The filter-bank interpretation of equation 4.12, usually called phase vocoder, is that of a filter bank analysis in which $X(k,n)$ is viewed as a function of the time index n for a fixed frequency k. In this case, the window $w(m)$ is seen a low-pass filter that determines all of the properties of the filter-bank. The design of the filter w is dominated by an important consideration, the sharper the filter frequency response cuts-off at the band edges, the longer its impulse response will be. In other words, to get sharp cut-offs with minimal overlap, we must use filters whose time response is very slow.

Historically, the phase vocoder comes from a long line of voice coding techniques which were developed primarily for speech processing [Dolson, 1986]. Indeed, the word "vocoder" is simply a contraction of the term "voice coder". There are many different types of vocoders. The phase vocoder was first described by Flanagan and Golden [Flanagan and Golden, 1966] in what is now a landmark in speech processing.

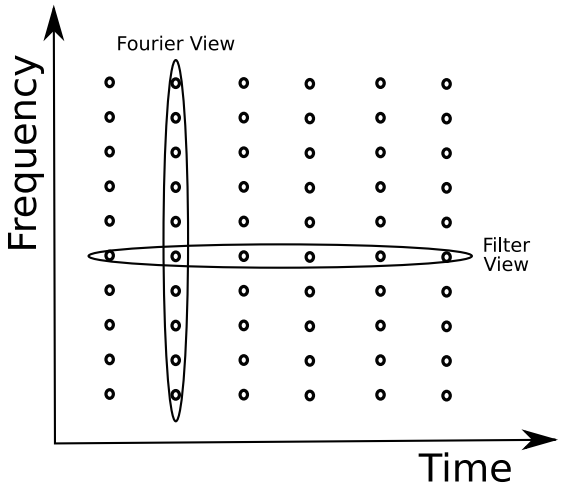

Figure 4.1: Filter-bank interpretation vs. Fourier transform interpretation. Adapted from Dolson [Dolson, 1986]

4.1.4 Phase Vocoder

The STFT provides information about a sound signal's spectral content at discrete time intervals. However, an equivalent way of viewing this information would be to consider each analysis bin

as the output of a band-pass filter. In this case $X(k,n)$ can be written as the linear convolution (denoted by *) of the signal $x(n)\exp\left[-j\left(\frac{2\pi}{N}\right)km\right]$ with the impulse response $w(n)$, i.e.,

$$X(k,n) = \left[x(n)\exp\left[-j\left(\frac{2\pi}{N}\right)km\right]\right] * w(n) \tag{4.14}$$

where $w(n)$ is a low-pass filter being applied to the signal $x(n)\exp\left[-j\left(\frac{2\pi}{N}\right)km\right]$. The modulation of $x(n)$ by $\exp\left[-j\left(\frac{2\pi}{N}\right)km\right]$ serves to shift the frequency spectrum of $x(n)$ at frequency $\omega_k = {}^{2\pi k}/_{N}$ down to 0 frequency. This operation is called heterodyning. Thus the STFT can be thought of as filtering the shifted spectrum of $x(n)$ in the region of frequency ω_k by the low-pass filter $w(n)$, usually called heterodyne filtering.

Yet another alternative way of viewing the phase vocoder as a filter bank is as follows. In equation 4.12, we change the variables to $n - m = l$ and get

$$X(k,n) = e^{-j\omega n}\sum_{l} w(l)\,x(n-l)\,e^{j\omega l} = e^{-j\omega n}\left[x(n) * w(n)\,e^{j\omega n}\right] \tag{4.15}$$

Now equation 4.15 can be viewed as first a modulation of the window to frequency ω, producing a band-pass filter $w(n)\,e^{j\omega n}$ followed by the filtering of $x(n)$ through this band-pass filter. The output of the filtering operation is then modulated back down to baseband by the complex exponential $e^{-j\omega n}$. Since each filter response is complex, the amplitude and phase of the output of each channel can be viewed as an amplitude and phase modulated complex sinusoidal

$$x(n) = \sum_{k=1}^{K} \hat{a}_k(n)\,e^{j\hat{\theta}_k(n)} \tag{4.16}$$

where $\hat{a}_k(n)$ and $\hat{\theta}_k(n)$ are calculated as in equations 4.7 and 4.8 respectively. In this case, c_k and d_k are the real and imaginary parts of the output of each filter.

Each filter, then, represents the time-varying energy in that particular frequency region, as defined by the operation of the heterodyne filtering technique. The use of the STFT in this fashion as an analysis/synthesis technique is called the phase-vocoder. The phase vocoder models a given signal as the sum of K sine waves, the parameters of which are determined by the STFT. These parameters include the time-varying amplitude, and phase of each sine wave, as calculated in equations 4.7 and 4.8.

The difference between the phase vocoder and the conventional channel vocoder is that the phase information is preserved in each channel and that we can guarantee that the output is exactly identical to the input [Moorer, 1979]. Mathematically, the phase vocoder is just an alternative representation of the STFT.

Flanagan and Golden [Flanagan and Golden, 1966] state that the conventional channel vocoder separates vocal excitation and envelope functions. The envelope functions are essentially the same as each $\hat{a}_k(n)$ as in equation 4.16 and they are band-limited because within any given filter band, the result of heterodyning and low-pass filtering is a signal whose highest frequency is equal to the cut-off frequency of the filter. In the phase vocoder, however, information about the excitation is mostly encoded in the phase derivative signals because the phase functions $\hat{\theta}_k(n)$ are generally not bounded.

Moorer [Moorer, 1979] remarks that the formulas used to convert the real and imaginary part of the output of the filters into the amplitude and phase representation are nonlinear and are thus non-band-limited. The phase derivative signal is given by

$$\dot{\theta} = \frac{c_k\dot{d}_{k-} - d_k\dot{c}_k}{c_k^2 + d_k^2} \tag{4.17}$$

Flanagan and Golden [Flanagan and Golden, 1966] remark that when the number of channels is sufficiently large, the information about excitation is conveyed primarily by the phase derivative signals, while a small number of broad analyzing channels results in amplitude signals that contain more information about the excitation and phase signals that contain more information about the envelopes. Qualitatively, therefore, the number of channels determines the relative amounts of excitation and spectral information carried by the amplitude and phase signals.

4.1.4.1 Phase unwrapping

A direct consequence of using equation 4.8 to estimate the phase signal is that the arc tangent function gives a discontinuous result everytime it "wraps around" 2π (that is, at each complete cycle). Since the frequency values are calculated directly from the phase values using equation 4.17, we nee to first add 2π everytime the phase completes a full cycle. This is usually termed "phase unwrapping" and there are many proposals in the literature of how to do it efficiently. See for example [Kaplan and Ulrych, 2007] for a review.

4.1.4.2 Resynthesis

If the center frequencies of the individual band-pass filters happen to align with the partials (which would probably have to be near harmonic because of the DFT), then the outputs of the phase vocoder analysis are essentially the time-varying amplitudes and frequencies of each partial. The filter-bank itself has three constraints. First the frequency response characteristics of the individual band-pass filters are identical (they are the same window w), except that each filter is centered at a different frequency. Second, these center frequencies are equally spaced across the entire spectrum from 0 Hz to half the sampling rate. Third, the individual band-pass frequency response is such that the combined frequency response of all filters in parallel is essentially flat across the entire spectrum. This ensures that no frequency component is given disproportionate weight in the analysis and that the phase vocoder is in fact an analysis-synthesis identity.

The number of filters must be sufficiently large to guarantee that there is never more than one partial within the passband of any single filter. For near harmonic sounds, this amounts to saying that the number of filters must be greater than the sampling rate divided by the fundamental frequency [Dolson, 1986]. For inharmonic and polyphonic sounds, the number of filters may need to be much greater. If this condition is not satisfied, the partials within a single filter will constructively and destructively interfere with one another, and the information about their individual frequencies will be coded as an unintended temporal variation in a single composite signal. In fact, several other problems might arise.

The detected partials are not allowed to vary outside the bandwidth of a given channel, otherwise they would be detected by more than one filter at the same time. This would certainly obscure the representation, and undermine the clarity of each component. Conversely, sometimes a partial can fall between the cracks of two analysis bins where it is not well-represented by either filter. This gives rise to representational difficulties: any sonic inputs that have a continuous spectrum or noisy components cannot be clearly represented or easily modified because they are not well-represented by a summation of sinusoids. Instrumental onsets or vocal fricatives, for example, fall into this category. Thus, while it is possible to perfectly reconstruct the input waveform, we are left with a model that presents a confusing representation of components which vary more than the bandwidth of one channel.

Yet another problem is that of phase dispersion, In time-scale modification, for example, the integration of the phase derivative and scaling of the unwrapped phase results in a loss of the original phase relation among sine waves, thus giving an objectionable "reverberant" quality characteristic of this method. Finally, the phase vocoder was formulated in the context of discrete sine

waves and hence was not designed for the representation of noise components of a sound.

The analysis stage of the original phase vocoder and its refinements views sine-wave components as outputs of a bank of uniformly-spaced bandpass filters. Rather than relying on a filter bank to extract the underlying sine-wave parameters, an alternate approach is to explicitly model and estimate time-varying parameters of sine-wave components by way of spectral peaks in the short-time Fourier transform [McAulay and Quatieri, 1986, Smith and Serra, 1985]. This approach lends itself to sinewave tracking through frequency matching, phase coherence through a source/filter phase model, and estimation of a stochastic component by use of an additive model of deterministic and stochastic signal components. As a consequence, the resulting sine-wave analysis/synthesis scheme resolves many of the problems encountered by the phase vocoder, and provides a useful framework for a large range of speech and music signal processing applications.

4.1.5 Classical Sinusoidal Modeling

The problem in analysis/synthesis is to take a waveform, extract parameters that represent a quasi-stationary portion of that waveform, and use those parameters or modified versions of them to reconstruct an approximation that is "as close as possible" to a desired signal. Furthermore, it is desirable to have a robust parameter extraction algorithm since the signal in many cases presents acoustic noise. The general identification problem in which the signal is to be represented by multiple sine waves is a difficult one to solve analytically [McAulay and Quatieri, 1986].

In the mid-1980s two independent solutions to the shortcomings of the phase vocoder were proposed. PARSHL by Julius Smith and Xavier Serra [Smith and Serra, 1985], and a sinusoidal model by McAulay and Quatieri [McAulay and Quatieri, 1986] are both intuitively-simple representations whereby the sound is modeled as the sum of a number of sinusoids. Both systems escape the band-limited nature of the phase-vocoder's time-varying filters, as the partials of the input sound are no longer bounded to a particular analysis channel and are free to vary across channels.

The input sound at time t is modeled as

$$x(t) = \sum_{k=1}^{K} A_k \cos(2\pi f_k t + \phi_k) \tag{4.18}$$

where A_k is the instantaneous amplitude of the k^{th} sinusoid, and ϕ_k and f_k are its initial phase and frequency, respectively. Whereas the phase-vocoder had a representation consisting of a fixed number of filters, the sinusoidal model can track an arbitrary number of partials, each of which is not constrained or obscured by the limits of a particular filter channel. The first step to obtain this representation is to detect any peaks or local maxima in the frequency spectrum, and to organize them into some number of time-frequency tracks. The detection of peaks is usually referred to as peak-picking while the process of defining sets of sine waves that will be continuously evolving in time is called peak matching.

4.1.5.1 Peak Picking

A peak is not always nicely resolved or clearly defined in the spectrum. This is especially true during onsets or with noisy signals. Therefore, in order to correctly identify and track prominent spectral peaks in the signal, the analysis depends more heavily on proper parameter settings than in the regular STFT. Additionally, due to the sampled nature of the spectrum, it may be difficult to determine the precise frequency location of a detected peak. The estimate will only be accurate to within $1/2$ of the spectral sampling period. Therefore, it may not always be so simple as to just

pick out the K greatest points in the spectrum; other techniques should be used to ensure proper (and accurate) selection.

In the peak identification process, there are several parameters used to control the operation of the algorithm, the most basic of which is a simple peak-height threshold, which is determined in relation to the relative power of nearby frequency components. This allows only the most prominent peaks to remain while others are removed. The location (frequency bin) of all remaining maxima in the frequency spectrum are recorded. Next, in order to determine the precise frequency of each detected peak, there are two different ways to proceed. The first utilizes the corresponding phase value of the located peak. The instantaneous frequency is ascertained by taking the derivative of the instantaneous phase value (unwrapping the phase as necessary) and adding it to the bin frequency. These extracted frequencies represent the precise location of each partial, which are then organized into each frequency track.

The other method involves a combination of zero padding and parabolic interpolation. Zero padding a windowed segment before Fourier analysis provides greater spectral resolution, which in turn minimizes the error in the estimate of a peak's spectral location. In this case, it also increases the resolution of the spectrum sufficiently to get a reasonable distance between the points required for parabolic interpolation. Parabolic interpolation, then, is used in order to refine the initial estimate and calculate the precise location of each peak. Interpolation such as this requires three points to identify each peak, and can give an estimate to within 0.1% accuracy. This method is usually preferred as it gives more robust results.

4.1.5.2 Peak Matching

If the number of peaks were constant from frame to frame, the problem of matching the parameters estimated on one frame with those on a successive frame would simply require a frequency-ordered assignment of peaks. In practice, however, the locations of the peaks will change as the pitch changes, and there will be rapid changes in both the location and the number of peaks corresponding to rapidly varying signal regions, such as at harmonic to noise-like transitions. In order to account for such rapid movements in the spectral peaks, the concept of "birth" and "death" of sinusoidal components is introduced.

The problem of matching spectral peaks in some "optimal" sense while allowing for this birth-death process is generally a difficult problem. One method that has proved to be successful is to define sine-wave tracks for frequencies that are successively "nearest-neighbors". The matching procedure is made dynamic by allowing for tracks to begin at any frame (a "birth") and to terminate at any frame (a "death"), events which are determined when successive frequencies do not fall within some "matching interval".

Once peaks from each frame have been identified and recorded, and their instantaneous frequency and phase determined, the algorithm attempts to place them into a number of frequency tracks. Essentially, each track works by finding the peak in the next frame that is closest to its current value. In this way, the track is updated at each time interval and can show each partial's trajectory. If a partial in the current frame is not found to be a continuation of any partial in the previous frame, a new track is created. Likewise, if a partial cannot be found to continue a certain track after several frames, the track is killed. This is better illustrated in figure 4.2.

Peak trajectories are determined from both noisy and harmonic components of the waveform, to give a sinusoidal representation for the entire sound. Therefore, the model makes no distinction between harmonic and non-harmonic components. Also, when there are large changes in frequency between frames, the tracks may be confused and could jump to following other peaks that are now closer in frequency. As Serra [Serra and Smith, 1990] points out, this would be unsuitable when we want the trajectories to follow just the harmonic part of the sound, but as there is only one component type used (that of the time varying sinusoid) to represent the given input, this is the

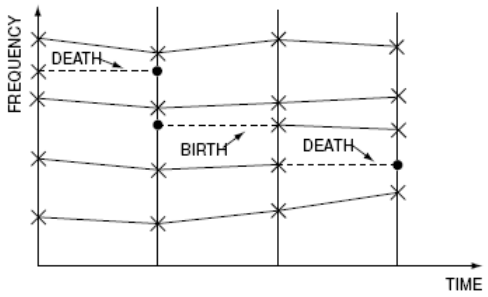

Figure 4.2: Depiction of the peak matching algorithm. The tracking of each sinusoidal partial is based on frequency-matching. Adapted from Quatieri and McAuley [Quatieri and McAuley, 2002].

only solution.

The result of the sinusoidal analysis for each spectral frame is a set of values that describe the partials contained in that frame. Each partial has a partial number, amplitude, frequency and phase value associated. For example, a frame where N partials are detected would contain the following.

$$\begin{array}{cccc} \text{partial number} & \text{amplitude} & \text{frequency} & \text{phase} \\ 1 & a_1 & f_1 & \phi_1 \\ 2 & a_2 & f_2 & \phi_2 \\ \vdots & \vdots & \vdots & \vdots \\ N & a_N & f_N & \phi_N \end{array} \tag{4.19}$$

where the first column contains the partial number, a_n is the amplitude of the n^{th} partial, f_n the frequency value, and ϕ_n the phase.

The sinusoidal model is useful as a starting point because it allows the extraction of several important characteristics of the input sound. Spectral shape, harmonicity and loudness are all easily identified in this representation, and the manipulation of these features is easily accomplished given a sinusoidal framework. But, while this model may be an improvement over the phase vocoder, it is still susceptible to some of the same shortcomings. For instance, it makes the assumption that every sound it models is composed of a number of (slowly-varying) sinusoids, each of which will vary no more than some given amount between frames. This is certainly not the case.

Many sounds have components that cannot be well-accounted for with sinusoids, with noise being one example. Attempting to model a sound which contains noise with a limited number of sinusoids sacrifices the fidelity of the sound. This is the major setback of this model. While it is possible to model noise components with a tracking vocoder using an unlimited number of sinusoids, the use of any sufficiently large number will require an even larger number of parameters to control them, which quickly makes this approach unwieldy.

Using fewer partials favors computation time and memory requirements, but it will introduce distortion and onset artifacts as non-harmonic components will not be well-represented. Spectral modeling synthesis proposes to solve this problem by representing the stable sinusoidal and noisy residual components independently.

4.1.6 Spectral Modeling Synthesis

Spectral modeling synthesis [Serra and Smith, 1990], hereafter referred to as SMS, was first proposed in the mid-eighties by Xavier Serra and Julius Smith. It grew out of the need for something more robust and flexible than previous systems such as PARSHL [Smith and Serra, 1985] could provide. Specifically, it addresses the need for a more robust representation of noisy components. The basic principle in SMS is that any sound may be said to be comprised of two components, a deterministic and a stochastic part.

By separating a sound into deterministic and stochastic components, a more flexible representation of the sound is created, which in turn facilitates modifications and changes to the sound itself. The SMS model closely mimics what is produced in a musical instrument or any other physical system. For instance, to produce pitched sounds, there must be some mode of vibration occurring. The deterministic or we could say harmonic and predictable component corresponds closely to this. Any other sound which is not accounted for by these primary vibrations, such as bow noise/breath noise, onset transients, etc, are modeled as residual data.

The flexibility this system affords is what attracts us to it. The deterministic component of SMS is based on a type of additive synthesis, where a number of time-varying sinusoids are used to model the harmonic spectrum of a given sound, as in PARSHL [Smith and Serra, 1985] and McAulay and Quatieri's model [McAulay and Quatieri, 1986] (see Section 4.1.5) and the stochastic part is modeled as filtered white noise.

In mathematical terms, this is expressed as

$$x(t) = \sum_{k=1}^{K} A_k(t) \cos \phi_k(t) + e(t) \tag{4.20}$$

where $A_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and phase of the k^{th} sinusoid, respectively, and $e(t)$ is the noise component [Serra and Smith, 1990]. The system then identifies peaks in the spectrum and places them into some number of trajectory-tracks as described in Section 4.1.5.2. In the context of SMS, these trajectories are referred to as guides and, as before, are used simply to organize partials in the input sound.

The main difference between this implementation and earlier systems is that it does not incorporate all of the selected peaks into guide layers. Doing so would attempt to fit spurious peaks resulting from noise data into the deterministic part of the model. However, SMS needs to separate these two components. It is for this reason that the peak-continuation algorithm in SMS is more sophisticated and geared solely to the deterministic part of the sound.

It is assumed that in a given input sound, the harmonic component is composed of quasi-stable, time-varying sinusoids. That is, partials which vary in frequency less than some user-predefined or expected amount. The peak-continuation algorithm places only these partials into guides, leaving unselected ones as part of the residual. Each guide is initialized by using the harmonic series of the detected pitch. This means that a partial is assumed to be close by to an integer multiple of the fundamental frequency.

In cases where no pitch can be detected, as with noisy inputs, guides are created dynamically when new peaks become available. As the guides progress forwards in time, the way each peak is assigned to them is determined by several factors. In harmonic sounds, since all the partials will evolve together, the fundamental frequency acts as the main control. In noisier signals, the system cannot rely on the fundamental for control, so the memory of each peak's trajectory is used to influence which partial provides the best fit to a given guide.

Depending on the signal's harmonicity, a combination of the previous trajectory and the current fundamental frequency is used to organize the peaks into guide layers. These guides then advance in time and look for the next set of partials to form trajectories. This provides a robust, yet

relatively simple representation of the deterministic component of the sound. What remains is to subtract this from the original input, and whatever remains can be deemed the residual, or stochastic component.

To perform the subtraction, we must first generate the deterministic component of the sound. Since the phase of each extracted partial is preserved, it is possible to simply add every sinusoid together and perform the subtraction from the original signal in the time-domain. However, this type of additive synthesis is computationally expensive. An efficient alternative would be to perform the subtraction in the spectral-domain using inverse-Fourier techniques [Rodet and Depalle, 1992]. In either case, the residual component must be characterized in the spectral domain.

The residual component can be approximated using the output of a time-varying filter, with white noise as the input. There are several ways in which one may characterize the filter. Perhaps the simplest, and certainly the most memory-efficient is with linear-predictive coding (LPC) analysis [Makhoul, 1975, Markel and Gray, 1976]. However, as the flexibility of the representation is forefront in design choices, another possibility is to approximate the spectrum using line segments. Here, the spectrum is divided up into a number of logarithmically-spaced sections, and the maxima in each is identified and connected. Using more points will increase the accuracy of the result, but is not strictly necessary as the gains in perceptual quality are minimal. Due to the flexibility of using line-segments to estimate the residual component, this method is often used over others such as LPC in this context.

SMS is a very good, robust, and general model that provides high-fidelity reproduction for a wide range of input sounds. Much work has been done with it in the area of sound morphing. However, the fact that it utilizes both a deterministic and a stochastic part makes the representation unwieldy. Components must be interpolated separately, and while this provides more possibilities for different, unusual types of interpolation, the simplicity of a single component type outweighs this novelty. Furthermore, the representation itself is far from perfect. As Serra himself points out [Serra and Smith, 1990], this model has problems with sounds that include noisy partials (for example, produced by a modulation). It would seem that, in practice, the assumed separation of the deterministic and stochastic components of a sound is rarely simple or clearly defined.

4.2 Interpolation of Sinusoidal Models

Most perceptually interesting sound transformations require processing techniques in the frequency domain rather than in the time domain because the ear is more sensitive to changes in the frequency domain. Serra states [Serra and Bonada, 1998] that an effective spectral representation should, ideally, provide high sound fidelity and flexibility, while minimizing memory consumption and computational requirements.

It should be noted that, among the many spectral models, variations in performance and the ability to accurately represent a given input can be attributed to how well the assumed model matches the process being analyzed. Therefore, to achieve good performance, we must carefully choose the appropriate analysis method and parameters to match.

Robust models can provide good performance for the generalized case. Therefore we need a representation of the sound in the spectral domain that is accurate, flexible, and robust to allow us to perform a wide variety of transformations. When the transformation in question is sound morphing, we must bear in mind that the interpolation principle dictates that we combine the parameters of the representation of the sounds used in the morph to obtain the representation of the morphed sound. The rest of this section is dedicated to the interpolation of sinusoidal models, while the rest of the chapter presents alternative models proposed in the literature to morph sounds.

4.2.1 Interpolation Procedure

When using sinusoidal analysis, the interpolation procedure consists in establishing correspondence between the two sets of partials and interpolating the amplitudes and frequencies of each pair using equation 2.1 or equivalently 2.4. The interpolated values constitute a new set of partials that define a frame of the morphed sound. After all the frames are processed, we simply resynthesize the morphed sound using the interpolated set of parameters.

The interpolation of frequency and amplitude values resulting from the sinusoidal analysis according to the interpolation principle introduced in chapter 1 implicitly assumes that, to be perceptually intermediate, the partials of the morphed sound should contain intermediate values of frequencies and amplitudes. Here we should notice that the phase value cannot be interpolated because it is not reasonable to expect a perceptually intermediate sound to have intermediate phase values. In other words, phase relations are nonlinear. Therefore, instead of interpolating the phase, we retrieve the phase values from the interpolated frequencies using equation 4.2.

In theory, we can achieve any of the transformations explained in chapter 3 with this rather straightforward interpolation procedure. The results can be perceptually convincing, as long as we take care of the temporal correspondence of frames before interpolating. Although the temporal correspondence problem is very important to achieve perceptually relevant results, establishing the correspondence between partials once two frames have been matched can be a surprisingly complicated problem, depending on the type of sound used in the morph. As we will see next, for harmonic sounds, correspondence is straightforward. For inharmonic sounds, however, the correspondence between partials can become a difficult problem to solve. Let us see why.

In part a) of figure 4.3 we see two sets of 4 partials and their partial numbers. One straightforward way of matching the partials one by one is by their partial numbers. In this case we could simply fade out (interpolate with zero) all higher partial numbers that find no match (if we suppose no correspondence between the number of partials). However, this simple algorithm does not take the values of the frequencies into account.

Using the partial number works if both sets of partials are (quasi-)harmonic and their fundamental frequencies are not too far apart (in terms of musical interval). The consequence can be a sweep in frequency when this is not the case. Let us imagine that one of the sounds being morphed is neatly harmonic (like a musical instrument sound), while the other is not, like a baby's cry. In this case, it might be perceptually more consistent to match partials whose frequency values are very close and fade all the others out, like illustrated by the dash-dot lines in part a) of figure 4.3.

Osaka [Osaka, 2005] dubbed this the optimal partner search problem and proposed an algorithm whose aim is to find the optimal solution to the problem of correspondence between two sets of partials derived from sinusoidal analysis with sound morphing in mind. The algorithm finds the optimal solution by minimizing the distance between the frequency intervals for all possible matches of partials (one-by-one). It is also possible to take the amplitudes of the partials into consideration. However, part b) of figure 4.3 shows that the solution can quickly become computationally intractable depending on the number of partials and on their positions on the frequency grid because of the combinatorial nature of the problem.

In part c) of figure 4.3 we see an alternative solution to the matching problem that avoids the combinatorial explosion. The idea is to divide the frequency grid into frequency bands and solve the problem inside each band. Part c) of figure 4.3 supposes that the bands are linearly spaced and that one of the sets of partials is harmonic, while the other is not. However, this solution can be successfully applied (and lead to more interesting results) when the frequency bands are perceptually related, such as the mel scale [Stevens et al., 1937], the bark scale or ERB.

Interpolation of sinusoidal modelling is amongst the most common approaches in the literature of sound morphing [Fitz et al., 2003, Fitz and Haken, 1996, Boccardi and Drioli, 2001, Hatch, 2004, Osaka, 2005, Osaka, 1995, Tellman et al., 1995, Williams and Brookes, 2007,

Williams and Brookes, 2009, Haken et al., 2006]. Tellman et al. [Tellman et al., 1995]. offer us one of the earliest descriptions of a morphing technique, which is based on a sinusoidal representation. The morphing scheme consists of interpolating the result of the Lemur [Fitz and Haken, 1996] analysis and involves time-scale modification to morph between different attack and vibrato rates.

More recently, Fitz [Fitz et al., 2003] presented a morphing technique also using a sinusoidal representation, and morphing is achieved again by simply interpolating the parameters of the model. Osaka [Osaka, 1995] also proposes to interpolate the parameters of sinusoidal analysis to morph between the sounds of musical instruments. Even though interpolation of sinusoidal models is a very popular approach to sound morphing, it is definitely not the only one found in the literature. Let us review some of them in the next sections.

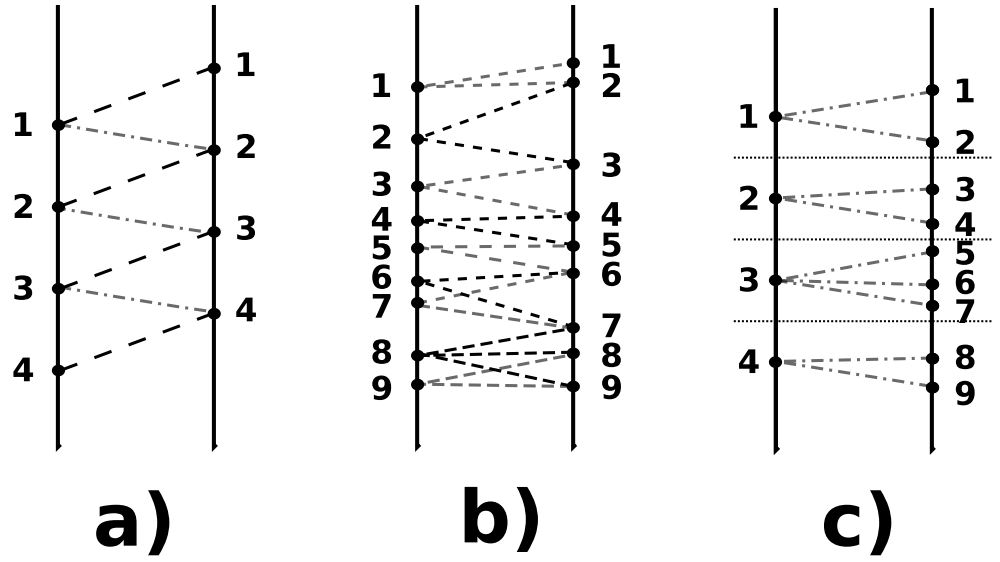

Figure 4.3: Optimal partner search problem. The figure illustrates different possible matches between pairs of partials in three situations. In part a) we suppose both sets of partials are harmonic, in part b) we suppose they are both inharmonic, and in part c) we suppose one of the sets is harmonic and the other inharmonic. Notice that this algorithm does not require one to one correspondence.

4.3 Other Approaches

I will briefly present alternative approaches to sound morphing that were proposed in the literature. Some of these approaches use a classic sound model and the novelty is restricted to the method for achieving the transformation. Others propose to model the sounds using a novel technique and morphing is achieved by simply interpolating the parameters of the proposed model.

4.3.1 Magnitude Spectrograms

Even though most previous work in sound morphing has used sinusoidal analysis, Malcolm Slaney and colleagues [Slaney et al., 1996] describe a creative approach to morphing in which they propose their own dedicated sound model and then they describe techniques based on magnitude

spectrograms. In this approach, sound morphing is accomplished by representing the sound in a multi-dimensional space that can be warped or modified to produce a desired result. After matching components of the sound, a morph smoothly interpolates the sound amplitudes to describe a new sound in the same perceptual space. Finally, the representation is inverted to produce a sound.

Figure 4.4 shows a block diagram of the approach proposed by Slaney et al. [Slaney et al., 1996]. It can be shown that this approach is similar to the source-filter model formalization of sound morphing, where the filter is the MFCC based spectral envelope, and the source is the residual.

Slaney states that, unlike image morphing, time is an important dimension of sound and it can be considered independently of the other sound dimensions. The morphs described here consider time separately from the other dimensions of the auditory signal. As will be shown, the separability of the temporal dimension simplifies all aspects of audio morphing.

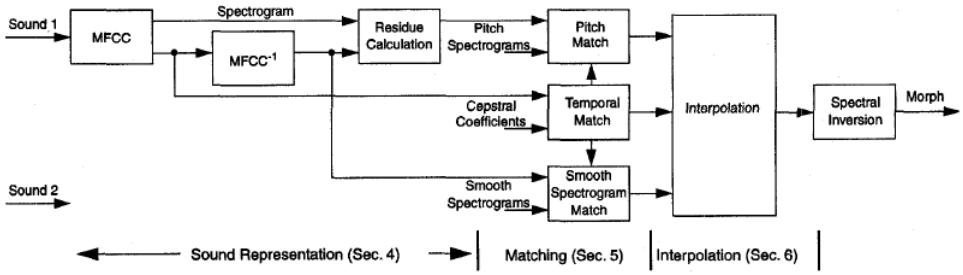

Figure 4.4: The three stages of audio morphing, representation, matching, and interpolation, are. shown. The signal path for representing sound 2 is not shown to simplify the drawing. After Slaney [Slaney et al., 1996]

4.3.1.1 Mel-Frequency Cepstral Coefficients

Conventional spectrograms can represent any sound, but cross-fading spectrograms does not produce convincing morphs when there are pitch changes because formants move with the harmonics and therefore simple scaling does not work [Slaney et al., 1996]. Slaney uses mel-frequency cepstral coefficients (MFCC) to separate the broad spectral characteristics of the sound from the pitch and voicing information. The MFCC coefficients are used in the initial temporal matching and to compute the smooth spectrogram. MFCC is computed by resampling a conventional magnitude spectrogram to match critical bands as measured by auditory perception experiments. After computing logarithms of the filter-bank outputs a low-dimensional cosine transform is computed.

In the approach proposed by Slaney et al. [Slaney et al., 1996], the MFCC representation is inverted to generate a smooth spectrogram for the sound. After applying the cosine transform again and undoing the logarithm we have a smooth estimate of the filter-bank output. The filter-bank output is then reinterpolated to get a spectrogram. The logarithmic transform and low quefrency cosine transform serve to filter out the pitch information in the spectrogram.

MFCC is good at modeling the overall spectral shape, but it doesn't include pitch. When we invert MFCC we get a rough approximation of the spectrogram, but without the pitch information. It would be nice if we could summarize all the information about pitch with a small number of scalars and then smoothly vary these numbers to get intermediate excitations. For example, we might use one number for the pitch and one to indicate the amount of voicing.

Unfortunately, this type of summarization is not sufficient as is seen in speech compression

systems. Simple LPC systems suffer from objectionable inaccuracies in the excitation. To provide acceptable reconstructions, a large codebook is needed to summarize the possible residues.

In sound morphing we use a spectrogram of the residue to code the pitch and voicing in the acoustic signal. A conventional shorttime spectrogram $S(\omega, t)$ encodes all the information in the signal and the smooth spectrogram $S_s(\omega, t)$ describes the overall spectral shape. Dividing the shorttime spectrogram, S, by the smooth spectrogram, S_s, gives us a "pitch" or residual spectrogram, $S_p(\omega, t)$, which describes the pitch and voicing information in the sound. The smooth and pitch spectrograms form the basis of their morphing technique. They recover the original spectrogram by multiplying the pitch and smooth spectrograms together.

4.3.1.2 Temporal Matching

Temporal matching is the requirement of correspondence in time domain. Matching is necessary so that we know which features of the first sound correspond to any particular feature of the second. Often a feature has moved and to affect a morph we need to slowly move the feature from where it is in the first sound to its position in the second.

There are many ways to perform the matching. Dynamic time warping and harmonic alignment are used to match features in audio morphing. Dynamic Time Warping (DTW) is used to find the best temporal match between the two sounds. Over the course of the morph, we want features that are common to both sounds to remain relatively fixed in time.

MFCC is often used in modern speech recognition systems as a distance metric and is used by Slaney et al. [Slaney et al., 1996] for the same purpose. Using DTW allows us to calculate the match between the two sounds so that later spectral stages have correspondence.

Sound morphs with different properties are created with different matching functions. In morphing between two versions of the same song, the melody is important. The temporal matching is done with a distance metric based on the dominant pitch. For other music (i.e. rap) we will want to consider the underlying rhythm.

4.3.2 Physical Models

Hikichi et al. [Hikichi, 2001] propose to use physical modeling to obtain morphed sounds between two different instruments. The approach is very straightforward because it simply proposes to interpolate the parameters of the two different physical models of the instruments.

Hikichi et al. state that, since the artificial instrument can have the same control parameters as the real one, the user can control its timbre more intuitively. They also say that most sound morphing proposals in the literature attack the problem from a signal processing point of view, especially using sinusoidal models. However, they conclude that the linear interpolation of the parameters of their physical model does not lead to perceptually linear morphed sounds. So they propose to construct MDS spaces using the source, target and morphed sounds to study how to warp the interpolation factor to obtain perceptually linear morphed sounds.

Naturally, this approach renders the results very difficult to obtain and to evaluate. Also, the warping function is model dependent and propably user dependent too, since it is subjective.

4.3.3 Gaussian Mixture Models

Boccardi et al. [Boccardi and Drioli, 2001] use Gaussian mixture models (GMM) to build the acoustic model of the source sound and then applies a set of conversion functions to transform from source to target sounds.

First they present how to use GMM to model musical instrument sounds. They state that when used to model the spectrum of a musical instrument sound such as a single sustained note, we may

say that the components of the GMM represent different portions of the sound (e.g., frames from the attack, the sustain, or the release portion).

However, depending on the data the model is trained with, it may represent the notes from the same instrument played with different intensities, or notes from different instruments, and so on. In other words, a conversion function which relies on this model is in principle able to classify the input sound frame to be transformed and to perform the transformation required for that frame.

Boccardi et al. use a sinusoidal plus residual model to represent the sounds, and focus on the transformation of the magnitude of the partials only. In other words, they do not model the differences of frequency and phase among the partials of the source and target sounds.

For this assumption to be considered reasonable, they also restrict the choice of the source and the target sounds to a set of compatible signals (e.g., morphing among piano notes with different spectral characteristics, morphing among sustained notes of wind or string instruments, etc.). Morphing is simply an interpolation of the parameters of the GMM representation of source and target sounds. They do not report on the perceptual impact of their method.

4.3.4 Wigner Distribution Analysis

Hope et al. [Hope and Furlong, 1997] propose a morphing algorithm based on the Wigner distribution analysis of the time-frequency contents of the sound instead of the traditional short-time Fourier transform (STFT). By basing such morphing tools on Wigner distributions of musical tones, rather than spectrograms, representational distortions would be minimized and detailed spectral and temporal features of real instrument tones would be made much clearer for the purposes of computational analysis and synthesis.

When attempting to develop a morphing algorithm, the intrinsic problems involved in the analysis of timbre must first be considered. That is, the sounds to be morphed must first be analyzed to determine those spectral and temporal features which characterize timbre. They make a good point about the representation, it is important to have a good representation of the sounds to be morphed in order to be able to blend the features encoded in this representation. Nonetheless, they compare representations of synthetic sounds using both transforms, and even hint at how this would impact when morphing musical instrument sounds.

Unfortunately, though, they do not present morphing results. It is not even clear if they suggest interpolating the parameters of the Wigner analysis of the sounds being morphed. Their main point is the comparison between the spectral resolution of both transforms. Due to the preliminary character of this work, the perceptual impact of such procedure would have to be investigated.

4.3.5 Neural Networks

Röbel [Röbel, 1998] describes a sound morphing technique that uses radial basis function (RBF) neural networks to model the sounds as dynamical systems and the morphing procedure consists in interpolating the attractors. He states that if the dynamical system, or the musical instrument, is observed by means of an output signal, the characteristics of this signal, or the sound, are closely related to the topology of the attractor.

In the context of real world sound signals the underlying system is always stable and, therefore, a stationary sound signal is always related to an attractor. The type of attractor depends on the musical instrument and its excitation. In most cases sounds obtained from chaotic attractors are considered noise and, therefore, are not used by classical musicians. Therefore, the use of musical instruments is often confined to periodic or quasi periodic attractors. However, musical instruments are not used with a stationary excitation. For slowly varying dynamics this situation

can be described by a system undergoing a parameter variation and, therefore, following a sequence of attractors.

He describes how to synthesize sounds from attractors and how to morph sounds based on homotopic mixing of dynamical systems [Röbel, 1998]. He presents morphing results using artificial synthetic sounds and two saxophone signals that have been played by a professional player with similar excitation and pitch difference of one half tone. Therefore he presents results of morphing across the pitch instead of timbre related dimensions of musical instrument sounds.

The morphing algorithm is then applied to obtain morphed saxophone signals for the set of fixed values of the interpolation factor α equally spaced between zero and one. He states that the results presented confirm that the attractors of the saxophone signals are smoothly transformed, while their topology is preserved. He does not comment on the perceptual impact of the method, though.

Finally, he ponders that if the model is operated in this regime, the large distance to the other units leads to a prediction function that is locally constant, which in turn produces the undesirable distortion. The problem has been addressed by formulating an additional constraint for the training algorithm, such that the width parameters of the RBF units are kept above a fixed value. With this constraint, he says, the distortions in the synthesized signals are no longer audible.

4.3.6 Wavelet Analysis

Ahmad et al. [Ahmad et al., 2009] propose to analyze the sounds with the discrete wavelet transform (DWT) and to further reduce the dimensionality of the representation by singular value decomposition (SVD). The morphing procedure is done by interpolation on the reduced SVD domain and synthesizing. They state that the perceptual impact is yet to be investigated.

Chapter 5

Hybrid Musical Instruments

The aim of this work is to morph musical instrument sounds across timbre dimensions to create the auditory illusion of hybrid musical instruments. We must understand the mechanisms underlying musical instrument sound perception to manipulate the features of musical instrument sounds that would create the desired effect. Therefore, this chapter revolves around the central concept of musical instrument sound perception, namely timbre. The concept of timbre is related to the subjective response to the perceptual qualities of sound objects and events [Handel, 1995]. The association between the term timbre and auditory object identification can be traced back to the origins of the word. In musical contexts, timbre is intrinsically associated with musical instrument identification [Fletcher, 1934]. Handel wrote that "we know that sound source identification is not reduced to waveform memorization because the intrinsic dynamic nature of the sources produces variations [Handel, 1995]." Nevertheless, the relationship between timbre and sound source identification is obscured by the myriad perceptual phenomena encompassed by the term timbre.

Timbre perception is very complex and not very well understood. In 1960, the American National Standards Institute – ANSI published a standard definition of timbre that was broadly accepted and adopted: "Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar [ANSI, 1960]." This definition implies that timbre, pitch and loudness are independent dimensions of sound perception. However, the perceptual dependence between timbre and those parameters is evident in some cases.

Timbre perception is inherently multidimensional, involving features such as the attack, spectral shape, and harmonic content. The recognition of musical instruments, for example, depends quite strongly on attack transients and on the temporal structure of the spectral envelope. The characteristic tone of a piano depends upon the fact that the notes have a rapid onset and a gradual decay. If a recording of a piano is reversed in time, the timbre is completely different. It now resembles that of a harmonium or accordion, in spite of the fact that the long-term magnitude spectrum is unchanged by time reversal.

The multidimensional nature of timbre perception makes it very difficult to isolate dimensions, associate verbal labels and find physical variables associated to these dimensions. A significant breakthrough in timbre research is due to the application of multidimensional scaling (MDS) techniques in musical instrument (dis)similarity experiments [Grey and Gordon, 1977, Krimphoff et al., 1994, Krumhansl, 1989, Handel, 1995, McAdams et al., 2005]. MDS techniques allow the investigator to propose a low dimensional space where the distances represent the perceptual (dis)similarity judgments. Figure 5.1 shows an example of such a space, usually referred to as timbre space. A consequence of the representation of musical instrument sound perception as timbre spaces is the possibility to associate acoustic correlates to each dimension. Physical variables

that present significantly high correlation with dimensions of timbre spaces are thought to capture the physical cues used in the mental representation of musical instruments. The features used in this thesis to guide the morphing transformation are acoustic correlates of salient dimensions of timbre spaces obtained in psychoacoustic studies.

One interesting characteristic of timbre spaces derives from the property of MDS techniques. MDS spaces can be chosen to be metric and orthogonal. The dimensions in orthogonal spaces are independent and the notion of distance is defined in metric spaces. A metric orthogonal timbre space would give independent acoustic correlates related to musical instrument timbre perception.

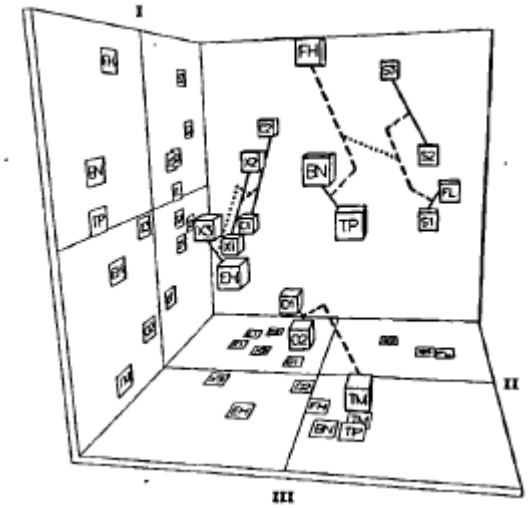

Figure 5.1: Example of multidimensional timbre spaces. After Grey [Grey and Gordon, 1977]

5.1 Timbre Perception

The classical model of musical sound proposed by the German physicist Hermann von Helmholtz in 1885 [Helmholtz, 1885] postulates that the amplitude of the vibration determines the force or loudness, and the period of vibration the pitch. Helmholtz concluded that quality of tone can therefore depend upon neither of these. The only possible hypothesis, therefore, according to Helmholtz, is that the quality of tone should depend upon the manner in which the motion is performed within the period of each single vibration. However, Helmholtz wondered to what extent the differences of musical quality can be reduced to the combination of different partial tones with different intensities in different musical tones. According to Helmholtz, in his time there was a general inclination to credit quality with all possible peculiarities of musical tones that were not evidently due to loudness and pitch. This was correct to the extent that quality of tone was merely a negative conception. What Helmholtz means by negative conception is that timbre was defined as what it is not, the qualities other than pitch, loudness, and duration. However, later studies [Seashore, 1938] showed that the picture was not so simple and that there are timbral variations associated with pitch changes (such as the registers of the clarinet) and loudness (such as brassy trumpet sounds).

If timbre is the characteristic of perception that allows the listener to identify the instrument that played the sound, how can there be timbral variations in one instrument due to, for example,

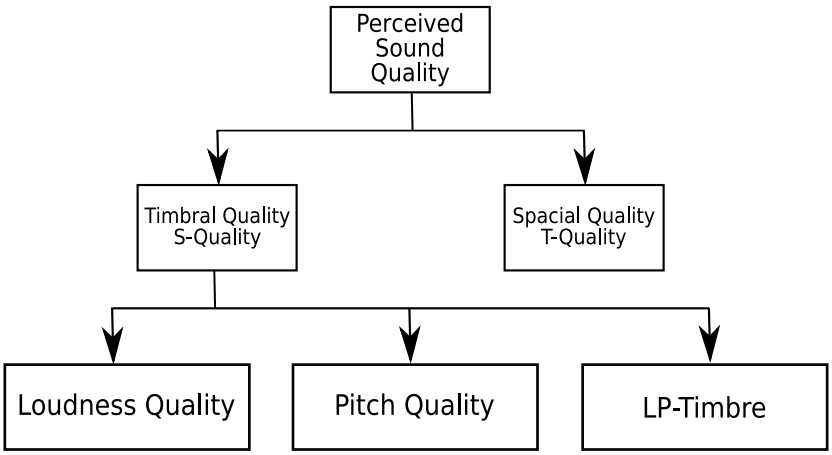

Figure 5.2: Timbre, tone color, and sound quality. The figure shows the hierarchical classification of sound quality that includes sound color and timbre as subsets. Adapted from [Letowski, 1992]

loudness? These variations exist, so in 1938 Carl Seashore says that the term must be refined. Seashore [Seashore, 1938] conjectures that tone quality has two fundamental aspects, timbre and sonance. Physically the timbre of the tone is a cross section of the tone quality, while sonance is the pattern of change in timbre. This implies that in order for sonance to exist, a sound must last long enough for patterns of change to be established.

Bregman [Bregman, 1990], on the other hand, states that the ANSI definition of timbre

> "... is, of course, no definition at all. For example, it implies that there are some sounds for which we cannot decide whether they possess the quality of timbre or not. In order for the definition to apply, two sounds need to be able to be presented at the same pitch, but there are some sounds, such as the scarping of a shovel in a pile of gravel, that have no pitch at all. We obviously have a problem: Either we must assert that only sounds with pitch can have timbre, meaning that we cannot discuss the timbre of a tambourine or of the musical sounds of many African cultures, or there is something terribly wrong with the definition."

Krumhansl [Krumhansl, 1989] points out that one of the major difficulties in finding a definition of timbre is generalizing the notion of timbre beyond the set of traditional orchestral instruments. Pratt and Doak [Pratt and Doak, 1976] refine the ANSI definition of timbre, proposing that

> "Timbre is that attribute of auditory sensation whereby a listener can judge that two sounds are dissimilar using any criteria other than pitch, loudness or duration."

Many authors advocate to abandon the term timbre since the classical definition is inadequate, yet widely accepted. Slawson [Slawson, 1985] proposes the term sound color to refer to a specific subset of sound qualities of musical instrument sounds and speech vowels. In the light of the source-filter model of sound production, sound color is associated with the filter. Letowski [Letowski, 1992], on the other hand, proposes a hierarchical classification of sound quality that includes timbre as subsets of sound qualities. Figure 5.2 shows a schematic of sound quality assessment according to Letowski.

Letowski proposes to differentiate between the terms timbre and sound quality using a the X-timbre system of timbre subspaces based on the concept of “residual dimension”. The concept

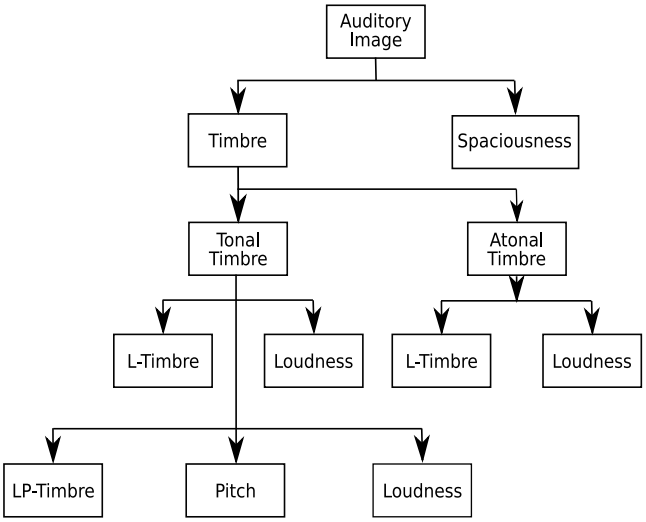

Figure 5.3: Basic elements of a hierarchical model of the auditory image based on the concept of "residual dimensions" called X-timbre. Adapted from [Letowski, 1992]

of "residual dimension" can be understood as follows. If differences along one dimension of timbre, such as loudness, dominate overall perception perception of sound and are not a desired object of assessment, such differences can be equalized, i.e., the said dimension can be excluded from assessment. After such a dimension is excluded from consideration, the auditory images are projected to $(N-1)$-dimensional space, and the differences among the images along the remaining dimensions of perceptual space become more pronounced. Letowski states that this operation can be extended on more than one dimension. The resulting $N-k$-dimensional space, where k is the number of dimensions removed from consideration, is called "residual space".

A basic descriptive model of an auditory image that utilizes the X-timbre system of labels is shown in figure 5.3. Letowski states that "the X-timbre method of labeling the timbral subspaces is very flexible and cannot be outgrown. He says it constitutes a solid frame of reference needed for the future development of more specific attributes of auditory image which are necessary for parametric sound assessment" [Letowski, 1992].

5.1.1 Timbre Revisited

The Helmholtz model of musical sound represents the most significant research work in musical acoustics in the XIX century. Since then, researchers have determined a more detailed model of musical instrument sounds. Digital recording allowed the investigators to show that the waveform, and hence the spectrum, change radically during the course of a sound.

In his "Computer study of trumpet tones [Risset, 1966]" Jean-Claude Risset discovered that each partial of a sound possesses a unique temporal amplitude envelope. This clearly contrasts with the Helmholtz model, where all partials present the same temporal envelope. Risset concluded the following regarding his studies: the spectrum of trumpet sounds is close to, but not perfectly harmonic. The higher harmonics become richer as we go up the intensity scale (dynamics). There is a rapid low amplitude quasi-random fluctuation in frequency. Higher partials attack later and take longer to reach maximum amplitude during attack. Finally, Risset detected a formant peak

around 1500 Hz. In subsequent studies Risset observed that the temporal evolution of the spectrum of trumpet sounds plays a fundamental role on the characteristic sound quality of the instrument.

Other researchers, following Risset's steps, systematically used the computer to analyze the temporal evolution of the spectrum of a variety of instrumental sounds. James Moorer and John Grey published computer analyses showing the temporal evolution of the spectrum of several instruments, including the violin [Moorer and Grey, 1977a], the clarinet and the oboe [Moorer and Grey, 1977b], and the trumpet [Moorer and Grey, 1977c]. Apart from the amplitude progression, the analyses determined the frequency variation of each partial. The frequency of each partial fluctuates during the course of the sound. This variation can be rather erratic during the attack portion. Resynthesis of the sound without the fluctuations in frequency produced a perceived difference in sound quality.

Grey [Grey and Gordon, 1978] investigated if simplifications/alterations of the temporal progression of the spectrum are perceptually detectable. Grey approximated the amplitude envelope of the partials of musical instrument sounds by straight line segments, and resynthesized trumpet sounds using the simplified data which were judged virtually identical to the original sounds. Grey concluded that these slight variations in the amplitude envelope of the partials do not contribute significantly to the perception of timbre.

When a a set of spectral components is presented to a listener, they might fuse into a single percept. One of the determinant factors is the onset asynchrony, which refers to the differences in onset times of partials. Fluctuations in the frequency of the partials are also necessary to give the perception of fusion. The perception of timbre involves correlating a number of factors, including the nature of the attack, the harmonic contents, and the tuning of the partials. To a certain extent, the loudness, pitch, and temporal aspects contribute to timbral characterization. Different researchers have suggested sets of independent dimensions for timbre classification. John Grey [Grey, 1975, Grey and Gordon, 1977] studied timbre using a set of 14 sounds from different musical instruments equalized in pitch, loudness, and duration, isolating the perception of dissimilarity to dimensions of sounds independent of pitch and loudness according to the classical definition of timbre. Then he presented the possible pairs of tones to listeners, asking them to rate the perceived dissimilarity between each pair. Finally, he applied a multidimensional scaling (MDS) technique to construct a three dimensional space that maps the dissimilarity judgment into distances. The result, shown in figure 5.1, is that sounds that are close in this space are perceived as similar in timbre.

5.1.2 Timbre Spaces

Since the pioneering work of Helmholtz [Helmholtz, 1885], MDS techniques figure among the most prominent when trying to quantitatively describe timbre. McAdams [McAdams et al., 2005] and Handel [Handel, 1995] independently propose comprehensive reviews of the early timbre space studies. Grey [Grey, 1975, Grey and Gordon, 1977] investigated the multidimensional nature of the perception of musical instrument timbre, constructed the three-dimensional timbre space shown in figure 5.1, and proposed acoustic correlates for each dimension. He concluded that the first dimension corresponded to spectral energy distribution (spectral centroid), the second and third dimensions were related to the temporal variation of the notes (onset synchronicity).

Krumhansl [Krumhansl, 1989] conducted a similar study using synthesized sounds and also found three dimensions related to attack, synchronicity and brightness. Krimphoff [Krimphoff et al., 1994] studied acoustic correlates of timbre dimensions and concluded that brightness is correlated with the spectral centroid and rapidity of attack with rise time in a logarithmic scale. McAdams [McAdams et al., 2005] conducted similar experiments with synthesized musical instrument timbres and concluded that the most salient dimensions were log rise time, spectral centroid and degree of spectral variation.

More recently, Caclin [Caclin et al., 2005] studied the perceptual relevance of a number of acoustic correlates of timbre-space dimensions with MDS techniques and concluded that listeners use attack time, spectral centroid and spectrum fine structure in dissimilarity rating experiments. McAdams [McAdams et al., 2005] and Caclin [Caclin et al., 2005] proposed acoustic correlates for each dimension they found in their timbre spaces, which capture perceptual aspects of musical instrument sounds.

5.1.2.1 Acoustic Correlates of Timbre Dimensions

Therefore, the features used to guide the morphing transformation are important because of their perceptual correlation. The acoustic correlates of timbre dimensions presented above used as features allow to monitor the perceptual impact of the manipulation of sounds. For example, the spectral envelope is important in timbre perception [Slawson, 1985]. LPCs are parameters that describe the spectral envelope but carry little information about how a sound whose envelope is described by them is perceived. The spectral centroid of the same spectral envelope, on the other hand, carries important perceptual information about the sound because it is directly related to one of the three most salient dimensions of timbre spaces. The spectral centroid is said to capture the perceptual dimension usually labeled brightness [Schubert and Wolfe, 2006]. Schubert [Schubert and Wolfe, 2006] compared two models that predict perceived timbral brightness in terms of the centroid of the frequency spectrum and concluded that brightness is much better correlated with frequency spectrum centroid than with the ratio of the centroid of the frequency spectrum to the fundamental frequency.

We should point out that most timbre spaces obtained with MDS techniques posses two important properties, they are orthogonal and metric. Orthogonality means that the dimensions are independent, i.e., the perception of attack-time, for example, is independent from the spectral centroid. In a metric space the notion of distance between elements is defined. This means that a sound that is perceived as twice as bright as another one will be twice as far from the reference.

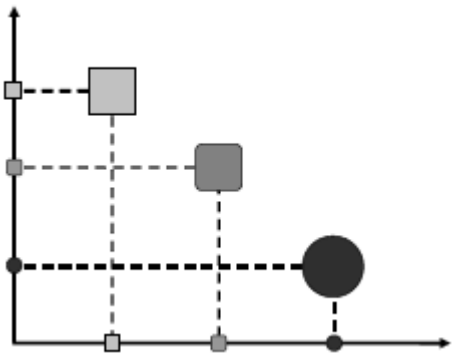

Figure 5.4: Orthogonal metric space. The figure illustrates an orthogonal metric space and shows two different sound objects as distinct points in this space. The projections of these sound objects on the axes represent the values of their features.

Figure 5.4 illustrates this idea with a two-dimensional abstraction of an orthogonal metric timbre space. Each dimension of this space would correspond to a dimension of timbre perception. Sonic features that are correlated to these dimensions capture aspects of timbre perception related to this abstract timbre space. Two perceptually different sound objects (the circle and the square) are represented in this space as two distinct points, as shown in figure 5.4 . The projections of

the points onto the axes of the space represent the different values of the features associated with each sound object. The sound object represented by the square with rounded corners is supposed to be perceptually intermediate between the circle and the square. Perceptually intermediate objects should occupy intermediate positions in this space, and therefore have intermediate values of features. Because the space is metric, if the intermediate sound is perceived as exactly halfway between the other two with respect to a dimension, its corresponding feature value will also be halfway between the other sounds'. A direct consequence is that sounds that are positioned linearly between two in such space have linearly varying feature values.

When morphing sounds, a perceptually intermediate sound should be positioned between source and target in the underlying timbre space, such that it has intermediate values of features. A morphed sound whose descriptors are halfway between source and target should be perceived in the middle of the two regarding the features captured by the descriptors. This thesis uses temporal and spectral features to guide the morphing transformation. The temporal features used in this work are log attack time, transition time, sustain time, release time, temporal centroid. The spectral features are the spectral shape descriptors, namely spectral centroid, spectral spread, spectral skewness, and spectral kurtosis. The spectral shape features are a measure of the distribution of spectral energy. The next section introduces the features used as guides in this work and presents how to calculate them.

5.2 Features

Many different types of acoustic signal features have been proposed for the task of sound description [Herrera et al., 1999]. Some of them come from the speech recognition community [Rabiner, 1993]. Others derive from previous studies on musical instrument sound classification [Foote, 1997, Scheirer and Slaney, 1997, Brown, 1998, Serra and Bonada, 1998, Brown, 1999, Jensen, 2001, Peeters and Rodet, 2002, Peeters and Rodet, 2003, Peeters, 2003, Martin and Kim, 1998, Wold et al., 1996] and from the results of psychoacoustical studies [Krimphoff et al., 1994, Misdariis et al., 1998, Peeters et al., 2000]. A systematic taxonomy is outside of the scope of this work, nevertheless, we can distinguish features according to four different points of view, the time scale, the time extent, the perceptual relevance, and the extraction process.

1. The time scale reflects how long the time support of the feature is. In other words, it says if the feature is an instantaneous value because it was calculated on a frame of the STFT, or a global value for the whole sound, such as the average of instantaneous features or a description of some global attribute like total duration or attack time. Global descriptors are computed once for the whole signal because they describe a feature related to an event that only happens once or their meaning is associated with the entire duration of the signal. Examples are attack time, temporal centroid, total duration, among others. Instantaneous descriptors are computed for each frame of the STFT, therefore they are time-localized and their values are considered relatively stable during the duration of the frame. An example would be spectral features such as the spectral centroid, that can vary between frames. The description of the variation of the values of the instantaneous descriptors itself can be considered a spectro-temporal descriptor.

2. The time extent reflects the fact that some descriptions apply only to part of the signal, for example the attack time, while others apply to the whole signal, like the loudness of a sound.

3. The perceptual relevance reflects the correlation between the feature value and a dimension of sound perception. For example, the fundamental frequency is highly correlated to pitch perception.

4. The extraction process of the feature basically describes where in the feature extraction flowchart shown in 5.5 the feature is calculated. The calculation of most features is rarely done on the sound signal and usually depends on the estimation of parameters of a model of the sound signal, such as temporal envelope, spectral envelope, sinusoidal modeling, and even more complicated perceptual models [Peeters, 2004]. So we can further distinguish among temporal, spectral, energy, harmonic or perceptual features.

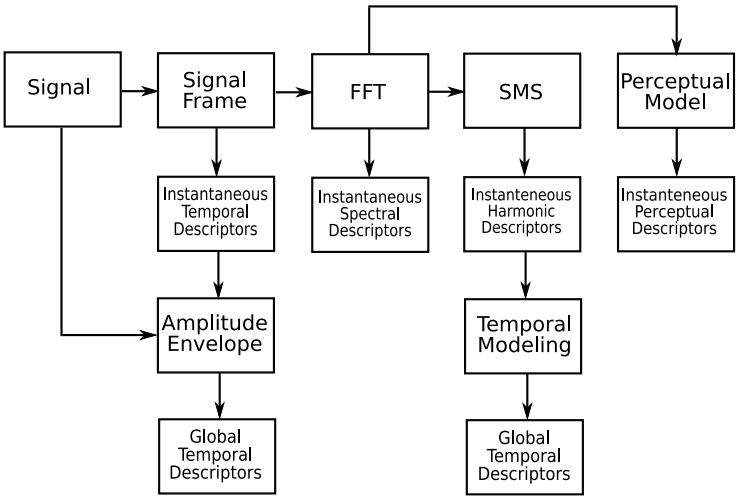

Figure 5.5: Descriptor extraction flowchart depicting the general descriptor extraction scheme.

Temporal features are usually calculated directly on the sound signal, such as the attack time or effective duration, but some temporal shape features might need some pre-processing, like the estimation of the temporal envelope. Temporal features are usually global because they apply to the whole sound signal. Spectral features are calculated on the spectrum of each frame of the STFT, and as such are usually instantaneous.

An example would be the spectral shape features used in this work. Perceptual features, on the other hand, may be either global or instantaneous because they are computed on a model of the human hearing process, such as a loudness model or a filter that mimics the response of the middle ear.

As 5.5 shows, most features need some estimation of parameters of the sound, such as the temporal or spectral envelope, or sinusoidal modeling for harmonic features. Most importantly, some features depend on a model such as the mel scale or the frequency response of the middle ear, explained in more detail in the next section.

5.2.1 Temporal Features

Here we estimate features related to the temporal evolution of spectral events. In other words, we are interested in characterizing the evolution of the spectral features of the sound in time, and estimate parameters that describe them. Usually, we segment the temporal evolution of musical instrument sounds into regions such as attack, steady state or sustain, decay, and release. The boundaries between these regions is blurred and not all sounds present all of them at the same time.

5.2.1.1 Temporal Centroid

The temporal centroid is the measure of the balance of energy distribution along the course of a tone and is calculated as follows

$$\tau = \frac{\sum_t ta(t)}{\sum_t a(t)} \tag{5.1}$$

where τ represents the temporal centroid, t is time, and $a(t)$ is the value of the amplitude envelope for time t. The temporal centroid has been shown [Skowronek and McKinney, 2006] to be especially important when comparing percussive and sustained sounds because that is when it varies more significantly, allowing us to distinguish between the two classes. Still, in the context of strictly sustained sounds, the attack times and temporal centroids vary significantly enough to be relevant

5.2.1.2 Log Attack Time

The attack is present in all sounds and psychoacoustic (dis)similarity studies discovered that it is one the most perceptually salient features of musical instrument sounds. Several studies [Caclin et al., 2005, Krimphoff et al., 1994, Krumhansl, 1989, Grey and Gordon, 1977, Handel, 1995, McAdams et al., 2006, McAdams et al., 2005] have shown that the attack time is perceived roughly on a logarithmic scale, like pitch and its counterpart fundamental frequency. This means that in order for a listener to perceive linear increments, the stimulus must be multiplied by the same factor. The log attack time (lat) is calculated as shown in 5.2, where at_1 stands for the beginning of the attack and at_2 for the end.

$$lat = \log(at_2 - at_1) \tag{5.2}$$

5.2.1.3 Length of Other Segments

I propose to estimate the length of all the other segments considered, namely, transition, sustain, decay, and release, without any warping. Caclin et al [Caclin et al., 2005] propose different warping formulas for each perceptually salient region considered in their study. The warpings used by Caclin et al. were derived by calibrating a perceptual experiment using synthetic tones. This work morphs recordings of acoustic musical instruments, which are much more complex than synthetic sounds, thus there is no reason to apply the warping functions proposed by Caclin et al. for synthetic sounds. Chapters 8 and 12 present respectively the model of the temporal events used in this work, and how to estimate the duration of each one of them.

5.2.2 Spectral Features

The spectral shape features are calculated on every frame, which permits to follow their temporal variation. Chapter 11 will show plots of the waveforms of musical instrument sounds used in this thesis along with the temporal variation of the spectral shape features.

5.2.2.1 Spectral Shape Descriptors

Spectral shape features measure the balance of the distribution of spectral energy. The spectral centroid is correlated with dimensions of timbre space obtained with MDS. Spectral skewness and kurtosis were shown to be significantly correlated with 2 out of 27 dimensions of 10 timbre spaces tested in a study of acoustic correlates of timbre dimensions [McAdams et al., 2006]. Among the many spectral features used in audio retrieval and classification , the spectral shape descriptors that

stand out among the most relevant are the standardized moments, calculated as if the magnitude spectrum were a probability distribution. So we define the frequencies k as the possible outcomes and the probabilities to observe them are given by

$$p(k) = \frac{|X(k)|}{\sum_k |X(k)|} \tag{5.3}$$

which is the normalized magnitude spectrum, such that the spectral moments become

$$\mu'_m = E\left[(p(k))^m\right] = \sum_k k^m p(k) \tag{5.4}$$

Following this definition [Peeters, 2004], the spectral shape descriptors are defined as the first four standardized moments of $p(k)$

5.2.2.2 Spectral Centroid

Spectral Centroid: it is the "center of mass" or barycenter of the spectrum and is related to the "brightness" of a sound. It is defined as the mean of $p(k)$.

$$\mu = \sum_k k p(k) \tag{5.5}$$

Notice that when we substitute $p(k)$ given in equation 5.3 into equation 5.5, this is equivalent to the definition in equation 8.2.

5.2.2.3 Spectral Spread

The spectral spread measures the spread of the spectrum around its mean value and is defined as the variance of $p(k)$.

$$\sigma^2 = \sum_k (k-\mu)^2 p(k) \tag{5.6}$$

The description of the magnitude spectrum with the spectral centroid and spread are analogous to using the center frequency and bandwidth to describe a formant peak of the spectrum (the obvious difference is that the spectral shape features are calculated on the whole range of frequencies of the magnitude spectrum).

5.2.2.4 Spectral Skewness

The spectral skewness measures the asymmetry of a distribution around its mean value. It is defined as the third standardized moment.

$$\gamma_3 = \frac{\sum_k (k-\mu)^3 p(k)}{\sigma^3} \tag{5.7}$$

As shown in figure 5.6, zero skewness indicates a symmetric distribution, positive skewness indicates more energy on the left, and negative skewness indicates more energy on the right. Naturally, the spectral distribution of most natural sounds tends to have positive skewness to reflect the fact that the spectral energy is concentrated on the lower frequency end of the spectrum.

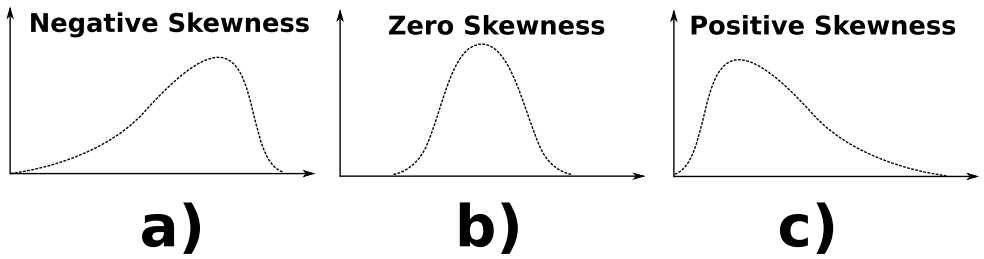

Figure 5.6: Skewness. The figure shows three distributions and their associated values of skewness.

5.2.2.5 Spectral Kurtosis

The spectral kurtosis gives a measure of the "peakdness" of $p(k)$. It is measured as the fourth standardized moment.

$$\gamma_4 = \frac{\sum_k (k-\mu)^4 p(k)}{\sigma^4} \tag{5.8}$$

The kurtosis of the Gaussian distribution is equal to 3 and is the reference value. So a normal distribution has kurtosis 3, a distribution flatter than the normal distribution has kurtosis smaller than 3, and a more "peaky" distribution has kurtosis greater than 3, as illustrated in figure 5.7.

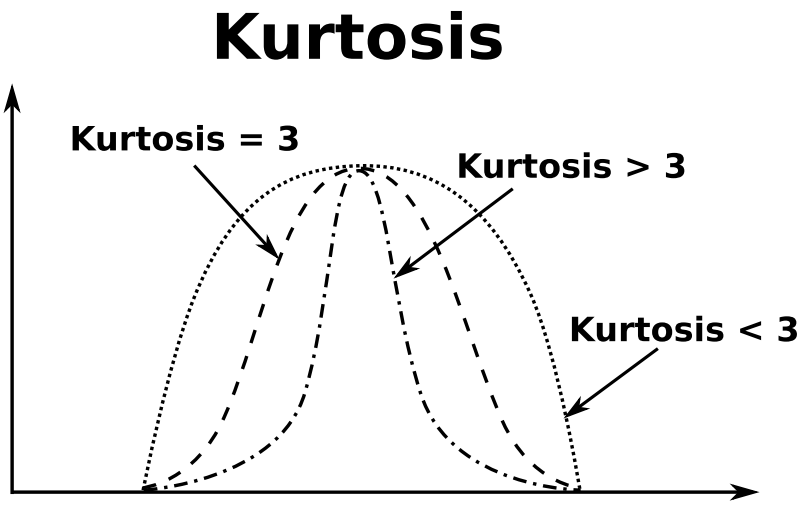

Figure 5.7: Kurtosis. The figure illustrates distributions with different values of kurtosis.

5.2.3 Calculation of Features

The spectral shape features can be calculated on different frequency and amplitude scales. Particularly, we know that the logarithmic frequency and amplitudes in dB (decibels, also logarithmic) are scales that are somewhat related to the perception of frequencies and their respective amplitudes. Thus we note that the spectral shape features can be calculated on the "perceptual spectrum" as

shown in figure 5.8. Figure 5.8 shows the magnitude spectrum expressed in linear amplitude and frequency scales on the left, the filter that simulates the response of the middle ear in the middle, and the "perceptual spectrum" on the right. The "perceptual spectrum" is calculated as follows.

1. The frequency and amplitude scales of the magnitude spectrum are warped;

2. The "mid ear filter" is applied to the warped magnitude spectrum.

The result of the operations described above is the perceptual magnitude spectrum shown in the figure. The amplitude is expressed in decibels (dB). The map from linear frequency in Hertz to mel frequency in Hertz is the following

$$\begin{aligned} f_m &= \frac{f_l}{f_b C} & f_l < f_b \\ f_m &= C\left[1 + \log\left(\frac{f_l}{f_b}\right)\right] & f_l > f_b \end{aligned} \tag{5.9}$$

where f_m is the frequency in mel, f_l is linear frequency, f_b is the linear breakpoint of the mel frequency, and C is the scaling factor of the normalized mel scale. The breakpoint of the mel frequency f_b is a parameter of the conversion. Usually 1000 Hz is considered reasonable. C is selected such that the center bin of the linear frequency vector translates into the center bin of the normalized mel scale. This gives

$$C = \frac{N/2}{\left(1 + \log\left(\frac{SRN^2}{2f_b}\right)\right)} \tag{5.10}$$

where SR is the sampling rate, and N is the number of frequency bins.

Finally, the spectral shape descriptors are calculated using the mel frequencies and amplitude values in dB. The result is the distribution of spectral energy calculated on the perceptual spectrum.

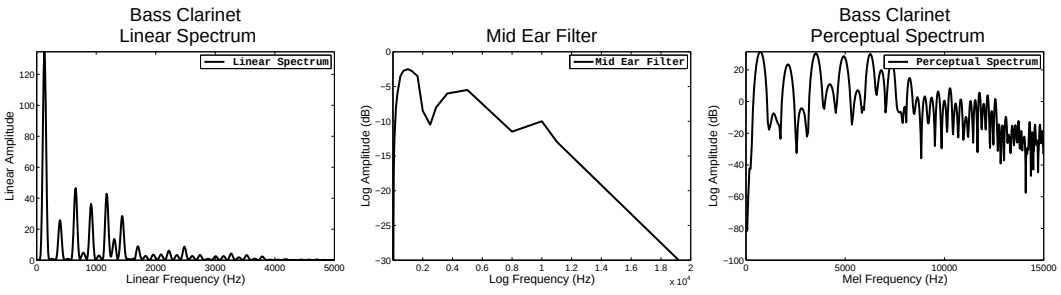

Figure 5.8: Perceptual Spectral Shape Descriptors

We should notice that the widely used mel frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980] are calculated in a similar way, using perceptually motivated models. But the coefficients encode information about the spectral envelope without explicitly using the actual curve, which we need to calculate the spectral shape descriptors. In their approach to sound morphing, Slaney et al. [Slaney et al., 1996] use an MFCC based spectral envelope to describe the overall shape of the spectrum, without pitch information. Terasawa [Terasawa et al., 2005] proposes to use 13 MFCCs to describe the color of a spectrum, or its spectral shape, instead of spectral shape descriptors. Chapter 13 presents a derivation of the analytic relationship between cepstral coefficients and the spectral shape descriptors developed in the context of this thesis.

5.2.3.1 The Mel Scale

Stevens, Volkmann and Newman [Stevens et al., 1937] and later Stevens and Volkmann [Stevens and Volkman, 1940] established a scale relating perceived pitch to frequency that they called the mel scale (short for melody scale). The mel scale encompasses the entire range of audible frequencies and it is neither strictly logarithmic as the musical-pitch scale, nor is it linear. Interestingly, it correlates well with a variety of other psychophysical and some physiological measures, including the relation of frequency to the distance of the point of maximum excitation along the basilar membrane. Among the methods used to derive the mel scale was the "fractionation" method, in which the listeners were asked to set the frequency of one sinusoid to a pitch that was some prescribed fraction of the pitch of another standard tone. The tone under the listener's control was to be set to one-third of the pitch of the standard, for example. Musicians, used to think in terms of octave equivalence and musical intervals, are usually hard-pressed to imagine how they might respond to such a task; however, the results from the few musicians who participated in the experiments did not differ significantly from those of non-musicians.

The mel scale was said [Stevens et al., 1937] to pertain to one aspect of musical pitch, its height, as distinguished from its pitch class or chroma. The distinction between these two aspects of pitch is common in Western music, and it seems to hold cross-culturally. It is perhaps clearest in the theory of atonal and dodecaphonic music. The 12-tone row or a smaller pitch set controls the chroma, whereas the height - the register - is chosen according to other considerations. A provocative and difficult question is whether composer's registral choices have in some fashion reflected the mel scale instead of the more obvious log-frequency scale. The interest is to investigate whether the mel scale serves as a measure of perceptual distance in pitch.

An alternative interpretation of the mel scale follows from the possibility that listeners in the original mel scale experiments were not judging a dimension of ordinary auditory sensation at all. Sine waves, after all, do not occur in nature, and the simplicity of their mathematical specification is not reflected in the sensations to which they give rise. Listeners in the mel-scale experiment may have been forced by the poverty of the acoustical stimulus to listen in a "reduced" manner. When faced with a sound that had neither genuine pitch nor color, listeners gave responses that, in effect, directly reflected some measure of distance along the basilar membrane and not the higher levels of auditory analysis that must underlie both pitch and color determinations in natural sounds. According to Slawson [Slawson, 1985], if this interpretation is correct, the mel scale pertains to an auditory process in the cochlea that underlies both pitch and sound color.

Part II

Estimation and Representation of the Source-Filter Model Parameters

Chapter 6

The Source-Filter Model

In this chapter the source-filter model of sound production will be presented. Historically, the source-filter model was developed to explain the mechanisms of speech production. The source–filter model of speech production models speech as a combination of a sound source, the vocal cords, and a linear acoustic filter, the vocal tract (and radiation characteristic). According to this model, speech is viewed as the result of passing a glottal excitation signal (source) through a time-varying linear filter that models the resonant characteristics of the vocal tract.

The development of the model is due, in large part, to the early work of Gunnar Fant [Fant, 1960]. While only an approximation, the model is widely used in a number of applications because of its relative simplicity. An important assumption that is often made in the use of the source-filter model is the independence of source and filter. Thus speech production is the result of a special interaction between source and filter in which the filter changes the source imprinting its resonant characteristics on its spectrum, but the frequencies of the partials are not affected by the interaction.

The assumption of independence between source and filter partially explains why the source-filter model is more rarely applied to model acoustic musical instruments, whose source and filter are strongly coupled. The coupling between source and filter for musical instruments can be easily understood as and interaction between source and filter in which the filter drives the source. In other words, the filter imposes not only its resonant characteristics to the source, but also the frequencies at which the source will resonate. In more musical terms the filter tunes the fundamental frequency or the pitch of the source.

Nevertheless, there are some special conditions under which the source-filter model can be applied to model the production of acoustic musical instrument sounds. Notably, when the sounds we are interested in have the same pitch (or fundamental frequency), to eliminate the dependency of the pitch from the model. These conditions will be presented in depth in this chapter, but we will first review the original source-filter model of speech production and adapt it to acoustic musical instruments.

The approach adopted in this chapter draws parallels between speech and musical instrument sounds, adopting techniques specifically developed for speech in a musical instrument sound model that can be used in morphing. For instance, the production of sound by the vibrating lips inside a mouthpiece is a very complex problem that is, in practice, closely analogous to the production of sounds by the vocal folds. So, this chapter presents the basic aspects of the source-filter model of speech production and explains how we can apply it to musical instrument sounds.

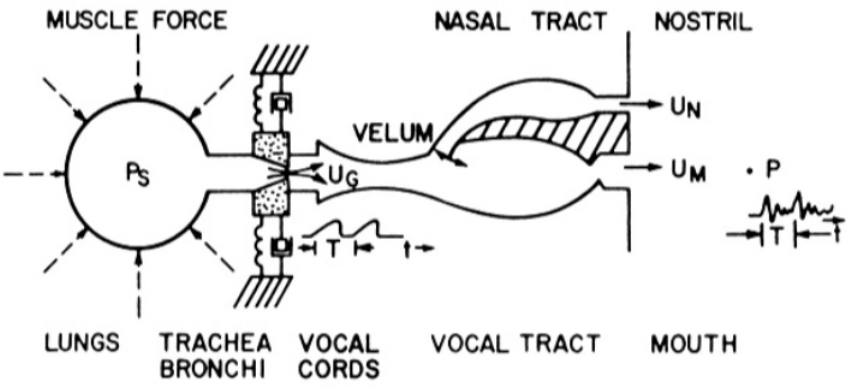

Figure 6.1: Schematized diagram of the vocal apparatus. After Rabiner [Rabiner and Schafer, 1978]

6.1 The Source-Filter Model for Speech

In studying the speech production process, it is helpful to abstract the important features of the physical system in a manner which ultimately leads to a realistic yet tractable mathematical model. Figure 6.1 shows a physically related schematic diagram of the vocal system. For completeness the diagram includes the sub-glottal system composed of the lungs, bronchi and trachea, a mechanical model of the vocal cords, including mass, spring and damping components, and a variable area set of tubes that model the vocal tract configuration. The sub-glottal system serves as a source of energy for the production of speech. The mechanical model of the vocal cords provides the excitation signal for the vocal tract. The resulting speech signal is simply the acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is shaped accordingly by the (time varying) vocal tract.

The vocal tract and nasal tract are shown in figure 6.1 as tubes of nonuniform cross-sectional area. As sound, generated as discussed above, propagates down these tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed with organ pipes or wind instruments. In the context of speech production, the resonance frequencies of the vocal tract tube are called formant frequencies or simply formants. The formant frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formant frequencies. Different sounds are formed by varying the shape of the vocal tract. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape varies.

Based on the discussion of the mechanisms for speech production/generation by humans, a simple linear model has evolved for characterizing speech signals. If we assume that the excitation signal is represented as $e(t)$, with Fourier transform $E(\omega)$, and the vocal tract transfer function is called $v(t)$, with Fourier transform $V(\omega)$, then the resulting speech waveform is the convolution of $e(t)$ and $v(t)$, i.e.,

$$s(t) = e(t) * v(t) \tag{6.1}$$

or, in the frequency domain

$$S(\omega) = E(\omega) V(\omega) \tag{6.2}$$

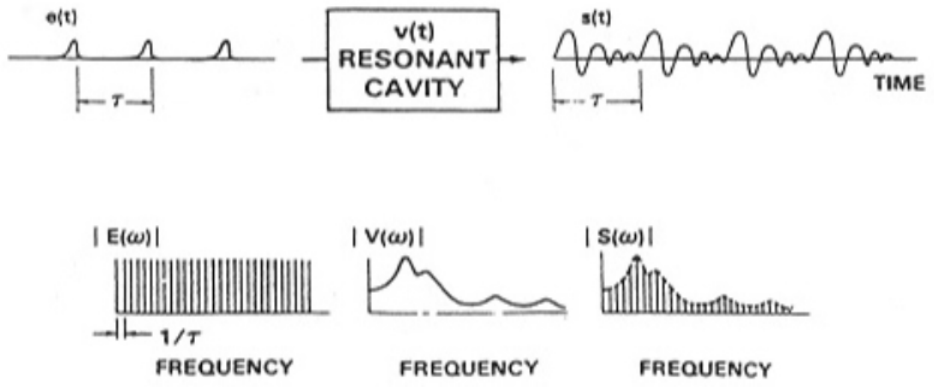

Figure 6.2: Linear model of speech production showing temporal and spectral representations of the source, vocal tract and resulting speech signal. After Rabiner [Rabiner and Schafer, 1978]

Figure 6.2 shows this linear model of speech production with plots of the temporal and spectral representations of the source, vocal tract and the resulting speech signal for a voiced section of speech. The excitation signal is represented as a periodic train of very short pulses with pulse spacing τ and with a flat frequency spectrum consisting of periodic impulses with spacing $1/\tau$ (Strictly speaking, the frequency spectrum of a periodic train of (finite width) pulses is not flat but has a spectral falloff beginning at a frequency related inversely to the width of the pulses. We will not be concerned with this effect at this time.) The vocal tract impulse response has a continuous Fourier transform that peaks at the resonances (formants) of the particular vocal tract configuration, as seen in the middle part of figure 6.2. The resulting speech waveform is also periodic, with period τ, and with a Fourier transform that is the product of the Fourier transforms of the excitation and the vocal tract impulse response; i.e., a line spectrum (with frequency spacing of $1/\tau$ and a shape that is representative of the vocal tract spectral shape, as shown in figure 6.2). The next section investigates how we can adapt the source-filter model to explain the production of acoustic musical instrument sounds from a theoretical point of view.

6.2 The Source-Filter Model for Acoustic Musical Instrument Sounds

Slawson [Slawson, 1985] proposes a theory of sound color and argues that it is associated with the filter in the source-filter model of sound production, which he states can be used to analyze a broad class of sounds. Slawson explains that, according to this model, a sound is produced when an object is struck, or excited, by some kind of mechanical energy and the object in turn changes the excitation in some manner. The mechanical excitation is called the source and the object itself is the filter. The independence of source and filter and, at the same time, their interaction or the modification of the source by the filter are the essential features of the source-filter model.

Slawson states that "Sound color is associated with the filter, not the source. To keep sound color constant, keep the filter constant." Since the theory of sound color is a general theory of the perception of sounds, Slawson also draws a parallel between the source-filter model and speech production. Accordingly, the main source in vowels is the action of the vocal chords and the main filter is the vocal tract, the cavity formed by the throat and mouth. In speech, the vocal chords (source) carry information related to intonation, while the vocal tract (filter) is responsible for

the resonant modes (used in speaker identification and voice conversion systems). For musical instrument sounds, the source is responsible for the pitch and is related to expressivity and effects such as vibrato, while the filter is the resonant cavity, usually the body of the instrument.

One important signal processing aspect of the source-filter model is how we represent the source and the filter mathematically and how we interpret this representation under the light of the theory of sound color. In this respect, the effect of the interaction between source and filter is modeled by the spectral envelope of the resultant sound, like shown in the middle of Figure 6.2 where it is represented by $|V(\omega)|$. Now, Slawson rephrases the previous statement as "To keep the color of a sound constant, keep its spectrum envelope constant." He interprets it as a more precise version of his previous rule because it says that, when using signal processing methods to transform audio signals, in order to keep the filter constant we should keep the effect of the filter constant [Slawson, 1985]. Let us see next under what conditions we can apply the source-filter model to acoustic musical instruments sounds

6.2.1 Acoustic Musical Instruments and Strong Coupling

The instruments of the orchestra produce sounds in a variety of ways. With the exception of some percussion instruments, however, they all share a mode of action which contrasts with that of the vocal tract being excited by the vocal chords and which fail to meet one of the requirements of the source-filter model. All the string and wind instruments have a source (the bow or a buzzing reed), and all have a filter (the string and resonant body or the horn). Theoretically, the model does not apply because of the way the source interacts with the filter in those instruments. The source-filter model supposes that the source and the filter are independent, but in most musical instruments the source is coupled to the filter.

Let us consider the clarinet, for example. We can buzz at any pitch on the mouthpiece alone. However, when the mouthpiece is inserted into the horn, only the pitch that has been fingered can be played (without extra effort). When playing in the standard way, if we attempt to buzz at A below middle C (220 Hz), but we have covered the holes in the instrument in such a way to produce G a whole step below, the G will sound rather than the A. The source may start the sound at its own frequency but immediately the filter, at its favored frequency, begins to act on the source, forcing a change.

This feedback from the filter to the source changes the picture of the source-filter model. The source is strongly coupled to the filter in the clarinet, which means that the clarinet source is not independent like the model assumes originally, it is driven by the filter. In vowel production the filter changes the amplitudes of the source components but does little else to the source. In other words, source and filter are weakly coupled. Therefore, to change the pitch of the voice, we change the pitch of the source. In the clarinet and most other musical instruments, the filter affects the source as well. In fact, the usual method of changing the pitch of a musical instrument is not to alter its source characteristics, but rather to change the filter (the effective length of the horn or of the string).

Figure 6.3 illustrates one possible difference between weakly and strongly coupled systems under a change of pitch of an octave. First we suppose that both systems happen to have the same original spectrum when they are playing the same pitch. We want to investigate the effect on the spectrum of playing a pitch an octave higher.

In the weakly coupled case, the filter - thus the spectral envelope - stays the same, emphasizing and attenuating a set of more sparsely distributed partials of the source that results in a pattern of relative amplitudes of the same partials quite different from the pattern at the lower pitch. In strongly coupled systems, on the other hand, the filter causes changes in pitch, so the filter itself must change. One possibility is for the spectral envelope to stretch out to higher frequencies, with

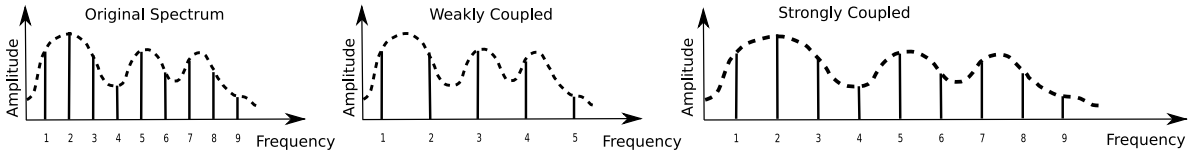

Figure 6.3: Changing pitch under weak and strong coupling. If the original spectrum is produced by a source-filter system that is weakly coupled, raising the pitch by an octave will result in the spectrum shown in the middle. If, on the other hand, the original spectrum is produced by a system with strong coupling, the filter itself changes with the change in pitch. If we assume that the only change in this filter is a proportional stretching, then raising the pitch by an octave results in the alternative spectrum shown on the right. Adapted from Slawson [Slawson, 1985].

the result that the partial amplitudes remain the same as they were at the lower fundamental frequency.

In this case, Slawson says that for strongly coupled systems we need a new rule concerning sound color. He proposes the following "to keep sound color constant, keep the relative intensities of the partials constant." This new statement about sound color and the original rule about the filter (spectral envelope) are contradictory, so Slawson proposes to investigate whether human listeners favor the first or second alternative by carrying out psychoacoustic experiments and suggests that the preponderance of the evidence favors the previous rule that says that to keep sound color constant we must keep the spectral envelope constant [Slawson, 1985].

It is hard, however, to simply reject evidence about the way musical instruments work and the term color has been used in reference to the sounds of musical instruments. So Slawson proposes to resolve the dilemma by closer analysis of the characteristics of musical instruments that do not change with pitch. In other words, we should look for subsystems of a musical instrument that are weakly coupled to the rest of the instrument and that may be largely responsible for the sound color of the instrument.

Let us suppose for a moment that we can represent the strongly coupled and weakly coupled components of musical instruments separately. Now we suppose that we can represent the strongly and the weakly coupled components of the source-filter model for musical instruments in cascade, like shown in figure 6.4. In this model, the source $e(t)$ corresponds to the blowing into the mouthpiece or the bowing of the strings. The highly nonlinear system S_c models the strongly coupled component, responsible for the pitch of the sound emitted by the instrument, such as the fingering for the clarinet. The periodic signal $x(t)$ is the result of the interaction between $e(t)$ and the S_c. The pitched signal $x(t)$ is fed into the weakly coupled component of the system, the linear shift-invariant (LSI) filter W_c that represents the resonator.

Let us see examples of what these separate strongly and weakly coupled components correspond to for acoustic musical instruments. Figure 6.5 illustrates the construction of the violin. When the violinist draws the hair of the bow across the strings, these are set in motion by the energy applied by the friction between the bow and the string. When the strings are bowed, they are set to vibrate with a fundamental frequency given by the length of the string (which may depend on the fingering). The strongly coupled component of the model corresponds to this part. The vibration of the strings containing the fundamental frequency and its overtones (harmonics) are modeled by the signal $x(t)$. The bridge, shown in figure 6.5, is responsible for transmitting the vibration of the strings to the body of the instrument, which is modeled as the weakly coupled component of the system W_c. We suppose that the modes of vibration (resonances) of the body of the instrument are independent (decoupled) from the rest of the system and that the filter W_c models its influence on the sounds produced by the instrument.

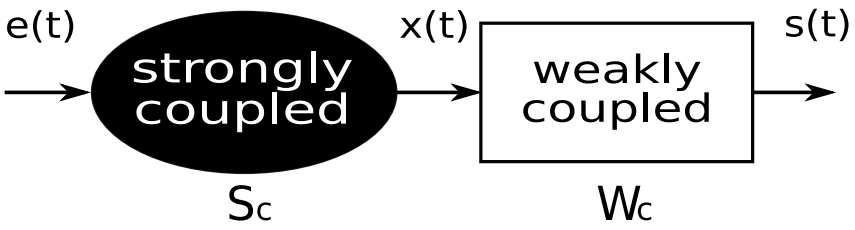

Figure 6.4: Independent representation of the strongly and weakly coupled components of the source-filter model. The figure illustrates the supposition that musical instruments can be represented as two separate components, the first highly nonlinear component accounts for the strongly coupled effect while the second linear component models the weakly coupled response.

There remain acoustic systems that do not fit the source-filter model very well and some perceptual regularity holds for sounds produced in those systems, so Slawson proposes to find another term, such as instrumental quality, for this possible psychological attribute and reserve sound color for cases in which the source and the filter are weakly coupled. So the conclusion is that musical instruments are devices in which sound color cannot be held invariant as the pitch changes. In this work, in order to respect the assumptions about decoupling of source and filter, I propose to morph musical instrument sounds with the same pitch.

6.3 Resonances and the Spectral Envelope

It is important to investigate the relationship between the spectral envelope and the physical properties of the filter, the resonant cavity or the body of the instrument. To characterize the resonances of tubes and cavities that can be used to approximate the body of musical instruments (and even the vocal tract) we would need the its impulse response. According to figure 6.4, the spectrum of the sound $s(t)$ is the result of the interaction between the pitched source $x(t)$ and the filter W_c. The spectral envelope curve models the result of this interaction when the source is not impulsive. Therefore, the formant peaks of the spectral envelope, the regions of higher energy content in the resultant spectrum, are only an approximation of the resonant modes when the instrument is excited by a periodic signal.

In our source-filter model depicted in figure 6.4, we suppose that the filter W_c is responsible for sound color, and thus the spectral envelope models its influence rather than that from S_c. Taking one step further, we will assume that the source $e(t)$ has a flat spectrum (such as white noise) and the influence of the system S_c changes only the frequency values of the partials, such that the effect of S_c on the amplitudes is negligible. This means that we assume the signal $x(t)$ has a flat magnitude spectrum and is periodic, with period defined by the system S_c.

When morphing musical instrument sounds, we are mainly interested in the relationship between the signals $x(t)$ and the resulting sound $s(t)$, so that we will model $x(t)$ as the source that drives the resonant filter W_c. The musical instrument sound $s(t)$ is separated into a sinusoidal component $s_s(t)$ plus a residual component $s_r(t)$ as follows

$$s(t) = s_s(t) + s_r(t) \tag{6.3}$$

where $s_r(t)$ can be seen resulting from the subtraction of the purely sinusoidal component $s_s(t)$ from the original sound $s(t)$ as follows $s_r(t) = s(t) - s_s(t)$. Both the sinusoidal component $s_s(t)$ and the residual component $s_r(t)$ are modeled as source and filter. The filter component of both is modeled via spectral envelope estimation, while the sources are modeled separately.

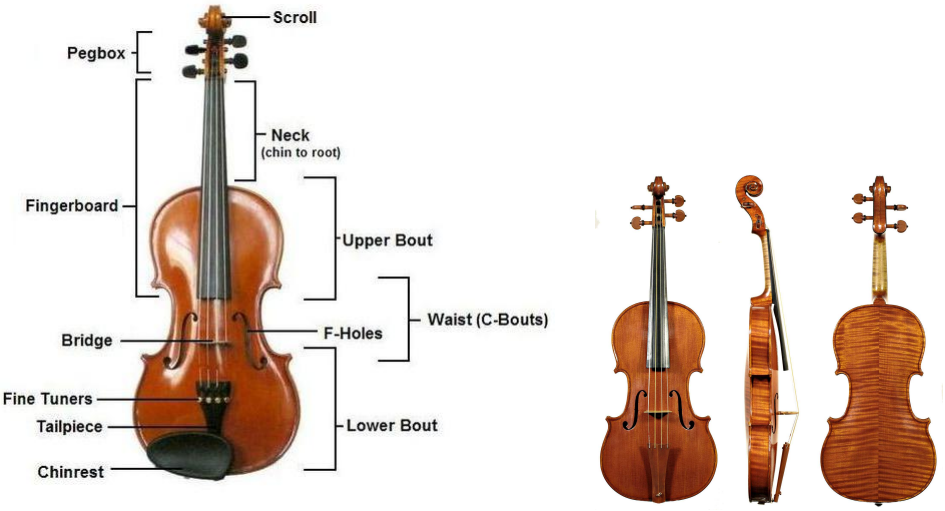

Figure 6.5: The construction of the violin. The figure names each part of the violin.

The source part of the sinusoidal component is modeled as sinusoids using sinusoidal analysis [McAulay and Quatieri, 1986, Serra and Smith, 1990], and the source part of the residual component is modeled as white noise. When we impose some constraints on the signal $s(t)$, the resonant filter W_c can be considered as an LSI system, such that linearity guarantees that the source signal $x(t)$ can also be decomposed into a sum of a sinusoidal plus a residual parts as follows

$$x(t) = x_s(t) + x_r(t) \tag{6.4}$$

When the filter W_c meets the conditions to be modeled as an LSI system, the following relation holds

$$s(t) = x(t) * W_c(t) = [x_s(t) + x_r(t)] * W_c(t) = x_s(t) * W_c(t) + x_r(t) * W_c(t) = s_s(t) + s_r(t) \tag{6.5}$$

As will be clear later in section 6.5, the source $x_s(t)$ is modeled as a sum of sinusoidal partials and the effect of the resonant filter W_c on the sinusoidal component $h_s(t)$ is modeled as the spectral envelope of $y_s(t)$. For the residual component, the effect of the resonant filter W_c is modeled as the spectral envelope of $y_r(t)$, while the residual source $x_r(t)$ that drives it is considered white noise.

It is important to notice that only when we interpolate both source and filter is the result morphing according to the formalization in section 3.3.3. Stylianou [Stylianou, 2008] gives a clear example for voice transformations. Still, there are authors that do not draw a line between sound morphing and cross-synthesis [Wen and Sandler, 2010], specially when using a source-filter model.

It is very easy to fall into the trap of confusing cross-synthesis and sound morphing. Using the formalization from chapter 3, we can view the difference between cross-synthesis and morphing more clearly using the vocabulary employed in the source-filter model. In cross-synthesis, the sources and filters are not changed in any way, they are estimated for both sounds and exchanged upon resynthesis. Morphing, on the other hand, requires an additional step (usually achieved by means of interpolation) to obtain a morphed version of the source and of the filter. The morphed source and filter are then used to resynthesize the result.

6.4 The Source-Filter Model from a Temporal Perspective

It is clearly not enough to only take spectral characteristics of the sounds into account when modeling the sounds of musical instruments. Contrary to the Helmholtz model, we know that musical instrument sounds have dynamic varying spectral features and that this temporal variation plays an important part in the perception of these sounds. One illustrative example would be playing a recorded piano note forward and backward. We know that the spectral contents are exactly the same, but due to the reversed order, they are perceived as two completely different sounds.

In the example of the reversed piano note, the major change between both sounds we hear can be explained by the temporal envelope. The temporal envelope describes how the energy (or amplitude in some instances) of the sounds evolves in time. This distribution of energy in time is a major factor in (dis)similarity judgments and it has been shown to affect the perception of "percussiveness" [Skowronek and McKinney, 2006]. Chapter 9 is dedicated to the estimation of the temporal envelope. In this section we will consider the temporal evolution of the source and filter components, while chapter 8 will present considerations on how the interaction between source and filter from a strictly temporal perspective affects the resulting sound, and specially how it can be used to include perceptually salient features of musical instrument sounds in our model

Even though there are other temporal factors that affect our perception of sounds (such as attack time, for instance), the temporal envelope remains among the most perceptually relevant. It can be said that the temporal envelope of a sound depends primarily on that of the excitation. In other words, the temporal evolution of the energy depends on how energy is supplied to the system (musical instrument).

Figure 6.6 shows a simplified schematic view of the temporal evolution of the excitation (dotted line) and the resulting temporal envelope followed by the sound (solid line) for two markedly distinct classes of excitation methods, namely step-like (part a) and impulse-like (part b). Step-like excitation corresponds to playing modes whose energy is supplied for some length of time before being interrupted, while impulse-like is when energy is supplied in a short burst. The former typically applies to sustained sounds resulting from bowed strings and blown instruments, and the latter to percussive excitations such as plucked strings or struck instruments, although blown or bowed staccato notes would probably be better described by it. The beginning and end of the energy supply for each excitation mode are highlighted by long arrows and the short arrow marks the maximum amplitude attained by the sound.

In chapter 8, we make a connection between the mainly physical events such as onset, attack, decay, sustain, release and offset and its model counterparts in connection with the excitation and resulting temporal evolution presented earlier. The idea is to find signal level manifestations of the physical gestures. Our main goal is to show that these events cannot be solely described by the amplitude envelope of most sounds, such that we need a more complete model to appropriately segment them.

6.5 Mathematical Modeling of Source and Filter

All musical instruments produce sound via the excitation of a vibrating structure. Woodwind, brass and percussion instruments radiate sound directly. However, stringed instruments radiate sound indirectly because the vibrating string itself radiates an insignificant amount of energy. Energy from the vibrating string therefore has to be transferred to the much larger area, acoustically efficient, radiating surfaces of the body of the instrument. The resultant modes of vibration are complex and involve the interactions and vibrations of all the component parts, such as strings, bridge, front and back plates, sound post, neck, and even the air inside the volume of the violin

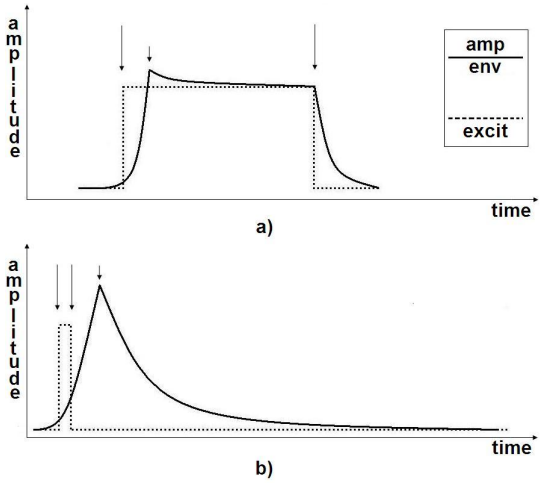

Figure 6.6: Simplified temporal evolution of the excitation (dotted line) and resulting amplitude envelope (solid line) for two distinct classes of excitation modes. In a) we see the typical excitation and amplitude envelope resulting from the step-like excitation (e.g., blown/bowed) and in b) for impulse-like excitation (e.g., plucked/struck). The beginning and end of the excitation are marked with long arrows, while the short arrow shows the maximum amplitude attained by the resulting amplitude envelope.

body.

Any vibrating structure presents a number of normal modes of vibration. Damping plays a major role on the nature of the normal modes and the normal modes can be described by the same equations of motion as a simple damped mass-spring resonator.

$$m_n \left(\frac{\partial^2 \xi_n}{\partial t^2} + \frac{\omega_n}{Q_n} \frac{\partial \xi_n}{\partial t} + \omega_n^2 \xi_n \right) = F(t) \tag{6.6}$$

where the effective mass m_n at the point p is defined in terms of the kinetic energy of the excited mode $\frac{1}{2} m_n \left(\partial \xi_n / \partial t\right)_p^2$, $\omega_n = 2\pi f_n$ is the eigenfrequency of free vibration of the excited mode in the absence of damping and Q_n is the quality factor describing its damping. We consider a local driving force $F(t)$ at point p, although it can be applied at any chosen point on the structure or distributed over the whole surface.

Typical driving forces are those acting on the bridge of a bowed or plucked string instrument and the pressure fluctuations at the input end of the air column of a blown woodwind or brass instrument. Such forces are generated by highly nonlinear excitation mechanisms. In contrast, the vibrations of the vibrating structure are generally linear with displacements proportional to the driving force. However, there are important exceptions for almost all types of instruments, when nonlinearity becomes significant at sufficiently strong excitation.

In any continuously bowed or blown musical instrument, feedback from the vibrating system results in a periodic driving force, which will not in general be sinusoidal. Nevertheless, by the Fourier theorem, any periodic force can always be represented as a superposition of sinusoidally varying, harmonically related partials, with frequencies that are integer multiples of the periodic repetition frequency. We can therefore consider the induced vibrations of any musical instrument in terms of the induced response of its vibrational modes to a harmonic series of sinusoidal driving forces.

6.5.1 Signal Processing Modeling of Source and Filter

The source-filter model of speech production views speech as the result of passing a glottal excitation signal through a time-varying linear filter that models the resonant characteristics of the vocal tract. A well known source-filter system is that based on linear prediction (LP) of speech, explained in detail in chapter 7. In its simplest form, a time-varying filter modeled as an autoregressive (AR) filter is excited by either quasi-periodic pulses (during voiced speech) or noise (during unvoiced speech). A more flexible representation of the excitation signal has been proposed for speech [McAulay and Quatieri, 1986] and musical instrument sounds [Serra and Smith, 1990] independently and is referred to as sinusoidal models (SM). In sinusoidal modeling, the excitation signal $x_s(t)$ is represented by a sum of sinusoids

$$x_s(t) = \sum_{k=0}^{K(t)} a_k(t) \exp[j\phi_k(t)] \tag{6.7}$$

where $a_k(t)$ and $\phi_k(t)$ are the instantaneous excitation amplitude and phase of the k^{th} sinusoid, respectively, and $K(t)$ is the number of sinusoids, which may vary in time. For speech and musical instrument sounds, a model where the sinusoids are harmonically related is a good approximation (even though they are quasi-harmonic), which leads to

$$\frac{d}{dt}\phi_k(t) = 2\pi k t f_0(t) \tag{6.8}$$

where $f_0(t)$ is the instantaneous fundamental frequency (which is the closest known acoustic correlate of pitch perception, discussed briefly in chapter 5. For both speech and musical instrument sounds, a further simplification of the excitation signal is convenient, assuming that the amplitude of the excitation signal $a_k(t)$ is constant over time (and equal to unity, i.e., $a_k(t) = 1$). Based on these simplifications, the time-varying linear filter that models the resonant characteristics of the vocal tract for speech and of the vibrating structure of the body of the musical instrument approximates the effects of the shape of the excitation and of the transmission characteristics of the resonant body. For example, for speech sounds, the filter accumulates the effects of superglottal cavities including radiation at the mouth opening and of the glottal pulse shape. The time-varying transfer function of the filter can be written as

$$H_s(f,t) = |H_s(f,t)| \exp[j\psi_s(f,t)] \tag{6.9}$$

where $|H_s(f,t)|$ and $\psi_s(f,t)$ are respectively the amplitude and phase of the system. The processing of speech and musical instrument sounds is usually done on a frame-by-frame basis, where each frame typically containing three periods of the waveform can be considered a stationary process [Stylianou, 2008]. In this case, inside a frame, the filter $H_s(f,t)$ is considered LSI. Then the output of the system can be viewed as the convolution of the impulse response of the LTI filter, $h_s(t)$, and of the excitation signal $x_s(t)$

$$s_s(t) = \int_0^t x_s(\tau)\, h_s(t-\tau)\, d\tau \tag{6.10}$$

Recognizing then that the excitation signal is just the sum of $K(t)$ eigenfunctions of the filter $H_s(f,t)$, the following model is obtained

$$s_s(t) = \sum_{k=0}^{K(t)} |H_s[f_k(t)]| \exp[j(\phi_k(t) + \psi_s(f_k(t)))] = \sum_{k=0}^{K(t)} A_k(t) \exp[j\theta_k(t)] \tag{6.11}$$

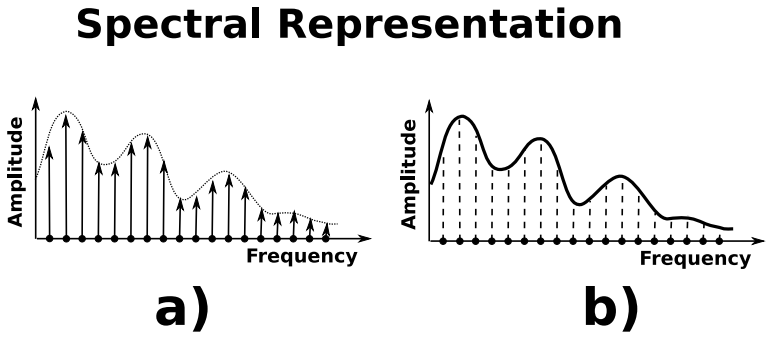

Figure 6.7: Spectral representation of partials. The figure shows the traditional sinusoidal representation with the frequency values and amplitudes tied to each other in part a). Part b) depicts our representation, where the amplitudes of the partials are represented independently with a spectral envelope model.

where $f_k(t) \approx k f_0(t)$, which are the eigenfrequencies of the filter $|H(f,t)|$. The amplitude $A_k(t)$ of the k-th harmonic is the system amplitude $|H[f_k(t)]|$, which is the eigenvalue. The phase $\theta_k(t)$ of the k-th harmonic is the sum of the excitation phase $\phi_k(t)$ and the system phase $\psi_s[f_k(t)]$ and is often referred to as the instantaneous phase of the k-th harmonic.

In our model, the filter is considered an LSI system and it is modeled as the spectral envelope of each frame, such that the amplitudes of the partials $A_k(t)$ are given by the spectral envelope curve, as shown in part b) of figure 6.7. Figure 6.7 compares the spectral representation of partials for the traditional sinusoidal modeling approach in part a), and for the source-filter model in part b). In sinusoidal modeling, each partial is assigned an amplitude and frequency values, while the source-filter modeling represents them intrinsically independently.

6.5.2 Estimation of Source and Filter

The estimation of the source and filter parts for both the sinusoidal and residual components is a key aspect of the method. The quality of the results depends largely on the accuracy of the representation. As mentioned earlier, the musical instrument sounds are first decomposed into a sinusoidal $s_s(t)$ and a residual component $s_r(t)$. Each component is modeled (and processed) separately as follows.

6.5.2.1 Sinusoidal Component

As stated earlier, the sinusoidal part is decomposed into a sinusoidal source $x_s(t)$ and the response of the resonance cavity W_c when excited by $x_s(t)$, $h_s(t)$ for each frame. The frequencies of the sinusoids $f_k(t)$ are estimated from the Fourier spectrum using quadratic interpolation [McAulay and Quatieri, 1986]. The filter response $h_s(t)$ is estimated as the spectral envelope of the Fourier spectrum $H_s(\omega)$. The spectral envelope estimation method used is extremely important. Wen and Sandler [Wen and Sandler, 2010] propose an algorithm based on the channel vocoder to model the filter part. However, for voice conversion tasks, Villavicencio [Villavicencio et al., 2006] showed that "true envelope" (TE) [Röbel and Rodet, 2005] outperformed the other spectral envelope estimation methods tested. TE can be interpreted as the best bandlimited interpolation of the spectral peaks [Villavicencio et al., 2007], minimizing the estimation error for the peaks of the spectrum. Thus "true envelope" was chosen to estimate the spectral envelope curve $H_s(\omega)$. Figure 6.8 shows the source-filter modeling from a spectro-temporal perspective.

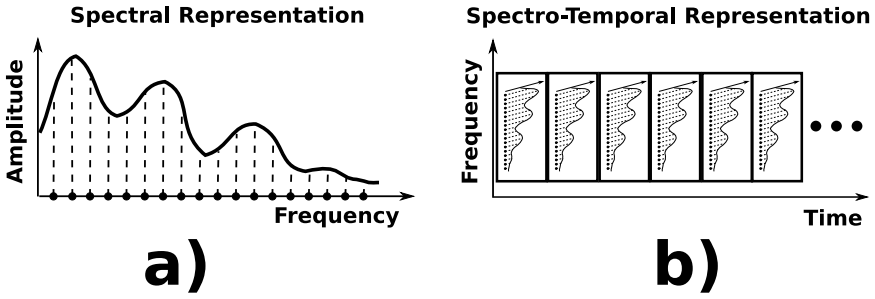

Figure 6.8: Spectro-temporal illustration of the source-filter model. In part a) we see the spectral representation of source and filter for one frame, while part b) represents the temporal view, where the frames are arranged in temporal succession.

6.5.2.2 Residual Component

The residual signal $s_r(t)$ is modeled as white noise driving the response of the system W_c. Thus, the source part $x_r(t)$ is white noise, and the response of the resonant cavity to the excitation $x_r(t)$ is modeled as the spectral envelope of $s_r(t)$. The spectral envelope of the residual component is calculated using linear prediction because it provides a better estimate for noise. In this case, the source part is modeled using a spectral envelope curve that follows the average energy of the magnitude spectrum rather than fit the amplitudes of the spectral peaks.

6.5.3 Filter Modifications

By filter modification we mean modification of the magnitude spectrum of the frequency response of the system $H(f,t)$. In this work, I propose to do this by modifying the spectral envelope of each frame because the filter is modeled as the spectral envelope. Modification of the spectral envelope is achieved by simply varying the parameters of the spectral envelope representation, as explained in chapter 13. Perceptually, the spectral envelope is usually associated with timbre, but since timbre is such multidimensional phenomenon and it is so complex to characterize, Slawson narrows down the perceptual relevance of the spectral envelope to what he calls sound color. So, for musical instrument sounds, manipulation of the spectral envelope corresponds to changing the sound color.

Since timbre is an important factor in sound source identification, radical changes in the spectral envelope parameters can lead to the sound resulting from the modified envelope not being recognized as the original musical instrument anymore. Sometimes this is the effect we want to achieve, but most of the time we want to be able to change the spectral envelope slightly in order to manipulate perceptual features of sounds that depend on the spectral envelope. In this case, we need to know the relation between the parameters of the spectral envelope and the corresponding features. Ideally, we want to be able to manipulate the features independently.

So, a key aspect of spectral envelope manipulation is what representation of the spectral envelope we will use and how the parameters of this representation relate to the spectral envelope curve. For example, the cepstral coefficients represent the amount of energy in each frequency band associated with the spectral envelope curve and its oscillations, while the line spectral frequencies [Itakura and Saito, 1970, Itakura, 1975, McLoughlin, 2008, Morris and Clements, 2002] are directly related to the peaks of the spectral envelope, as will be clearer in chapter 7. In conclusion, depending on what we want to manipulate, one representation might be more appropriate than the others.

Chapter 7

Spectral Envelope Estimation

This chapter is dedicated to spectral envelope estimation. This chapter reviews the most popular techniques of spectral envelope estimation based on linear prediction and cepstral smoothing. The spectral envelope is among the most important characteristics of musical instrument sounds because it is perceptually related to musical instrument identification [Brown, 1999] and timbre perception [Krumhansl, 1989, Krimphoff et al., 1994, Caclin et al., 2005, McAdams et al., 2005]. Helmholtz [Helmholtz, 1885] was among the first to investigate the relationship between the relative amplitudes of the partials and timbre perception for musical instrument tones (pitched sounds or notes). According to Slawson [Slawson, 1985], the spectral envelope of weakly coupled systems corresponds to the aspect of sound perception referred to as sound color. Terasawa [Terasawa et al., 2005] even proposed that MFCCs model well sound color, suggesting that there are 13 colors (or dimensions) of timbre perception related to the spectral envelope.

Spectral envelope estimation is a very important part of the morphing algorithm because it models the filter in the SF model. The next section presents a formalization of spectral envelopes, intended to make the presentation of the estimation techniques clearer and more readily comparable. First, the estimation of spectral envelopes is addressed, followed by the conversion between different spectral envelope models, and the manipulation of spectral envelopes.

Next, spectral envelope estimation techniques based on linear prediction are presented. The estimation of traditional linear prediction coefficients (LPC) is explained, together with the discrete version of the technique, discrete all-pole (DAP). Then, spectral envelope estimation techniques based on cepstral smoothing are discussed. The basic cepstral smoothing technique is explained, which is based on the cepstrum and homomorphic systems. The discrete cepstrum (DC) and the so called "true envelope" (TE) estimation techniques are also presented. The chapter ends with a discussion of some of the alternative spectral envelope representations, such as line spectral frequencies (LSF), and the conversion from cepstral to linear prediction based representations.

One example where model conversion is desireable is the transmission of speech using codings where the linear prediction coefficients (LPCs) obtained from the estimation of the spectral envelope are usually transmitted as line spectral frequencies (LSFs) because LPCs do not quantize well [Soong and Juang, 1984]. Finally, when the intention is to manipulate the parameters of a given spectral envelope representation to obtain a desired transformation, we should keep in mind what kind of information the parameters of a given spectral envelope representation encode and how changes in the values of these parameters reflect on the spectral envelope curve they represent.

7.1 Formalization of Spectral Envelopes

Generally, when we talk about estimating spectral envelopes we mean estimating the values of the parameters of a spectral envelope model that lead to a spectral envelope curve that fits the magnitude spectrum optimally according to some criterion. Before discussing techniques to do it, it is useful to define spectral envelope curves, spectral envelope models and spectral envelope model parameters. Spectral envelope curves are functions of frequency like the Fourier spectrum. A spectral envelope model consists of a set of parameters and a spectral envelope map that gives the spectral envelope curve when applied to the parameters. More formally,

$$S(\boldsymbol{\sigma}) = H(\omega) \tag{7.1}$$

where S is the spectral envelope map applied to the vector of parameters $\boldsymbol{\sigma}$, and $H(\omega)$ is the resulting spectral envelope, a function of angular frequency ω. The spectral envelope curve $|H(\omega)|$ is then the absolute value of $H(\omega)$. We should notice that the parameters of a given spectral envelope model are not necessarily frequency values.

Some spectral envelope maps are invertible, such that we can recover the parameters from the spectral envelope with the application of its inverse S^{-1}. This operation can be defined as $\boldsymbol{\sigma} = S^{-1}(H(\omega))$. There is an analogy between spectral envelopes and filters. The spectral envelope curve $|H(\omega)|$ corresponds to the magnitude frequency response of the filter, and $\boldsymbol{\sigma}$ to the parameters that specify the frequency response, also called filter coefficients. The map S is simply the mathematical operations needed to obtain the spectral envelope from the parameters, such as the inverse Fourier transform to obtain the frequency response of the filter from its coefficients.

Each spectral envelope model or representation requires a particular map, which, in turn, acts on a given set of parameters whose range and characteristics (real, nonnegative integer, complex) depend on the model adopted.

7.1.1 Estimation of the Spectral Envelope

Herman von Helmholtz [Helmholtz, 1885] is among the first to speculate that timbre perception depends on the relative amplitudes of the partials. This provides us with a first definition of the spectral envelope curve, that is, a curve that connects the peaks of the magnitude spectrum.

Early attempts to estimate the spectral envelope rely on this definition and use a simple piecewise linear approximation connecting the prominent peaks of the spectrum [Burred et al., 2010]. The parameters of this representation of the spectral envelope are the parameters of each straight line fitted to connect two spectral peaks. Even though a piecewise-linear approximation does not give a smooth spectral envelope curve, it has proved to be accurate enough in musical instrument recognition tasks [Burred et al., 2010]. Another possible piecewise approximation that generates a smooth spectral envelope curve is obtained by polynomial splines [Hahn et al., 2010], where now the parameters are naturally the coefficients that define the splines.

D'haes [D'haes and Rodet, 2003] defines the spectral envelope as "a function of frequency that matches the amplitudes of the individual partials in the spectrum". Burred [Burred et al., 2010], on the other hand, prefers the definition "a smooth curve that approximately matches the peaks of the spectrum." From this perspective, spectral envelope estimation can be considered an interpolation of the amplitudes of the spectral peaks. The discrete all-pole (DAP) and discrete cepstrum (DC) estimation methods are based on this principle.

The estimation of the spectral envelope is intimately linked to the source-filter model [Hahn et al., 2010, Klapuri et al., 2010, Laroche and Meillier, 1998, Wen and Sandler, 2010] because it corresponds to the identification of the parameters of the filter via deconvolution. The main goal of this deconvolution between source and filter by means of spectral envelope estimation is to eliminate the harmonic structure of the spectrum, which is associated with the source.

There are two main classes of deconvolution methods, linear prediction (also called auto-regressive) [Markel and Gray, 1976, Makhoul, 1975] and cepstral smoothing (also called homomorphic deconvolution) [Oppenheim and Schafer, 1968, Oppenheim, 1964]. Each of these deconvolution classes unfolds in a number of spectral envelope estimation techniques, some of which will be reviewed in this chapter such as "true envelope."

Perhaps the most important aspect of any spectral envelope estimation technique is the accuracy with which the spectral envelope curve represents the magnitude spectrum. Among the possible ways of measuring the accuracy of estimation, a simple error measure between the peaks of the magnitude spectrum and the spectral envelope curve at those points is enough for many applications [Villavicencio et al., 2006]. However, the parameters of the most accurate spectral envelope estimation might not be the most suitable representation when we want to manipulate the spectral envelope to perform transformations. In this case, we require an additional step to perform the conversion of spectral envelope representation.

7.1.1.1 Conversion Between Spectral Envelope Representations

When we choose to use an alternative representation of a spectral envelope to manipulate it, we need to convert the parameters of the current model into the equivalent set of parameters of the target representation. This operation is the spectral envelope model conversion. The conversion can be done directly between the parameters of two representations or using the spectral envelope curve to perform the operation.

Some conversion techniques can be regarded simply as a different representation of the parameters in the same domain, linear prediction or cepstral, for example. Line spectral frequencies (LSF) are just a more convenient way of representing linear prediction coefficients (LPC) for some applications [McLoughlin, 2008, Morris and Clements, 2002, Itakura and Saito, 1970]. The same can be said of reflection coefficients (RC), among many other possibilities.

A challenging conversion is between cepstral and linear prediction representations of spectral envelopes. At the end of this chapter we will review some alternative spectral envelope representations, along with techniques to convert between linear prediction and cepstral based spectral representations directly between the coefficients and via the spectral envelope curve.

Supposing we have two sets of parameters $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ corresponding to two different spectral envelope models with maps S_1 and S_2. The conversion between the coefficients uses a map T between the coefficients $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ as follows

$$\boldsymbol{\sigma}_2 = T\left(\boldsymbol{\sigma}_1\right) \tag{7.2}$$

The conversion operation via the spectral envelope curve is defined next. When we estimate the spectral envelope curve $|H(\omega)|$ with the first model, we obtain a set of parameters $\boldsymbol{\sigma}_1$ for which the following relationship holds

$$S_1\left(\boldsymbol{\sigma}_1\right) = H(\omega) \tag{7.3}$$

Using the property of invertibility of some spectral envelope representations, we can always recover the set of parameters $\boldsymbol{\sigma}_2$ that correspond to a spectral envelope function $H(\omega)$ via the inverse operator S_2^{-1}, which gives us $S_2^{-1}\left(H(\omega)\right) = \boldsymbol{\sigma}_2$.

Spectral envelope model conversion can be considered as a spectral envelope manipulation technique whose main requirement is to preserve the spectral envelope curve as accurately as possible. Naturally there have been proposals to measure spectral distortion or spectral distance [Itakura and Saito, 1968]. Among the measures of spectral distortion, the Itakura-Saito (IS) distance [Itakura and Saito, 1968] stands out as a very popular choice mainly because it is said to capture perceptually related information from the magnitude spectrum (related to the spectral

envelope curve and to the balance of spectral energy as measured by the spectral shape features). Chapter 13 makes use of the IS distance to evaluate the result of the spectral envelope conversion technique used in this work, which uses the spectral envelope curve to perform the conversion.

7.1.1.2 Manipulation of Spectral Envelopes

The manipulation of spectral envelopes is usually intended to change the spectral envelope curve associated with a given model. These manipulation techniques are sometimes called spectral envelope transformation and spectral envelope morphing is one of them. Spectral envelope transformations are usually attained by changes in the values of the parameters of a given spectral envelope model rather than changes in the values of the spectral envelope curve, such that the transformed spectral envelope curve is smooth.

This procedure usually assumes that S is a continuous map between the space of parameters $\boldsymbol{\sigma}$ and the space of spectral envelopes functions $H(\omega)$ associated. Intuitively, small changes in the input values (the parameters) of a continuous map (such as the addition of a small perturbation vector) result in small changes in the output. More formally,

$$S(\boldsymbol{\sigma}_p) \approx S(\boldsymbol{\sigma}_p + \delta_p) \tag{7.4}$$

where δ_p represents a small perturbation vector around point $\boldsymbol{\sigma}_p$. There are many possible ways of performing such changes and it usually depends on the type of information encoded in the parameters and the intention of the transformation. In this work we are interested in spectral envelope morphing, which involves finding intermediate representations between spectral envelopes. We are interested in the spectral envelope curve associated with this intermediate representation. But first, let us see how to estimate spectral envelopes in this chapter. Chapter 13 is dedicated to morphing spectral envelopes.

7.2 Linear Prediction

The roots of linear prediction lie in the analysis of the outputs of dynamic systems regarded as time series, and treated mostly from a statistical approach [Makhoul, 1975]. In linear prediction the signal is modeled as a linear combination of its past values and present and past values of a hypothetical input to a system whose output is the given signal. In the frequency domain, this is equivalent to modeling the signal spectrum by a pole-zero spectrum [Makhoul, 1975].

Linear prediction is used to find a parametric model of a system in terms of a signal it is supposedly the output of. The general linear prediction model states that a signal $s(n)$ can be considered to be the output of a system with some unknown input $u(n)$ such that the following relation holds

$$s(n) = -\sum_{k=1}^{p} a(k)\, s(n-k) + G \sum_{l=0}^{q} b(l)\, u(n-l) \tag{7.5}$$

where $b(0) = 1$ and $a(k)$, $1 \leqslant k \leqslant p$, $b(l)$, $1 \leqslant l \leqslant q$, and the gain G are the parameters of the hypothesized system. Equation (7.5) says that the "output" $s(n)$ is a linear function of past outputs and present and past inputs. That is, the signal $s(n)$ is predictable from linear combinations of past outputs and inputs. Equation (7.5) can also be specified in the frequency domain by taking the z transform on both sides of (7.5). If $H(z)$ is the transfer function of the system. Then we have

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b(l)\, z^{-l}}{1 + \sum_{k=1}^{p} a(k)\, z^{-k}} \tag{7.6}$$

where the z transform of the sequence $s(n)$ is defined as

$$S(z) \triangleq \sum_{n=-\infty}^{\infty} s(n) z^{-n} \tag{7.7}$$

such that $H(z)$ is the z transform of $u(n)$. $H(z)$ in equation (7.6) is the general zero-pole model. The roots of the numerator and denominator polynomials are respectively the zeros and poles of the model. There are two special cases of the model that are of interest

- all-zero model: $a(k) = 0,\ 1 \leqslant k \leqslant p$
- all-pole model: $b(l) = 0,\ 1 \leqslant l \leqslant q$

The all-zero model is known in the statistical literature as the moving average (MA) model, and the all-pole model is known as the autoregressive (AR) model. The pole-zero model is then known as the autoregressive moving average (ARMA) model [Makhoul, 1975]. In speech processing, we generally use a source-filter model to represent the way speech sounds are produced [Rabiner and Schafer, 1978, Rabiner, 1993]. The source-filter model can use an all-pole system derived from linear prediction to analyze discrete signals. The model can also be extended to musical instruments [Klapuri et al., 2010, Laroche and Meillier, 1998, Hahn et al., 2010, Wen and Sandler, 2010].

7.2.1 Parameter Estimation

Here we will review the methods to fit the parameters of of the linear prediction model, namely, the autocorrelation and covariance methods, for a deterministic and random signals.

7.2.1.1 All-Pole Model

In the all-pole model, we assume that the signal $s(n)$ is given as a linear combination of past values and some input $u(n)$

$$s(n) = -\sum_{k=1}^{p} a(k) s(n-k) + Gu(n) \tag{7.8}$$

where G is a scalar gain. The transfer function $H(z)$ in equation (7.6) now reduces to

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a(k) z^{-k}} \tag{7.9}$$

Given a particular signal $s(n)$, the problem is to determine the predictor coefficients $a(k)$ and the gain G in some manner.

7.2.1.2 Method of Least Squares

Here we will present the derivation using the least squares approach assuming first that $s(n)$ is a deterministic signal and then that $s(n)$ is a sample from a stochastic process, following the standard derivation as presented by Makhoul [Makhoul, 1975]. Firstly, we assume that the input $u(n)$ is totally unknown, which is the case in many applications [Rabiner and Schafer, 1978]. Therefore, the signal $s(n)$ can be predicted only approximately from a linearly weighted summation of past samples. Let this approximation of $s(n)$ be $\tilde{s}(n)$, where

$$\tilde{s}(n) = -\sum_{k=1}^{p} a(k) s(n-k) \tag{7.10}$$

Then the error between the actual value $s(n)$ and the predicted value $\tilde{s}(n)$ is given by

$$e(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^{p} a(k)\, s(n-k) \tag{7.11}$$

$e(n)$ is also known as the residual. In the method of least squares the parameters $a(k)$ are obtained as a result of the minimization of the mean or total squared error with respect to each of the parameters.

The analysis will be developed along two lines. First, we assume that $s(n)$, is a deterministic signal, and then we give analogous derivations assuming that $s(n)$ is a sample from a random process.

7.2.1.3 Deterministic Signal

Denote the total squared error by E, where

$$E = \sum_{n} e^2(n) = \sum_{n} \left(s(n) + \sum_{k=1}^{p} a(k)\, s(n-k) \right)^2 \tag{7.12}$$

The range of the summation in equation (7.12) and the definition of $s(n)$ in that range are important. However, we will minimize E without specifying the range of summation. E is minimized by setting

$$\frac{\partial E}{\partial a(i)} = 0 \tag{7.13}$$

with $1 \leqslant i \leqslant p$. From equations (7.12) and (7.13) we obtain the set of equations

$$\sum_{k=1}^{p} a(k) \sum_{n} s(n-k)\, s(n-i) = -\sum_{n} s(n)\, s(n-i) \tag{7.14}$$

also with $1 \leqslant i \leqslant p$. Equations (7.14) are known in the least squares terminology as the normal equations. For any definition of the signal $s(n)$, equations (7.14) form a set of p equations in p unknowns which can be solved for the predictor coefficients $\{a(k), 1 \leqslant k \leqslant p\}$ which can minimize E in equation (7.12).

The minimum total squared error, denoted E_p, is obtained by expanding equation (7.12) and substituting into equation (7.14). The result can be shown to be

$$E_p = \sum_{n} s^2(n) + \sum_{k=1}^{p} a(k) \sum_{n} s(n)\, s(n-k) \tag{7.15}$$

We shall now specify the range of summation over n in equations (7.12), (7.14) and (7.13). There are two cases of interest, which will lead to two distinct methods for the estimation of parameters, namely the *autocorrelation method* and the *covariance method*. First we will assume that the error in equation (7.12) is minimized over the infinite duration $-\infty < n < \infty$, which leads to the autocorrelation method.

7.2.1.4 Autocorrelation Method

Equations (7.14) and (7.13) then reduce to

$$\sum_{k=1}^{p} a(k)\, R(i-k) = -R(i) \tag{7.16}$$

with $1 \leqslant i \leqslant p$, and

$$E_p = R(0) + \sum_{k=1}^{p} a(k) R(k) \tag{7.17}$$

where

$$R(i) = \sum_{n=-\infty}^{\infty} s(n) s(n+i) \tag{7.18}$$

is the autocorrelation function of the signal $s(n)$. Note that $R(i)$ is an even function of the index i, i.e.,

$$R(-i) = R(i) \tag{7.19}$$

Since the coefficients $R(i-k)$ form what often is known as an autocorrelation matrix, this method is generally called the autocorrelation method. An autocorrelation matrix is a symmetric Toeplitz matrix[1].

In practice, the signal $s(n)$ is known over only a finite interval, or we are interested in the signal over only a finite interval. One popular method is to multiply the signal $s(n)$ by a window function $w(n)$ to obtain another signal $s'(n)$ that is zero outside some interval $0 \leqslant n \leqslant N-1$, that is

$$s'(n) = \begin{cases} s(n) w(n), & 0 \leqslant n \leqslant N-1 \\ 0, & \text{otherwise} \end{cases}. \tag{7.20}$$

the autocorrelation function is then given by

$$R(i) = \sum_{n=0}^{N-1-i} s'(n) s'(n+i) \tag{7.21}$$

where $i \geqslant 0$. The shape of the window function $w(n)$ can be of great importance.

7.2.1.5 Covariance Method

In contrast with the autocorrelation method, here we assume that the error E in equation (7.12) is minimized over a finite interval, say, $0 \leq n \leq N-1$. Equations (7.14) and (7.15) then reduce to

$$\sum_{k=1}^{p} a(k) \varphi(k,i) = -\varphi(0,i), \ 1 \leq i \leq p \tag{7.22}$$

$$E_p = \varphi(0,0) + \sum_{k=1}^{p} a(k) \varphi(0,k) \tag{7.23}$$

where

$$\varphi(i,k) = \sum_{n=0}^{N-1} s(n-i) s(n-k) \tag{7.24}$$

is the covariance of the signal $s(n)$ in the given interval. The coefficients $\varphi(k,i)$ in equation (7.22) form a covariance matrix, and, therefore, we shall call this method the covariance method. From equation(7.24) it can be easily shown that the covariance matrix $\varphi(i,k)$ is symmetric, i.e.,

[1] A Toeplitz matrix is one where all elements along each diagonal are equal.

$\varphi(i,k) = \varphi(k,i)$. However, unlike the autocorrelation matrix, the terms along each diagonal are not equal. This can be seen by writing from 7.24

$$\varphi(i+1,k+1) = \varphi(i,k) + s(-i-1)\,s(-k-1) - s(N-1-i)\,s(N-1-k) \tag{7.25}$$

Note from equation 7.25 also that values of the signal s, for $-p \leq n \leq N-1$ must be known: a total of $p+N$ samples. The covariance method reduces to the autocorrelation method as the interval over which n varies goes to infinity.

7.2.1.6 Random Signal

If the signal s_n is assumed to be a sample of a random process, then the error $e(n)$ in equation (7.11) is also a sample of a random process. In the least squares method, we minimize the expected value of the square of the error. Thus

$$E = \Xi\left(e^2(n)\right) = \Xi\left(s(n) + \sum_{k=1}^{p} a(k)\,s(n-k)\right)^2 \tag{7.26}$$

where Ξ represents the expected value. Applying equation (7.13) to (7.26), we obtain the normal equations

$$\sum_{k=1}^{p} a(k)\,\Xi\left((n-k)\,s(n-i)\right) = -\Xi\left(s(n)\,s(n-i)\right),\ 1 \leq i \leq p \tag{7.27}$$

The minimum average error is then given by

$$E_p = \Xi\left(s^2(n)\right) + \sum_{k=1}^{p} a(k)\,\Xi\left(s(n)\,s(n-k)\right) \tag{7.28}$$

Taking the expectations in equations (7.27) and (7.28) depends on whether the process $s(n)$ is stationary or nonstationary.

For a stationary process $s(n)$, we have

$$\Xi\left(s(n-k)\,s(n-i)\right) = R(i-k) \tag{7.29}$$

where $R(i)$ is the autocorrelation of the process. Equations (7.27) and (7.28) now reduce to equations identical to equations (7.16) and (7.17), respectively. The only difference is that here the autocorrelation is that of a stationary process instead of a deterministic signal. For a stationary (and ergodic) process the autocorrelation can be computed as a time average. Different approximations have been suggested in the literature for estimating $R(i)$ from a finite known signal $s(n)$. One such approximation is given by equation (7.21). Using this estimate in the stationary case gives the same solution for the coefficients $a(k)$ as the autocorrelation method in the deterministic case.

For a nonstationary process $s(n)$, we have

$$\Xi\left(s(n-k)\,s(n-i)\right) = R(n-k,n-i) \tag{7.30}$$

where $R(t,t')$ is the nonstationary correlation between times t and t'. $R(n-k,n-i)$ is a function of the time index n. Without loss of generality, we shall assume that we are interested in estimating the parameters $a(k)$ at time $n=0$. Then equations 7.27 and 7.28 reduce to

$$\sum_{k=1}^{p} a(k) R(-k,-i) = -R(0,-i) \tag{7.31}$$

$$E'_p = R(0,0) + \sum_{k=1}^{p} a(k) R(0,k) \tag{7.32}$$

In estimating the nonstationary autocorrelation coefficients from the signal $s(n)$, we note that nonstationary processes are not ergodic, and, therefore, one cannot substitute the ensemble average by a time average. However, for a certain class of nonstationary processes known as locally stationary processes, it is reasonable to estimate the autocorrelation function with respect to a point in time as a short-time average. Examples of nonstationary processes that can be considered to be locally stationary are speech and musical instrument sounds.

In a manner analogous to the stationary case, we estimate $R(-k,-i)$ by $\varphi(i,k)$ in equation (7.24). Using this approximation for the nonstationary autocorrelation leads to a solution for the parameters $a(k)$ in equation (7.31) that is identical to that given by equation (7.22) in the covariance method in the deterministic case. Note that for a stationary signal $R(t,t') = R(t-t')$, and therefore, the normal equations (7.31) and (7.32) reduce to (7.16) and (7.17).

7.2.1.7 Gain Computation

Since in the least squares method we assumed that the input was unknown, it does not make much sense to determine a value for the gain G. However, there are certain interesting observations that can be made. Equation (7.11) can be rewritten as

$$s(n) = \sum_{k=1}^{p} a(k) s(n-k) + e(n) \tag{7.33}$$

Comparing equations (7.8) and (7.33) we see that the only input signal $u(n)$ that will result in the signal $s(n)$ as output is that where $Gu(n) = e(n)$. That is, the input signal is proportional to the error signal. For any other input $u(n)$, the output from the filter $H(z)$ will be different from $s(n)$. However, if we insist that whatever the input $u(n)$, the energy in the output signal must equal that of the original signal $s(n)$, then we can at least specify the total energy in the input signal. Since the filter $H(z)$ is fixed, it is clear from the above that the total energy in the input signal $Gu(n)$ must equal the total energy in the error signal, which is given by E_p in equations(7.17) or (7.23), depending on the method used.

7.2.1.8 Computations of Predictor Parameters

In each of the two formulations of linear prediction presented in the previous section, the predictor coefficients $a(k)$, $1 \le k \le p$, can be computed by solving a set of p equations with p unknowns. These equations are (7.16) for the autocorrelation (stationary) method and (7.22) for the covariance (nonstationary) method. There exist several standard methods for performing the necessary computations, e.g., the Gauss reduction or elimination method and the Crout reduction method. These general methods require $p^3/3 + \mathcal{O}(p^2)$ operations (multiplications or divisions) and p^2 storage locations. However, we note from equations (7.16) and (7.22) that the matrix of coefficients in each case is a covariance matrix. Covariance matrices are symmetric and in general positive semidefinite, although in practice they are usually positive definite. Therefore, equations (7.16) and (7.22) can be solved more efficiently by the square-root or Cholesky decomposition method. This method requires about half the computation $p^3/6 + \mathcal{O}(p^2)$ and about half the storage $p^2/2$ of

the general methods. The numerical stability properties of this method are well understood and it is considered to be quite stable.

Further reduction in storage and computation time is possible in solving the autocorrelation normal equations (7.16) because of their special form. Equation (7.16) can be expanded in matrix form as

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \tag{7.34}$$

Note that the $p \times p$ autocorrelation matrix is symmetric and the elements along any diagonal are identical (i.e., a Toeplitz matrix). Levinson [Makhoul, 1975] derived an elegant recursive procedure for solving this type of equation. The procedure was later reformulated by Robinson [Markel and Gray, 1976]. Levinson's method assumes the column vector on the right hand side of equation (7.34) to be a general column vector. By making use of the fact that this column vector comprises the same elements found in the autocorrelation matrix, another method attributed to Durbin [Rabiner and Schafer, 1978] emerges which is twice as fast as Levinson's. The method requires only $2p$ storage locations and $p^2 + \mathcal{O}(p)$ operations: a big saving from the more general methods. Durbin's recursive procedure can be specified as follows:

$$E_0 = R(0) \tag{7.35}$$

$$k(i) = -\left[R(i) + \sum_{j=1}^{i-1} a^{i-1}(j) R(i-j)\right] / E_{i-1} \tag{7.36}$$

$$a^i(i) = k(i) \tag{7.37}$$

$$a^i(j) = a^{i-1}(j) + k(i)\, a^{i-1}(i-j), \quad 1 \le j \le i-1 \tag{7.38}$$

$$E_i = \left(1 - k^2(i)\right) E_{i-1} \tag{7.39}$$

Equations (7.35) through (7.39) are solved recursively for $i = 1, 2, \ldots, p$. The final solution is given by

$$a(j) = a^p(j), \quad 1 \le j \le p \tag{7.40}$$

Note that in obtaining the solution for a predictor of order p, one actually computes the solutions for all predictors of order less than p. It has been reported [Makhoul, 1975] that this solution is numerically relatively unstable. However, most researchers have not found this to be a problem in practice.

It should be emphasized that, for many applications, the solution of the normal equations (7.17) or (7.23) does not form the major computational load. The computation of the autocorrelation or covariance coefficients require pN operations, which can dominate the computation time if $N \gg p$, as is often the case.

The solution to equation (7.34) is unaffected if all the autocorrelation coefficients are scaled by a constant. In particular, if all $R(i)$ are normalized by dividing by $R(0)$, we have what are known as the *normalized autocorrelation coefficients* $r(i)$

$$r(i) = \frac{R(i)}{R(0)} \tag{7.41}$$

which have the property that $|r(i)| \leq 1$. This can be useful in the proper application of scaling to a fixed point solution to equation (7.34).

A byproduct of the solution in equations (7.35) through (7.39) is the computation of the minimum total error E_i at every step. It can easily be shown that the minimum error E_i decreases (or remains the same) as the order of the predictor increases. E_i is never negative, of course, since it is a squared error. Therefore, we must have

$$0 \leq E_i \leq E_{i-1},\ E_0 = R(0) \tag{7.42}$$

If the autocorrelation coefficients are normalized as in equation (7.41), then the minimum error E_i is also divided by $R(0)$. We shall call the resulting quantity the normalized error V_i

$$V_i = \frac{E_i}{R(0)} = 1 + \sum_{k=1}^{i} a(k)\, r(k) \tag{7.43}$$

From equation (7.42) it is clear that $0 \leq V_i \leq 1,\ i \geq 0$. Also, from equations (7.39) and (A.33), the final normalized error V_p is

$$V_p = \prod_{i=1}^{p} \left(1 - k_i^2\right) \tag{7.44}$$

The intermediate quantities $k_i,\ 1 \leq i \leq p$, are known as the *reflection coefficients*. In the statistical literature, they are known as *partial correlation coefficients*. k_i can be interpreted as the (negative) partial correlation between s_n and s_{n+i} holding $s_{n+1} \cdots s_{n+i-1}$ fixed. The use of the term "reflection coefficient" comes from transmission line theory, where k_i can be considered as the reflection coefficient at the boundary between two sections with impedances Z_i and Z_{i+1}. k_i is then given by

$$k_i = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i} \tag{7.45}$$

The transfer function $H(z)$ can then be considered as that of a sequence of these sections with impedance ratios given from equation (7.45) by

$$\frac{Z_{i+1}}{Z_i} = \frac{1 + k_i}{1 - k_i},\ 1 \leq i \leq p \tag{7.46}$$

The same explanation can be given for any type of situation where there is plane wave transmission with normal incidence in a medium consisting of a sequence of sections or slabs with different impedances. In the case of an acoustic tube with p sections of equal thickness, the impedance ratios reduce to the inverse ratio of the consecutive cross-sectional areas. This fact has been used in speech analysis [Rabiner and Schafer, 1978]. Because of the more familiar "engineering interpretation" for k_i, we shall refer to them as reflection coefficients.

Beside the direct methods for the solution of simultaneous linear equations, there exist a number of iterative methods. In these methods, one begins by an initial guess for the solution. The solution is then updated by adding a correction term that is usually based on the gradient of some error criterion. In general, iterative methods require more computation to achieve a desired degree of convergence than the direct methods. However, in some applications [Rabiner, 1993] one often has a good initial guess, which might lead to the solution in only a few iterations. This can be a big

saving over direct methods if the number of equations is large. Some of the iterative methods are the gradient method, the steepest descent method, Newton's method, conjugate gradient method and the stochastic approximation method.

Up till now we have assumed that the whole signal is given all at once. For certain real time applications it is useful to be able to perform the computations as the signal is coming in. Adaptive schemes exist which update the solution based on every new observation of the signal. The update is usually proportional to the difference between the new observation and the predicted value given the present solution. Another application for adaptive procedures is in the processing of very long data records, where the solution might converge long before all the data is analyzed. It is worth noting that Kalman filtering notions are very useful in obtaining adaptive solutions.

7.2.1.9 Optimal Number of Poles

One of the important decisions that usually have to be made in fitting all-pole models is the determination of an "optimal" number of poles. It is a nontrivial exercise to define the word "optimal" here, for as we have seen, the fit of the model "improves" as the number of poles p increases. The problem is where to stop. Clearly we would like the minimum value of p that is adequate for the problem at hand, both to reduce our computation and to minimize the possibility of ill-conditioning (which increases with p since the normalized error for order p, called V_p, decreases).

If the signal spectrum is an all-pole spectrum with p_0 poles, then we know that $V_p = V_{p_0},\ p \geq p_0$, and $k_p = 0,\ p > p_0$, i.e., the error curve remains flat for $p > p_0$. Therefore, if we expect the signal spectrum to be an all-pole spectrum, a simple test to obtain the optimal p is to check when the error curve becomes flat. But, if the signal is the output of a p_0-pole filter with white noise excitation, then the suggested test will not work, because the estimates of the poles are based on a finite number of data points and the error curve will not be flat for $p > p_0$. In practice, however, the error curve will be almost flat for $p > p_0$. This suggests the use of the following threshold test

$$1 - \frac{V_{p+1}}{V_p} < \delta \tag{7.47}$$

This test must succeed for several consecutive values before one is sure that the error curve has actually flattened out.

Appendix A discusses the stability of the all-pole filter, the frequency domain counterpart of the time domain formulation presented here, and the error analysis. To make the text flow more smoothly, I'll refer the interest reader to the appendices for more details about these topics, and consider them to be out of the scope of the text. They are, nevertheless, important conceptually.

7.3 Discrete All-Pole Model

It has been known for some time that linear prediction (LP) suffers from drawbacks that are especially manifested during voiced segments of speech. Specifically, the peaks of LP spectral estimates during these segments are highly biased towards the pitch harmonics (the partials), especially for high pitched sounds and voices [Makhoul, 1975]. El-Jaroudi and John Makhoul [El-Jaroudi and Makhoul, 1969] pointed out that the drawbacks of LP are inherent to its error criterion. To overcome these drawbacks, they proposed a new all-pole method based on a discrete form of the Itakura-Saito distance measure, shown in equation (7.83).

The new method, which they call discrete all-pole (DAP) modeling, overcomes the well know limitations of LP and generally gives better all-pole spectral envelopes that are less biased towards the pitch harmonics. In DAP, they approximate the spectrum of voiced speech, which has its

energy located approximately at the harmonics of the fundamental frequency, by a discrete spectrum. The problem of finding the spectral envelope is then reduced to fitting an all-pole spectrum to a finite set of spectral points so as to minimize the discrete form of the Itakura-Saito (IS) distance measure. They present an algorithm to compute the optimal envelopes and show that DAP modeling produces generally better fitting spectral envelopes than LP.

7.3.1 Limitations of Linear Prediction

As presented in Section 7.2, the basic concept of LP is to predict the present value of a signal based on its previous p values subject to an error measure. Normally, the error criterion used is a least squares distance measure between the actual and predicted values, like equation (A.17) expresses. For a set of discrete frequencies $\omega_m \in \Omega$, equation A.17 becomes

$$E_{LP} = \frac{1}{N} \sum_{m=1}^{N} \frac{P(\omega_m)}{\hat{P}(\omega_m)} \tag{7.48}$$

where, just like before, $P(\omega)$ is the spectrum of the original signal $s(n)$, and $\hat{P}(\omega)$ is the spectrum of the all-pole envelope approximating it, defined in equation (A.15). Here, $P(\omega_m)$ and $\hat{P}(\omega_m)$ are discrete versions of $P(\omega)$ and $\hat{P}(\omega)$ obtained by evaluating them at frequencies ω_m. It is important to notice that the frequencies ω_m include both positive and negative values and they can be arbitrary and need not be equally spaced [El-Jaroudi and Makhoul, 1969].

Just like for the continuous spectra in Section 7.2.1.1, the minimization of E_{LP} with respect to the predictor coefficients expressed in equation (7.13) leads to the normal equations (7.16) and the prediction error (7.17), where we express the autocorrelation of the discrete spectrum $P(\omega_m)$ as in equation (A.11) in its discrete form (A.23).

An important interpretation of LP is that, by minimizing E_{LP}, we are matching the autocorrelation of the continuous LP envelope $\hat{P}(\omega)$ to that of the given discrete spectrum $P(\omega_m)$, as expressed in Appendix A.1.3. This is equivalent to setting equation (A.23) equal to equation (A.18), the autocorrelation of the discrete spectrum $P(\omega_m)$. For discrete or harmonic spectra, El-Jaroudi [El-Jaroudi and Makhoul, 1969] discusses a typical behavior of LP analysis that consists in mismatching the original envelope $\hat{P}(\omega)$ or not fitting properly the original spectrum $P(\omega)$. El-Jaroudi states that there is a unique all-pole envelope, which he assumes to be the original $P(\omega)$, that perfectly fits the discrete spectrum $P(\omega_m)$ and that LP applied to $P(\omega_m)$ fails to recover it. Finally, he concludes by stating that, for discrete spectra $P(\omega_m)$, the LP error measure expressed in equation (7.48) possesses an error cancellation property that makes it select an envelope other than the only one which passes through all the spectral points.

Following El-Jaroudi [El-Jaroudi and Makhoul, 1969], I will briefly demonstrate why it is unreasonable to expect LP to recover the original envelope from the discrete spectral samples. From equation (A.11), the autocorrelation corresponding to the original all-pole filter with spectrum $P(\omega)$ is expressed as R_o, so that we have the relation

$$P(\omega) = \sum_{l=-\infty}^{\infty} R_o(l)\, e^{-j\omega l} \tag{7.49}$$

which is simply the inverse Fourier transform of equation (A.23). The autocorrelation R corresponding to the discrete samples of the synthesis envelope is defined in equation (A.23). By substituting equation (7.49) into (A.23), we obtain

$$R(i) = \frac{1}{N} \sum_{m=1}^{N} \sum_{l=-\infty}^{\infty} R_o(l)\, e^{-j\omega m(l-i)},\ \forall i \tag{7.50}$$

which is the relation between R and R_o. Equation (7.50) shows the aliasing that occurs in the autocorrelation domain whenever a spectral envelope is sampled at a discrete set of frequencies. For the periodic excitation case, the frequencies ω_m will be equally spaced at $\omega_m = {}^{2\pi(m-1)}/_{N}$, and equation (7.50) reduces to

$$R(i) = \sum_{l=-\infty}^{\infty} R_o(i - lN), \ \forall i \tag{7.51}$$

Now, as stated above, by minimizing E_{LP}, LP matches the autocorrelation $\hat{R}_{LP}$ of the continuous LP envelope $\hat{P}(\omega)$ to $R(i)$,that of the given discrete spectrum $P(\omega_m)$. This means that

$$\hat{R}_{LP}(i) = R(i) = \sum_{l=-\infty}^{\infty} R_o(i - lN) \neq R_o, \ 0 \leq i \leq p \tag{7.52}$$

In other words, since the autocorrelation corresponding to the LP envelope will always be an aliased version of R_o (for the discrete spectrum case), the LP envelope $\hat{P}(\omega_m)$ will not equal the original envelope $\hat{P}(\omega)$. It is also important to note that LP produces a unique all-pole model given a set of autocorrelations, which means that the original all-pole is not a possible solution to the normal equations (7.16).

To improve upon the LP estimate, researchers have devised methods with either a different error criterion or with added constraints to regular LP [El-Jaroudi and Makhoul, 1969]. El-Jaroudi briefly reviews some of these methods and then presents their own solution, which consists in adopting the IS distance measure.

Appendix B presents the properties of the error measure using the Itakura-Saito (IS) distance. Again, the interested reader should check it.

7.4 Cepstral Smoothing

In order to understand the cepstral smoothing technique to estimate spectral envelopes, we need to understand the cepstrum and what kind of information it represents. Nowadays, cepstrum analysis is considered as part of the techniques used in homomorphic systems [Oppenheim and Schafer, 1968, Oppenheim, 1969, Oppenheim et al., 1968]. The advantages of the cepstral representation are numerous: it was found to provide a perceptually-realistic distance measure for assessing the similarity of the spectral envelope of sounds [D'haes and Rodet, 2003], making it a natural candidate for speech/speaker recognition problems; it usually provides smooth envelopes (by contrast with autoregressive envelope modeling), which is a desirable feature in the context of sound synthesis. Also in the music domain, cepstral coefficients have been extensively used in numerous applications such as the retrieval of similar audio tracks [Aucouturier et al., 2005], instrument identification [Brown, 1999], content based audio retrieval [Foote, 1997], synthesis [Schwarz and Rodet, 1999], and they are currently investigated for automated estimation of control parameters for musical synthesis algorithms [D'haes and Rodet, 2003].

In a fascinating review article [Oppenheim and Schafer, 2004], Oppenheim and Schafer tell us that, historically, the term cepstrum [Bogert et al., 1963] was coined independently from the theory of homomorphic systems. To suggest what prompted the invention of the term cepstrum, note that a signal with a simple echo can be represented as

$$x(n) = s(n) - \alpha s(n - \tau) \tag{7.53}$$

The power spectral density of such signal is given by

$$|X(k)|^2 = |S(k)|^2 \left[1 + \alpha^2 + 2\alpha\cos(2\pi k\tau)\right] \tag{7.54}$$

Thus, we see from equation (7.54) that the spectral density of a signal with an echo has the form of an envelope (the spectrum of the original signal) that modulates a periodic function of frequency (the spectrum contribution of the echo). By taking the logarithm of the spectrum, this product is converted to the sum of two components; specifically

$$\tilde{X}(k) = \log|X(k)|^2 = \log|S(k)| + \log\left[1 + \alpha^2 + 2\alpha\cos(2\pi k\tau)\right] \tag{7.55}$$

Thus, $\tilde{X}(k)$ viewed as a waveform has an additive periodic component whose "fundamental frequency" is the echo delay τ . In conventional analysis of time waveforms, such periodic components show up as lines or sharp peaks in the corresponding Fourier spectrum. Therefore, the "spectrum" of the log spectrum would likewise show a peak when the original time waveform contained an echo. This new "spectral" representation domain was not the frequency domain, nor was it really the time domain. So, looking to forestall confusion while emphasizing connections to familiar concepts, Bogert et al. chose to refer to it as the quefrency domain, and they termed the spectrum of the log of the spectrum of a time waveform the cepstrum. While most of the terms in the glossary at the end of the original paper have faded into the background, the term cepstrum has survived and become part of the digital signal processing lexicon. In the early 1960s, totally unrelated to, and independent of, the work by Bogert et al., Alan Oppenheim was pursuing his doctoral research on a class of nonlinear signal processing techniques inspired by the concept of homomorphic (i.e., linear in a generalized sense) mappings between algebraic groups and vector spaces. His dissertation, "Superposition in a Class of Nonlinear Systems" [Oppenheim, 1964] completed at MIT in May, 1964, developed a theory for nonlinear signal processing referred to as homomorphic systems. The use of such systems for signal processing was termed homomorphic filtering.

7.4.1 Homomorphic Systems

Homomorphic systems for convolution obey a generalized principle of superposition. The principle of superposition for conventional linear systems is

$$\mathcal{L}[x(n)] = \mathcal{L}[x_1(n) + x_2(n)] = \mathcal{L}[x_1(n)] + \mathcal{L}[x_2(n)] = y_1(n) + y_2(n) = y(n) \tag{7.56}$$

and

$$\mathcal{L}[ax(n)] = a\mathcal{L}[x(n)] = ay(n) \tag{7.57}$$

where $\mathcal{L}$ represents the linear operator and a is a scalar constant. The principle of superposition simply states that if an input signal is composed of a linear combination of elementary signals, then the output is a linear combination of corresponding outputs. A direct result of the principle of superposition is the fact that the output of a linear time-invariantsystem can be expressed as the convolution sum

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)\,x(k) = h(n) * x(n) \tag{7.58}$$

The * symbol will henceforth denote the operation of discrete-time convolution. By analogy with the principle of superposition for conventional linear systems, we can define a class of systems which obey a generalized principle of superposition where addition is replaced by convolution

$$\mathcal{H}[x(n)] = \mathcal{H}[x_1(n) * x_2(n)] = \mathcal{H}[x_1(n)] * \mathcal{H}[x_2(n)] = y_1(n) * y_2(n) = y(n) \tag{7.59}$$

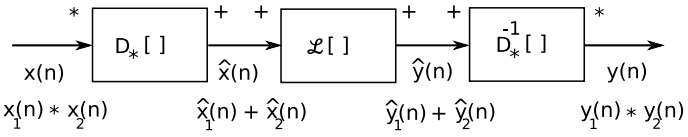

Figure 7.1: Canonic form for system for homomorphic deconvolution.

Systems that have the property expressed by equation ((7.59)) are termed "homomorphic systems for convolution". A homomorphic filter is simply a homomorphic system with the property that one component passes through the system unchanged, while the undesired component is removed; i. e., $y_1(n) = \delta(n)$. Any homomorphic system can be represented as a cascade of three homomorphic systems, as depicted in figure ((7.1))

The first system takes inputs combined by convolution and transforms them into an additive combination of corresponding outputs. The second system is a conventional linear system obeying the principle of superposition as given in equation ((7.56)). The third system is the inverse of the first system; i.e., it transforms signals combined by addition back into signals combined by convolution. The importance of the existence of such a canonic form for homomorphic systems lies in the fact that the design of such systems reduces to the problem of the design of a linear system. The system $D_*[]$ is called the characteristic system for homomorphic deconvolution and it is fixed in the canonic form of figure 7.1. Likewise, its inverse is also a fixed system. The characteristic system for homomorphic deconvolution obeys a generalized principle of superposition where the input operation is convolution and the output operation is ordinary addition. The properties of the characteristic system are defined as

$$D_*[x(n)] = D_*[x_1(n) * x_2(n)] = D_*[x_1(n)] + D_*[x_2(n)] = \widehat{x}_1(n) + \widehat{x}_2(n) = \widehat{x}(n) \tag{7.60}$$

Likewise, the inverse characteristic system D_*^{-1} is defined as

$$D_*^{-1}[\hat{y}(n)] = D_*^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] = D_*^{-1}[\hat{y}_1(n)] * D_*^{-1}[\hat{y}_2(n)] = y_1(n) * y_2(n) = y(n) \tag{7.61}$$

The mathematical representation of the characteristic system is dependent on the fact that we require that if the input is a convolution

$$x(n) = x_1(n) * x_2(n) \tag{7.62}$$

then the z-transform of the input is the product of the corresponding z-transforms.

$$X(z) = X_1(z) X_2(z) \tag{7.63}$$

From equation (7.60), it is clear that the z-transform of the output of the characteristic system must be an additive combination of z-transforms. Thus, the frequency domain behavior of the characteristic system for convolution must have the property that if a signal is represented as a product of z-transforms at the input, then the output must be a sum of corresponding output z-transforms. One approach to the representation of such a system is depicted in figure 7.2. This approach is based on the fact that the logarithm of a product can be defined so that it is equal to the sum of the logarithms of the individual terms. That is

$$\widehat{X}(z) = \log[X(z)] = \log[X_1(z) X_2(z)] = \log[X_1(z)] + \log[X_2(z)] \tag{7.64}$$

The logarithm must be defined so that it has the property that the logarithm of a product is equal to the sum of the logarithms, which is not always uniquely true for complex numbers. Here we

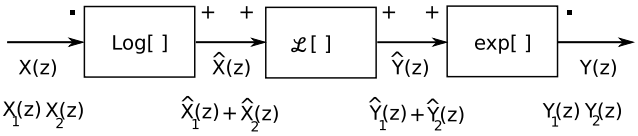

Figure 7.2: Frequency domain representation of a homomorphic system for convolution.

will be primarily concerned with ensuring it is valid when evaluated upon the unit circle (i.e., for $z = e^{j\omega}$), so an appropriate definition of the complex logarithm is

$$\widehat{X}\left(e^{j\omega}\right) = \log\left|X\left(e^{j\omega}\right)\right| + j\arg\left[X\left(e^{j\omega}\right)\right] \tag{7.65}$$

In this equation the real part causes no particular difficulty. However, problems of uniqueness arise in defining the imaginary part, which is simply the phase angle of the z-transform evaluated on the unit circle. One approach to dealing with the problem of uniqueness of the phase angle is to require that the phase angle be a continuous odd function of ω (this is not equivalent to the principal value of $\log(z)$, the one whose imaginary part lies in the interval $(-\pi, \pi]$). Ronald Schafer [Schafer, 1968] presents a detailed discussion of the conditions under which equations (7.64) and (7.65) hold.

7.4.2 Cepstrum

Historically, the cepstrum has its roots in the general problem of the deconvolution of two or more signals. This literature is rich and varied and encompasses linear prediction, predictive deconvolution, inverse filtering, and general deconvolution. Childers [Childers et al., 1977] presents a list of references on deconvolution methods in his comprehensive review of cepstrum-based processing. In what follows in this chapter, we will see that the power (or equivalently real) cepstrum was first developed for echo detection, while the (complex) cepstrum is concerned with the deconvolution of two signals (in speech processing the signals are usually a basic or fundamental wavelet and a train of impulses [Markel and Gray, 1976, Rabiner and Schafer, 1978]).

7.4.2.1 Historical Background

Allan Oppenheim tells us [Oppenheim and Schafer, 2004] that it was a fortuitous discussion in 1965 between Jim Flanagan of Bell Telephone Laboratories and himself that connected the work going on at MIT to the development of the cepstrum at Bell Laboratories. After hearing about homomorphic deconvolution from Oppenheim, Flanagan noted that the characteristic system for homomorphic convolution was reminiscent of the spectrum of the log of the spectrum (i.e., the cepstrum) as proposed by Bogert et al. Furthermore, he suggested looking at work by Michael Noll [Noll, 1964, Noll, 1967] in the Journal of Acoustical Society of America. Noll credits Manfred Schroeder (who was aware of the work of Bogert et. al.) with suggesting to him that it might be interesting to apply cepstrum analysis on a short-time basis to speech signals. In the Journal of Acoustical Society of America papers [Noll, 1967, Noll, 1973, Noll, 1964], Noll applied the cepstrum as a basis for pitch detection. The problem of pitch detection is very similar to detecting echo times in the sense that the basic speech model consists of representing speech as the convolution of the vocal tract impulse response with the quasi-periodic train of glottal pulses.

7.4.2.2 The Power Cepstrum

The power cepstrum was first described by Bogert et al. [Bogert et al., 1963] in 1963 as a heuristic technique for finding echo arrival times in a composite signal. Basically, these authors defined the

cepstrum (which we term the power cepstrum to avoid confusion with the complex cepstrum) of a function as the power spectrum of the logarithm of the power spectrum of that function.

These authors quickly showed (as we saw above in the opening of this section) that the effect of a delayed echo will manifest itself as a ripple in the log spectrum. The "frequency" of this ripple is easily determined by calculating the spectrum of the log spectrum wherein this "frequency" will appear as a peak. However, the units of "frequency" of this ripple in the log spectrum are in units of time; thus, the independent variable (abscissa) in the spectrum of the log spectrum is time. Other parameters were also observed to undergo similar transformations of units. To avoid confusion, Bogert et al. [Bogert et al., 1963] introduced the following now classical paraphrased terms according to a syllabic interchange rule

frequency ..... quefrency

spectrum ..... cepstrum

phase ..... saphe

amplitude ..... gamnitude

filtering ..... liftering

harmonic ..... rahmonic

period ..... repiod

along with others. Today the two most prevalent terms are cepstrum and quefrency, e.g., filtering in the cepstrum domain is usually called just that and not "liftering" as suggested by Bogert et al. [Bogert et al., 1963], but this can and often does lead to confusion. In practice the power cepstrum is effective if the wavelet and the impulse train, whose convolution comprise the composite data, occupy different quefrency ranges. In actuality, the power cepstrum does not exist for most signals; it is meaningful only when defined in a sampled data sense (as is the complex cepstrum) although attempts to extend it exist [Childers et al., 1977]. Thus the following definition is offered: the power cepstrum of a data sequence is the square of the inverse z-transform of the logarithm of the magnitude squared of the z-transform of the data sequence, as shown in equation 7.66. When this definition is evaluated on the unit circle, the result (except for the normalization factors associated with the power spectrum) is the same as that obtained with the Fourier transform. Thus we may write the power cepstrum as

$$\tilde{x}(nT) = \left[Z^{-1}\left(\log|X(z)|^2\right)\right]^2 = \left\{\frac{1}{2\pi j}\oint_C \log|X(z)|\, z^{n-1}dz\right\}^2 \tag{7.66}$$

where $X(z)$ is the z-transform of the data sequence $x(nT)$. Alternately, the definition could be changed to use the forward z-transform and/or the final squaring could be changed to magnitude squared. In actuality the final squaring operation in equation (7.66) is unnecessary and is frequently omitted for several reasons, but it has been used here to provide historical continuity with [Bogert et al., 1963]. When we omit the final squaring in equation (7.66) and evaluate it on the unit circle (which is equivalent to using the DFT to perform the calculation), the result is usually called real cepstrum [Noll, 1973, Kemerait, 1971].

The waveform of the basic wavelet cannot be recovered by processing the power cepstrum since the phase information is discarded. This latter situation is corrected with the complex cepstrum which we discuss in the next subsection along with the inversion process. The power cepstrum has been applied to seismic data [Bogert et al., 1963], sonar [LeBlanc, 1969],

speech [Noll, 1964, Noll, 1967, Noll, 1973, Noll, 1964], and the electroencephalogram (EEG) [Kemerait, 1972]. Its statistical properties have also been examined [Hassab and Boucher, 1976]. It is hopefully beneficial to point out that alternate viewpoints and, thus, subsequent terminologies have arisen since the original paper by Bogert et al. [Bogert et al., 1963]. These viewpoints have led to what might well be considered two lines of investigation:

1. the use of varying degrees of spectral whitening;
2. the attempts to devise methods for obtaining the phase relations of the wavelet with respect to the reference signal [Cohen, 1970].

We have seen that the occurrence of an echo in the time domain signal leads to what amounts to a spectral modulation (or ripple) in the frequency domain. The spectral whitening approach to echo detection considers the application of the logarithm a severe spectral whitener (rather than a mechanism to transform the product of two functions into the sum of the logarithm of the two functions as Bogert et al. intended).

7.4.2.3 The Complex Cepstrum

The complex cepstrum is an outgrowth of homomorphic system theory developed by Oppenheim [Oppenheim and Schafer, 1968, Oppenheim et al., 1968, Oppenheim, 1964, Oppenheim, 1969]. In fact, the power cepstrum is also a specific application of homomorphic system theory. The complex cepstrum has been investigated extensively [Kemerait, 1971, Kemerait, 1972, Noll, 1973, Schafer, 1968]. Since the complex cepstrum retains the phase information of the composite data, it can be used not only for echo detection but also wavelet recovery; this process is also known as homomorphic deconvolution or homomorphic filtering and has since been applied to seismic data [Cohen, 1970], speech [Oppenheim et al., 1968, Schafer, 1968, Oppenheim, 1969, Oppenheim and Schafer, 1968, Schafer and Rabiner, 1970], image processing [Oppenheim et al., 1968], and EEG analysis [Kemerait, 1971, Kemerait, 1972]. Formally, we define the complex cepstrum of a data sequence as the inverse z-transform of the complex logarithm of the z-transform of the data sequence [Childers et al., 1977, Schafer, 1968], i.e.,

$$\hat{x}(nT) = \frac{1}{2\pi j} \oint_C \log[X(z)]\, z^{n-1} dz \tag{7.67}$$

where $\hat{x}(0) = \log[x(0)]$ and $X(z)$ is the z-transform of the data sequence $x(nT)$. Frequently, $\hat{X}(z)$ is used to denote the $\log X(z)$; then $\hat{x}(nT)$, the complex cepstrum, is the inverse z-transform of $\hat{X}(z)$. The contour of integration lies within an annular region in which $\hat{X}(z)$ has been defined as single valued and analytic. If we have the convolution of two sequences as expressed in equation (7.62) in the time domain or in equation (7.63) in the frequency domain, then using the convention $\hat{X}(z) = \log X(z)$ we obtain equation (7.64). Further, if $\hat{x}_1$ and $\hat{x}_2$ occupy different quefrency ranges, then the complex cepstrum can be liftered (filtered) to remove one or the other of the convolved sequences. Since the phase information is retained, the complex cepstrum is invertible. Thus if $\hat{x}_2$ is rejected from $\hat{x}$ by liftering, then $\hat{x} = \hat{x}_1$ and we may then z-transform, exponentiate, and inverse z-transform to obtain the sequence $\hat{x}_1$, i.e., $\hat{x}_1$ and $\hat{x}_2$ have been deconvolved.

7.4.2.4 Phase Unwrapping

The computation of the complex cepstrum is complicated by the fact that the complex logarithm is multivalued. If the imaginary part of the logarithm is computed module 2π, i.e., evaluated as its principal value, then discontinuities appear in the phase curve. This is not allowed since $\log[X(z)]$ is the z-transform of $\hat{x}(nT)$ and thus must be analytic in some annular region of the z-plane. This problem may be rectified by making the following observations:

1. The imaginary part of $\log[X(z)]$ must be a continuous and periodic (evaluated on the unit circle) function of ω with period $2\pi/T$ since it is the z-transform of $\hat{x}(nT)$.
2. Since it is required that the complex cepstrum of a real function be real, it follows that the imaginary part of $\log[X(z)]$ must be an odd function of ω.

Subject to these conditions we may compute the unwrapped phase curve as follows [Schafer, 1968] (provided the phase is sampled at a rate sufficiently great to assure that it never changes by more than π between samples [Childers et al., 1977]): a correction sequence $C(k)$ is added to the modulo 2π phase sequence $P(k)$ where $C(k)$ is

$$C(0) = 0 \tag{7.68}$$

$$C(k) = \begin{cases} C(k-1) - 2\pi, & \text{if } P(k) - P(k-1) > \pi \\ C(k-1) + 2\pi, & \text{if } P(k-1) - P(k) > \pi \\ C(k-1), & \text{otherwise} \end{cases} \tag{7.69}$$

Alternately, the phase may be unwrapped by computing the relative phase between adjacent samples of the spectrum. These phases may be added to achieve a cumulative (unwrapped) phase for each point. Both methods have the drawback that the computation must be done sequentially. It is also noted that if the phase never changes by more than $\pi/2$ between samples, the phase modulo π could be computed and unwrapped with algorithms similar to the above. This is interesting since it is slightly easier to calculate the phase modulo π than the phase modulo 2π (the arctangent algorithm is simpler) and many signals have this property (though noise generally does not) [Childers et al., 1977].

Several other phase unwrapping procedures have been discussed, e.g., integrating the phase derivative [Schafer, 1968], an adaptive numerical integration procedure [Tribolet, 1977], and factorization of the z-transform [Steiglitz and Dickinson, 1977].

Phase unwrapping is unnecessary for the class of minimum phase signals, i.e., a sequence whose z-transform has no poles or zeros outside the unit circle, which implies that $\hat{x}(nT) = 0$ for $n < 0$ [Schafer, 1968]. The complex cepstrum of such a sequence is zero at negative quefrencies. Further, for $n > 0$ the complex cepstrum is identical to the real cepstrum (except for a factor of 2 and the squaring operation); for $n = 0$ the two cepstra are identical.

7.4.2.5 Relationship Between the Complex and Power Cepstra

Clearly the complex and power cepstra are closely related. The simple formal relationship can be obtained from equation (7.66) as follows:

$$\tilde{x}(n) = \left\{Z^{-1}\left[\log\left(X(z)\,X^*(z)\right)\right]\right\}^2 = \left\{Z^{-1}\left[\log\left(X(z)\right) + \log\left(X^*(z)\right)\right]\right\}^2 \tag{7.70}$$

Assuming that $x(n)$ is real and evaluating its z-transform on the unit circle, we find $X^*(z) = X\left(z^{-1}\right)$, thus we may write

$$\tilde{x}(n) = \left\{\frac{1}{2\pi j}\oint_C \log|X(z)|\, z^{n-1}dz + \frac{1}{2\pi j}\oint_C \log\left|X\left(z^{-1}\right)\right| z^{n-1}dz\right\}^2 \tag{7.71}$$

Letting $z' = z^{-1}$, we obtain

$$\tilde{x}(n) = \left\{\frac{1}{2\pi j}\oint_C \log|X(z)|\, z^{n-1}dz + \frac{1}{2\pi j}\oint_C \log|X(z')|\, z'^{-n-1}dz'\right\}^2 \tag{7.72}$$

Then by definition of the complex cepstrum in equation (7.67) we have

$$\tilde{x}(n) = [\hat{x}(n) + \hat{x}(-n)]^2 \tag{7.73}$$

Thus the power cepstrum is four times the square of the even part of the complex cepstrum. This also follows from the fact that the power cepstrum is the square of the inverse transform of twice the real part of the log spectrum; and, as was noted earlier, the power cepstrum contains no phase information. Equation (7.73) is of value since the power cepstrum is often superior to the complex cepstrum for echo arrival time estimation [Kemerait, 1972]. This is apparently due to the fact that the linear phase contribution (to be discussed below) of the imaginary part of the logarithm tends to mask the echo delay. There are probably other phase unwrapping errors as well as noise errors which contribute to this observation. A wavelet recovery (homomorphic filtering) system can easily compute both the power and complex cepstra. Finally, as was noted earlier, if the squaring operation in equation (7.73) is not performed, then the homomorphic filtering system can be used to obtain an estimate of the log power spectrum and in turn the power spectrum of the basic wavelet. Note that if this is one's objective (and not wavelet recovery), then the problems associated with phase unwrapping are not encountered.

7.4.2.6 The Fourier Transform Formulation of the Complex Cepstrum

The complex cepstrum can be defined as the inverse transform of the complex logarithm of the Fourier transform of the sequence $x(n)$

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}\left(e^{j\omega}\right) e^{j\omega n} d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\left(e^{j\omega}\right)\right| e^{j\omega n} d\omega + \frac{j}{2\pi}\int_{-\pi}^{\pi} \arg\left[X\left(e^{j\omega}\right)\right] e^{j\omega n} d\omega \tag{7.74}$$

the sequence $c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\left(e^{j\omega}\right)\right| e^{j\omega n} d\omega$ is usually termed the real cepstrum and can be shown to be equal to the even part of the complex cepstrum $\widehat{x}(n)$ because of the symmetry properties of the Fourier transform.[2]

The real cepstrum can be defined as the inverse transform of the logarithm of the magnitude Fourier transform of the sequence $x(n)$

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)| e^{j\frac{2\pi}{N}kn} \tag{7.75}$$

The doctoral dissertation "Echo Removal by Discrete Generalized Linear Filtering" [Schafer, 1968] by Ronald Schafer at MIT in 1968, focused on the issues of the discrete-time formulation of the complex cepstrum, phase computation, recursion relations, and applications to echo removal from speech. In our development of the complex cepstrum, a variety of alternate implementations of the complex cepstrum and relationships between the power cepstrum and complex cepstrum were developed, both in general and for minimum-phase, maximum-phase and all-pass sequences. The work also led to the interpretation in the cepstral domain of the Hilbert transform relationship between Fourier transform magnitude and phase for minimum phase signals.

The cepstral coefficients contain frequency information about the log magnitude spectrum, such that each coefficient can be interpreted as a measure of the energy of the log-magnitude spectrum on increasing frequency bands. It is possible to obtain an estimate of the spectral envelope by liftering (filtering in the cepstral domain) the higher order coefficients, keeping only the coefficients that contain lower-frequency information. This technique, termed cepstral smoothing, can give a good

[2] If $x(n)$is real, $|X(\omega)|$is even and $\arg[X(\omega)]$ is odd. If $x(n)$ is even, $X(\omega)$ is even. The real part of the Fourier transform is the Fourier transform of the even part of the sequence $x(n)$

approximation of the spectral envelope of $|X(k)|$ when used iteratively in a method called true envelope estimation [Röbel et al., 2007].

Appendix 7.4 presents the phase cepstrum and discusses some operations such as windowing and zero-padding in the cepstral domain.

7.4.3 Cepstral Smoothing

The cepstrum can be used to estimate the spectral envelope of speech or musical instrument sounds (or any other signal) in a technique called cepstral smoothing. Cepstral smoothing can be interpreted in the light of the source-filter model, where the (real or complex) cepstrum is used to deconvolve the pitch information (source) from the spectral envelope (filter) by liftering, or simply filtering the cepstrum. The basic idea is to eliminate any quefrencies above that corresponding to the fundamental period of the signal. Using the real cepstrum as defined in equation 7.75 and regarding the log magnitude spectrum as a signal, we can interpret each cepstral coefficient as a measure of the energy present in discrete frequency bands of that signal. Low-pass filtering the cepstrum (also called liftering) would result in a smoother version of the log magnitude spectrum, given by

$$C(k) = \sum_{n=0}^{N-1} w(n)\, c(n) \exp\left(\frac{-j2\pi kn}{N}\right) \tag{7.76}$$

where $C(k)$ is the smoothed spectrum (corresponding to the spectral envelope estimation) and $w(n)$ is a low-pass window in the cepstral domain usually defined as

$$w(n) = \begin{cases} 1, & |n| < n_c \\ 0.5, & |n| = n_c \\ 0, & |n| > n_c \end{cases} \tag{7.77}$$

where n_c is the cutoff quefrency. If we only want to represent the spectral envelope, discarding information about the partials we should set the cutoff quefrency below the period of the signal. One major drawback of this operation is that we discard spectral energy when setting cepstral coefficients to zero. The result is a smooth curve $C(k)$ that is always below the peaks of the log magnitude spectrum. Figure 7.3 illustrates the cepstral smoothing technique. In figure 7.3 we see the original spectrum and the resultant spectral envelope curve. We should notice that, even though the spectral envelope curve is a smooth curve that follows the amplitude of the magnitude spectrum, it does not approximately match the peaks. Therefore, cepstral smoothing does not give satisfactory results according to some definitions of the spectral envelope curve. We will see in this chapter that the "true envelope" estimator uses cepstral smoothing in an iterative procedure to overcome this problem. In the "true envelope" estimation technique, the aim is to fit a spectral envelope curve that approximately matches the peaks of the log magnitude spectrum. But first, we will see an alternative way of using the cepstrum to estimate the spectral envelope of discrete spectra in a method called discrete cepstrum.

7.5 Discrete Cepstrum

The discrete cepstrum is a technique to solve the problem of estimating a continuous frequency-envelope when the value of this envelope is specified only at discrete frequencies. This problem arises naturally in sinusoidal analysis/synthesis systems in which the signal is modeled as the discrete sum of sinusoids. Such a continuous envelope is needed for example for pitch-scale modifications of speech signals (because the amplitudes of the modified harmonics must be extrapolated from the knowledge of the original harmonic amplitudes).

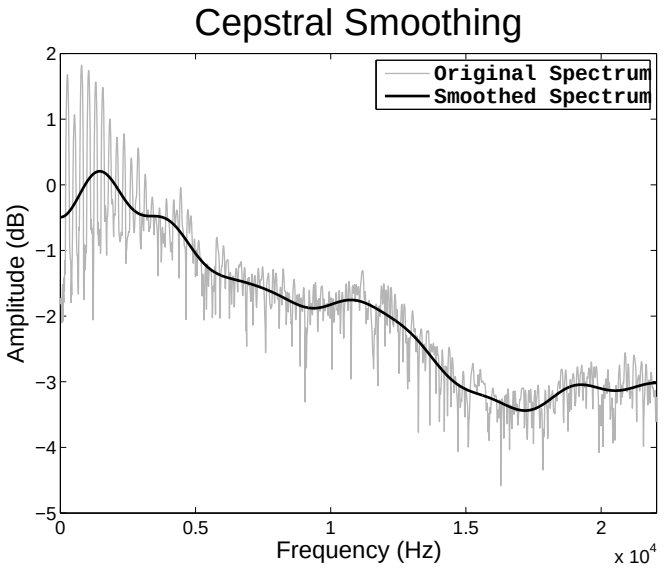

Figure 7.3: Cepstral Smoothing. The figures illustrates the cepstral smoothing technique. In the figure we see the log magnitude spectrum and the resultant spectral envelope curve.

Let us suppose that $X(\omega)$ is the power spectrum of the signal to be analyzed, $S(\omega)$ is the power spectrum of the hypothetical source, and $d(Y,Z)$ is a spectral distance. Now let us consider the class C of power spectra $P(\omega)$ that are candidates to model the filter, such that $X(\omega) \sim S(\omega) P(\omega)$. We search for the parameters p_m of $P(\omega)$ in C that define a model that minimizes $d(X, SC)$. When we consider that $S(\omega) = 1$, $\forall\omega$ and define $d(Y,Z)$ by the quadratic error between the log spectra and the class C of power spectra $P(\omega)$ defined by

$$P(\omega) = \prod_{k=0}^{L-1} e^{p_k cos\omega k} \tag{7.78}$$

we obtain the cepstrum because

$$\log|P(\omega)| = \sum_{k=0}^{L-1} p_k \cos(\omega k) = \sum_{k=0}^{L-1} (2 - \delta_{k0})\, c_k \cos(\omega k) = c_0 + 2\sum_{k=1}^{L-1} c_k \cos(2\pi f k) \tag{7.79}$$

which is consistent with the definition of the real cepstrum given by equation (7.75). When applied to discrete spectra this method gives erroneous results if the order of the model is not negligible when compared to the number of spectral peaks. Now, we assume that the power spectra $S(\omega)$ and $X(\omega)$ are defined on the same discrete set $\Omega = \{\omega_n, n = 1 \ldots N\}$, such that they can be described as a set of partials at frequencies ω_n with amplitudes s_n and x_n, respectively. This can be written as

$$S(\omega) = \sum_{n=1}^{N} s_n \delta(\omega - \omega_n) \tag{7.80}$$

and

$$X(\omega) = \sum_{n=1}^{N} x_n \delta(\omega - \omega_n) \tag{7.81}$$

where $\delta(\omega)$ denotes the Dirac delta distribution and the spectral envelope domain considered is defined by equation (7.78). Next we adopt the distance given by the quadratic error between the log spectra with spectral weights h_n that are strictly positive real numbers and that are used to obtain a better fit at certain discrete frequencies.

$$\epsilon(c) = \sum_{n=1}^{N} h_n \left[\log|P(\omega_n)| - \log(x_n)\right]^2 = \sum_{n=1}^{N} h_n \left[\sum_{k=0}^{L-1} (2 - \delta_{k0})\, c_k \cos(\omega k) - \log(x_n)\right]^2 \tag{7.82}$$

Galas [Galas and Rodet, 1990] states that this error measure is "rather pertinent from the perceptual point of view," probably because it uses the logarithm of the power spectrum. D'haes [D'haes and Rodet, 2003] states that comparing spectral envelopes is very interesting since it is related to the timbral similarity between two short time spectra in a trivial way. The fact that the perceived loudness of a human listener is approximately logarithmic with the signal amplitude suggests that the square difference between the log magnitude spectra can be used to express the perceived similarity. This difference, computed for two spectral envelopes $|H_1(\omega)|$ and $|H_2(\omega)|$ defined by two vectors of cepstrum coefficients c_1 and c_2 respectively, is equivalent to the Euclidean distance between these vectors. We should bear in mind that there are other perceptually motivated spectral distance measures, such as the Itakura-Saito distance, which is a measure of the perceptual difference between a spectrum $X(\omega)$ and its approximation $\tilde{X}(\omega)$. The Itakura -Saito distance is defined as

$$d\left(X(\omega), \tilde{X}(\omega)\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{X(\omega)}{\tilde{X}(\omega)} - \log\left(\frac{X(\omega)}{\tilde{X}(\omega)}\right) - 1\right] d\omega \tag{7.83}$$

It is important to notice that the Itakura-Saito distance is not symmetric, which means that $d\left(X(\omega), \tilde{X}(\omega)\right) \neq d\left(\tilde{X}(\omega), X(\omega)\right)$.

So the discrete cepstrum estimation can be formulated as an error minimization problem by imposing the condition that the partial derivatives of ϵ with respect to the cepstral coefficients c must be equal to zero, resulting in

$$\frac{\partial \epsilon(c)}{\partial c_k} = \sum_{n=1}^{N} h_n \left(\sum_{k=0}^{L-1} (2 - \delta_{k0})\, c_k \cos(\omega k)\right) \cos(k\omega_k) = 0 \tag{7.84}$$

This formulation leads to a system that can be solved as a matrix equation of the form $Ac = a$, with

$$A_{ij} = \sum_{k=0}^{N-1} h_k (2 - \delta_{k0}) \cos(i\omega_k) \cos(j\omega_k) \tag{7.85}$$

P is the vector of cepstral coefficients that we are looking for, and B is the vector given by

$$a_{ij} = \sum_{k=0}^{N-1} h_k \log(x_k) \cos(i\omega_k) \tag{7.86}$$

We can compute A more efficiently by using an intermediate vector r defined as

$$r_i = \frac{1}{2} \sum_{k=0}^{N-1} h_k \cos(i\omega_k) \tag{7.87}$$

and then

$$a_{ij} = r_{i+j} - r_{i-j} \tag{7.88}$$

Galas [Galas and Rodet, 1990] proposes to solve the matrix equation using the Cholesky algorithm and observes that the results depend on the order selection, that is, on the number L of cepstral coefficients. D'haes [D'haes and Rodet, 2003] observes that since the cepstrum coefficients are computed from a set of linear equations, the computation of L coefficients requires at least an equal number of detected peaks. Overfitting occurs when the number of coefficients equals the number of peaks. This can easily be avoided by lowering the number of coefficients. However, when too few coefficients are used, a low pass filtered envelope is obtained that fails to match the peaks accurately. Galas [Galas and Rodet, 1990] states that one of the reasons for this phenomenon is that the formulation assumes that we know precisely the position of the spectral peaks. He proposes to replace the position (ω_i, x_i) of every spectral peak by a probability distribution $Pr_n(\omega, x)$, and then replace the first error condition by its mathematical expectation. Assuming $s_n = h_n = 1$, $\forall n$, this can be expressed as

$$\epsilon(c) = \sum_{n=0}^{N-1} \iint Pr_n(\omega, x) \left[\log|P(\omega_n)| - \log x_n\right]^2 d\omega dx \tag{7.89}$$

Galas states [Galas and Rodet, 1990] that if $Pr_n(\omega, x)$ has no particular properties, computation of the parameters can be done by using a sampling of $Pr_n(\omega, x)$. Every spectral peak (ω_n, x_n) is then replaced by a set of peaks (ω_k, x_k) with weights $h_k = Pr_k(\omega_k, x_k)$. Particularly, if we choose a Gaussian distribution for $Pr_n(\omega, x)$, it is possible to compute explicitly the corresponding matrix equation $AC = B$. Later on, Cappé [Cappé and Moulines, 1996] proposed a regularization technique that tries to overcome many of the shortcomings of the original discrete cepstrum formulation.

Appendix 7.5 contains yet another proposal for computing the discrete cepstrum using regularized estimation [Cappé and Moulines, 1996, Cappé et al., 1995]. However, we will see in the next section that "true envelope" estimation leads to better results when we want a spectral envelope curve that approximately matches the peaks of the magnitude spectrum.

7.6 True Envelope

The true envelope estimator [Villavicencio et al., 2007] has been shown to outperform linear prediction [Makhoul, 1975] or cepstral methods such as discrete cepstrum [Cappé and Moulines, 1996] both in terms of accuracy and ease of model order selection. Recently the iterative procedure has been significantly improved such that the computational costs are similar to the costs of the Levinson recursion such that real time processing can be achieved [Villavicencio et al., 2007]. True envelope estimation is based on cepstral smoothing of the log amplitude spectrum and the resulting estimation can be interpreted as the best band limited interpolation of the major spectral peaks in such a way that the peak matching is maximized and inter-peak valleys are avoided [Villavicencio et al., 2007].

7.6.1 True Envelope Estimation

Let $X(k)$ be the K-point DFT of the signal frame $x(n)$ and $C_i(k)$ the smoothed spectrum at iteration i. The algorithm then iteratively updates the resulting spectral envelope $A_i(k)$ with the maximum of the original spectrum and the current spectral envelope $C_{i-1}(k)$

$$A_i(k) = \max\left(\log|X(k)|, C_{i-1}(k)\right) \tag{7.90}$$

and applies cepstral smoothing to $A_i(k)$ to obtain $C_i(k)$. The procedure is initialized setting $A_0(k) = \log|X(k)|$ and starting the cepstral smoothing to obtain $C_0(k)$. Figure 7.4 illustrates the "true envelope" estimation at different iterations i. On the left-hand side we see the original magnitude spectrum $\log|X(k)|$ and the smoothed spectrum $C_i(k)$ at the indicated iteration. On the right-hand side, we see the smoothed magnitude spectrum corresponding to $A_i(k)$ used in the next iteration.

7.6.2 Optimal Order Selection

The order of the cepstral representation of the spectral envelope is the number of cepstral coefficients we keep in the cepstral smoothing procedure, and as such is proportional to the fundamental frequency of the original signal. The optimal order should give a spectral envelope that follows the overall shape of the filter without representing the harmonic structure of the spectrum. In order to estimate the optimal order, we use the source-filter model and think of the spectrum as the result of the interaction of two components, represented by the source, an input signal that contains information about the frequencies of the partials and the filter that shapes the source spectrum. According to this model, the spectral envelope represents the filter that has been excited by the source. For near harmonic sources, the resulting spectrum will be quasi-harmonic. In terms of the interaction between source and filter, we can think of the resulting harmonics sampling the filter with a sampling rate that depends on the fundamental frequency of the source spectrum. According to the sampling theorem, we must sample the filter with at least twice the maximum frequency present in that signal. If we assume that the spectral envelope should not contain information about the harmonic structure of the spectrum, the maximum frequency present in that signal is the fundamental frequency F_0, such that the related Nyquist frequency (assuming a sampling rate of F_S) is $F_S = 2F_0$. This formula provides a simple way of selecting the cepstral order because higher sampling frequencies would reveal (maybe partially) information about the harmonic structure of the spectrum and lower sampling frequencies would smooth out the spectral envelope, not revealing information about the (formant) peaks. We can therefore postulate the near optimal order of the cepstrum given only that the maximum frequency difference between two spectral peaks that carry envelope information is known. If the difference between those peaks is Δ_F then the cepstral order should be

$$\hat{O} = \frac{F_S}{2\Delta_F} = \alpha\frac{F_S}{F_0},\ \alpha = 0.5 \tag{7.91}$$

While the optimal order, that is the order that provides an envelope estimate with minimum error, depends on the specific properties of the envelope spectrum, the order selection according to equation 7.91 is reasonable for a wide range of situations and the resulting error is generally rather close to the one obtained with the optimal order.

In this work I have chosen to estimate the spectral envelopes using true envelope and manipulate them with line spectral frequencies (LSFs). The next section introduces LSFs and how to convert from the cepstral coefficients resulting from the true envelope (TE) estimation to the LSFs representation used to manipulate the spectral envelopes.

7.7 Alternative Spectral Envelope Representations

Makhoul [Makhoul, 1975] proposes the following list of possible sets of parameters that characterize uniquely the all-pole filter $H(z)$ or its inverse $A(z)$.

1. Impulse response of the inverse filter $A(z)$, i.e., predictor parameters $a(k)$, $1 \leq k \leq p$. Note that the first $p+1$ coefficients uniquely specify the filter.

2. Autocorrelation coefficients of $a(k)$, $\rho(i)$, $0 \leq i \leq p$, as defined in equation A.20.

3. Spectral coefficients of $A(z)$, $\Gamma_i = \rho(0) + 2\sum_{j=1}^{p} \rho(j) \cos \frac{2\pi ij}{2p+1}$, $0 \leq i \leq p$ where $\rho(i)$ are as defined in equation A.20. In other words, Γ_i is obtained from $\{\rho(i)\}$ by a discrete Fourier transform.

4. Cepstral coefficients of $A(z)$ as defined in equation 7.74. There is an iterative method for the computation of the cepstral coefficients directly from the predictor coefficients. I will reproduce the derivation in section 7.7.2.

5. Poles of $H(z)$ or zeros of $A(z)$, denoted $z(k)$, $1 \leq k \leq p$, where $\{z(k)\}$ are either real or form complex conjugate pairs. Conversion of the roots to the s plane can be achieved by setting each root $z(k) = e^{s(k)T}$, where $s(k) = \sigma(k) + j\omega(k)$ is the corresponding pole in the s plane, and T is the sampling period. If the root $z(k) = z_r(k) + jz_i(k)$, where $z_r(k)$ and $z_i(k)$ are respectively the real and imaginary parts of $z(k)$, then we have $\omega(k) = {}^1/_T \arctan({}^{z_i(k)}/_{z_r(k)})$ and $\sigma(k) = {}^1/_{2T} \log\left(z_r^2(k) + z_i^2(k)\right)$.

6. Reflection coefficients k_i, $1 \leq i \leq p$, which are obtained as a byproduct of the solution of the autocorrelation normal equations, as in equation 7.36, or from the backward recursion A.6.

Some of the preceding sets of parameters have $p+1$ coefficients while others have only p coefficients. However, for the latter sets the gain G needs to be specified as well, thus keeping the total number of parameters as $p+1$ for all the cases. For purposes of data transmission, one is usually interested in recovering the predictor coefficients from the parameters that are chosen for transmission. Each representation has its own properties that depend on the application. Quantization and interpolation will be discussed later in section 7.7.3.

It is unknown whether Makhoul omitted line spectral frequencies (LSFs) for some reason or was just unaware of their proposal by Itakura in 1975 [Itakura, 1975], the same year Makhoul published his seminal paper on linear prediction. It was known that the linear predictor coefficients (LPC) of speech signals can be transformed into a "pseudo" vocal tract area function whose boundary conditions are a complete opening at the lips and a matching resistance termination at the glottis. Itakura realized that if the boundary condition at the glottis is replaced by a complete opening or a complete closure, all the poles of the resulting system function will move onto the unit circle in the z plane. Using this fact, Itakura proposed to describe the original LPCs by two sets of pole frequencies corresponding to the two new boundary conditions at the glottis, or a set of frequency-residue pairs corresponding to either set of poles. This representation is called line spectral frequencies (LSF) or line spectral pairs (LSP).

7.7.1 Line Spectral Frequencies

LSFs collectively describe the two resonance conditions arising from an interconnected tube model of the human vocal tract. This includes mouth shape and nasal cavity, and forms the basis of the underlying physiological relevance of the linear prediction representation [Paliwal, 1992]. The two resonance conditions are those that describe the vocal tract being either fully open or fully closed

at the glottis, respectively. The model in question being constructed from a set of equal-length but different diameter tubes, with the source end either closed or open. The two conditions give rise to two sets of resonant frequencies, with the number of resonances in each set being determined by the number of joined tubes (which in turn is a function of the order of the analysis system). The resonances of each condition are the odd and even line spectra, respectively, and are interleaved into a monotonically increasing set of LSFs. In reality, the human glottis opens and closes rapidly during voiced speech: it is neither fully closed nor fully open over an analysis frame. Hence actual resonances occur at frequencies located somewhere between the two extremes of odd and even LSP condition. Nevertheless, this relationship between vocal resonance and LSP position endows them with a significant interpretation which we will build upon as we review the representation.

Figure 7.5 illustrates LSPs overlaid on a power spectrum plot. The 10 vertical lines were drawn at the LSFs, and show the odd (solid) and even (dashed) frequencies being interleaved. Both the lines and the spectrum were derived from the same set of linear prediction parameters which were in turn obtained from 10th-order linear predictive analysis of a 20ms frame of voiced speech. Apart from the natural interleaving of the line frequencies, it is notable that peaks in the underlying spectrum of figure 7.5 tend to be bracketed by a narrow pair of lines. By contrast, local minima in the spectrum tend to not have LSPs overlaid nearby. This relationship between line location and spectral resonance is one reason for the popularity of LSPs for the analysis, classification and transmission of speech.

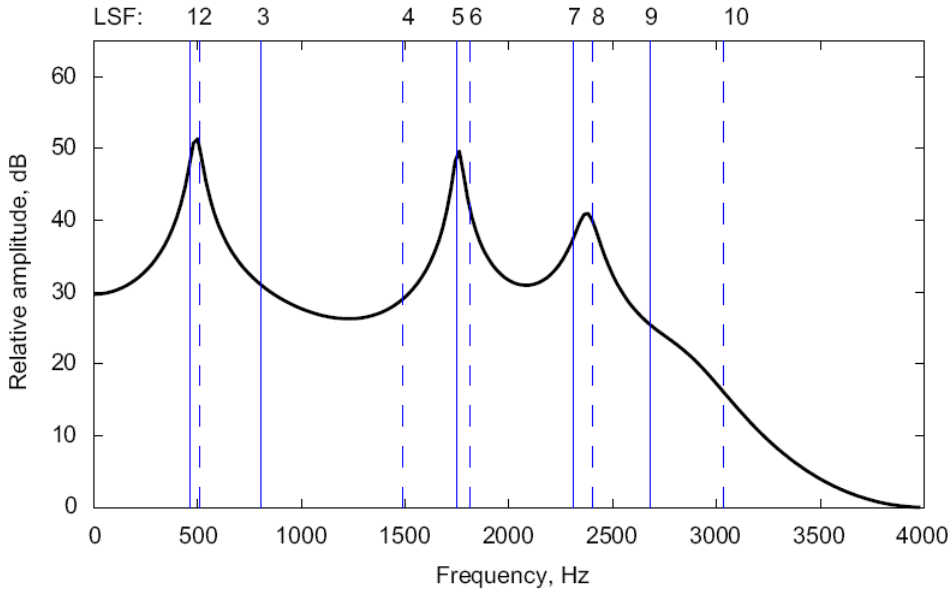

Figure 7.5: An example LPC spectrum overlaid with the corresponding vertical LSP frequencies. Odd lines are drawn solid and even lines are drawn dashed. After McLoughlin [McLoughlin, 2008]

Concerning vocabulary, the abbreviation LSP refers in some sources to Line Spectrum Pair polynomials and in others to line spectral polynomials. Moreover, the operation yielding LSP polynomials has been called LSP transformation, LSP decomposition or simply LSP. Since there is virtually no difference in meaning, we will use all forms interchangeably. The LSFs, which refer to the angle (frequency) of the zeros of LSP polynomials, are sometimes loosely used to refer to LSP methods in general, but we prefer to use LSFs to denote the actual angle or frequency only. The LSP decomposition is based on a transformation of a polynomial to its symmetric

and antisymmetric parts, and many different names have appeared to denote these symmetries. According to our understanding the following terms are equivalent: symmetric, selfreciprocal, and palindromic polynomial. The same terms apply for the antisymmetric polynomial but with a prefix of either "anti", "skew", or "conjugate". A polynomial $A(z)$ that has all its zeros inside the unit circle is said to be minimum-phase, or equivalently, its inverse $A^{-1}(z)$ is said to be stable. If it has all zeros outside the unit circle, it is maximum-phase and if the zeros are on the unit circle it is sinusoidal.

7.7.1.1 The LSP Representation

LSFs are derived from the linear predictive coding (LPC) filter representing vocal tract resonances in analyzed speech, for M^{th}-order analysis

$$A(z) = 1 + \sum_{m=1}^{M} a_m z^{-m} \tag{7.92}$$

Following the formalization in Appendix E.1, a polynomial $A_s(z)$ of order M is said to be palindromic if it has real coefficients $\{a_s(m)\}$ and the following relation holds

$$A_s(z) = z^{-M} A_s\left(z^{-1}\right) \tag{7.93}$$

Similarly, a polynomial $A_a(z)$ is antipalindromic if it has real coefficients $\{a_s(m)\}$ and the following relation holds

$$A_a(z) = -z^{-M} A_a\left(z^{-1}\right) \tag{7.94}$$

We will define two $(M+1)^{th}$-order polynomials related to $A(z)$ which we shall name $P(z)$ and $Q(z)$. These polynomials represent an interconnected tube model of the human vocal tract. They correspond in turn to complete closure at the source end of the interconnected tubes and a complete opening, defined by the $(M+1)^{th}$ extra term. The two polynomials are created from the LPC polynomial with an extra feedback term being positive to model energy reflection at a completely closed glottis, and negative to model energy reflection at a completely open glottis

$$P(z) = A(z) + z^{-(M+1)} A\left(z^{-1}\right) \tag{7.95}$$

$$Q(z) = A(z) - z^{-(M+1)} A\left(z^{-1}\right) \tag{7.96}$$

The roots of these two polynomials are the set of LSFs, ω_k. These relate back to the palindromic and antipalindromic polynomials.

In the original model, the source end is the glottis, which is neither fully open nor fully closed during the period of analysis, and thus the actual resonance conditions encoded in $A(z)$ are a linear combination of the two boundaries. In fact this is simply stated as

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{7.97}$$

It can be shown that the complex roots of the polynomials will lie on the unit-circle in the z-plane if the original LPC filter was stable [McLoughlin, 2008], and alternate in order around the unit circle. It is also important to note that any equivalent size set of roots that alternate in this way around and on the unit circle will represent a stable LPC filter, and we shall consider the implications of this for LSP-based adjustment and processing.

According to equation E.3, if we denote the set of complex roots as $\{\phi_k\}$, then the LSFs are determined from equations 7.95 and 7.96

$$\omega_k = \arctan\left(\frac{\Re\{\phi_k\}}{\Im\{\phi_k\}}\right) \tag{7.98}$$

ω_k are then the LSFs expressed in radians.

The polynomials $P(z)$ and $Q(z)$ have trivial zeros at $z = \pm 1$, like stated in section E.1. Conversion from LSFs back to LPCs is a simple process [McLoughlin, 2008], since we can easily use the ordered LSFs ω_k to recreate the polynomials that they are roots of, namely, if M is even, $P(z)$ and $Q(z)$ can be written as

$$P(z) = \left(1 + z^{-1}\right) \prod_{k=2,4,\cdots,M} \left(1 - 2z^{-1}\cos\omega_k + z^{-2}\right) \tag{7.99}$$

$$Q(z) = \left(1 - z^{-1}\right) \prod_{k=1,3,\cdots,M-1} \left(1 - 2z^{-1}\cos\omega_k + z^{-2}\right) \tag{7.100}$$

when M is odd, the relations are given by

$$P(z) = \prod_{k=2,4,\cdots,M} \left(1 - 2z^{-1}\cos\omega_k + z^{-2}\right) \tag{7.101}$$

$$Q(z) = \left(1 - z^{-2}\right) \prod_{k=1,3,\cdots,M-1} \left(1 - 2z^{-1}\cos\omega_k + z^{-2}\right) \tag{7.102}$$

so that we can easily retrieve $A(z)$ from the polynomials $P(z)$ and $Q(z)$.

7.7.1.2 Properties of Line Spectral Pair Polynomials

Bäckström gives a comprehensive review of the properties of LSFs [Backström and Magi, 2006]. In brief, if $A(z)$ is minimum phase, the roots of $P(z)$ and $Q(z)$ are on the unit circle, are real, interleaved with each other, and always lead to stable envelopes when arranged in ascending order [Backström and Magi, 2006]. Both palindromic and antipalindromic polynomials, as expressed in equations 7.93 and 7.94 are linear-phase FIR filters when interpreted as transfer functions [Backström and Magi, 2006]. It follows that if z_m is a root of a palindromic or antipalindromic polynomial, then also $z = z_m^{-1}$ must be a root. Zeros of palindromic and antipalindromic polynomials can be in one of the four possible categories

1. root quadruples symmetric to the unit circle and real axis $\left[z_m, \bar{z}_m, z_m^{-1}, \bar{z}_m^{-1}\right]$, where $\bar{z}_i$ stands for the complex conjugate of z_i;
2. root pairs on the unit circle summetric to the real axis $[z_m, \bar{z}_m]$;
3. root pairs on the real axis symmetric to the unit circle $\left[z_m, z_m^{-1}\right]$;
4. trivial zeros at $z_m = \pm 1$.

Like equations 7.99, 7.100 and 7.101, 7.102 suggest, simple trivial zeros of palindromic and antipalindromic polynomials depend on the order M of the polynomial.

Bäckström [Backström and Magi, 2006] defines the relation $\doteqdot$, such that $F_1(z) \doteqdot F_2(z)$ means that $F_1(z)$ and $F_2(z)$ interlace on the unit circle and $F_1(z)$ has a root-pair closer to the angle zero, or $z = 1$. The formal definition is as follows: Two polynomials with real coefficients $F_1(z)$ and $F_2(z)$ (not necessarily of the same order) are interlaced on the unit circle if

1. all zeros $z = z_m$ of $F_k(z)$ are on the unit circle, that is, $F_k(z_m) = 0 \Leftrightarrow |z_m| = 1$;

2. zeros of $F_1(z)$ and $F_2(z)$ are simple and distinct, with the exception of possible simple trivial zeros at $z = \pm 1$;

3. the N non-trivial zeros $z \neq \pm 1$ of $F_1(z)$ and $F_2(z)$ are interlaced on both the upper and lower halves of the unit circle, that is, the zeros $z_i^{(j)} = \exp\left(j2\pi\omega_k^{(m)}\right)$ of $F_k(z)$ have $-\pi < \cdots < 2\pi\omega_{N/2-1}^{(2)} < 2\pi\omega_{N/2-1}^{(1)} < 2\pi\omega_{N/2}^{(2)} < 2\pi\omega_{N/2}^{(1)} < 0$ and $0 < 2\pi\omega_{N/2+1}^{(1)} < 2\pi\omega_{N/2+1}^{(2)} < 2\pi\omega_{N/2+2}^{(1)} < 2\pi\omega_{N/2+2}^{(2)} < \cdots < \pi$. Note that N is always even because we have omitted trivial zeros.

Next, Bäckström [Backström and Magi, 2006] proves the most famous property of LSP, the *intra-model interlacing* property, that says that if $A(z)$ is a polynomial with real coefficients and all its roots are inside the unit circle, then the roots of the LSP polynomials defined in equations 7.95 and 7.96 are interlaced on the unit circle $P(z) \div Q(z)$. Conversely, if the zeros of two polynomials with real coefficients of the same degree, one palindromic and the other antipalindromic, are interlaced, then their sum always has all zeros within the unit circle.

Since the roots of LSP polynomials lie on the unit circle, they can, in principle, be readily found. Moreover, the zeros of the LSP polynomials define the polynomial unambiguously up to scaling and we can reconstruct the LSP polynomials from their zeros (and scaling coefficients) and thereby obtain the original $A(z)$ as well. The zeros can, in turn, be represented by their angles only, since they lie on the unit circle. Finally, the angles are bounded and if the ordering property is ensured, the minimum phase property of the reconstructed $A(z)$ is retained. It is therefore this theorem that justifies the use of LSP in speech coding.

LSFs also present the useful tendency to be located where the peaks of the envelope they represent are. Figure 7.5 shows that each pair tends to be close together when near a peak of the spectral envelope and far apart when not, depicting another useful property of LSFs. The closer the line spectrum pair is, the narrower the peak.

7.7.1.3 Modification of Line Spectral Frequencies

Based on these properties of LSFs, McLoughlin [McLoughlin, 2008] exemplifies how we can manipulate the LSFs to produce small changes in the shape of the spectral envelope and Morris [Morris and Clements, 2002] presents a method for modifying formant peak locations and bandwidths in the line spectrum domain. Figure 7.6 shows the original spectral envelope in grey and a modified envelope (solid line) with its corresponding LSFs. Since there are LSF pairs that correspond roughly to specific spectral peaks, we generally can make changes to a specific peak without changing much the overall spectral envelope.

A comparison of the original and final spectra shows differences in the immediate frequency regions of the lines that were changed most. From these observations it has been found possible to alter the values of particular LSPs to change the underlying spectral information which they represent. Methods used to achieve these alterations are shown to have great potential for enhancement of speech in the presence of noise. To illustrate some of the modifications possible, figure 7.6 plots the original spectrum of figure 7.5 in gray, and a power spectrum derived from an altered set of LSPs as drawn. The LSP adjustments made to cause these spectral changes were namely: The separation of line pair {1:2} has been increased, resulting in a wider, lower amplitude spectral peak between them. The separation of line pair {5:6} has been decreased, and the pair has also been translated slightly upward in frequency, causing a sharper peak between them, now at a higher frequency. Line 10 has been moved closer to the Nyquist frequency of 4 kHz, inducing a spectral peak at that frequency. All of the line alterations shown were performed manually.

It was probably Paliwal again [Paliwal, 1992] who first reported that the effects, on the underlying spectrum, of modifying a line are predominantly confined to the immediate frequency region of that line. However, amplitude changes in one region will always cause compensatory power redistribution in other regions. Despite this, as long as line alterations are minimal, the effects on other spectral regions can be limited. This is saying that, for small movements, and small movements only, localized spectral adjustments can be made through careful LSP manipulation. The example of figure 7.5 shows a spectrum of voiced speech. The three spectral peaks represent formants, and as such we can see that the operations we performed have affected those formants. In fact, LSP operations have demonstrably altered formant bandwidths and positions. The LSP operations to derive the changes shown in figure 7.5 can be formalized as follows. If ω_k are the LSP frequencies and $\omega\prime_k$ the altered frequencies, then narrowing line pair {k:k+1} by degree α would be achieved by

$$\omega_k^{'} = \omega_k + \alpha\left(\omega_{k+1} - \omega_k\right) \tag{7.103}$$

$$\omega_{k+1}^{'} = \omega_{k+1} - \alpha\left(\omega_{k+1} - \omega_k\right) \tag{7.104}$$

and increasing the frequency of line k by degree γ may be achieved with

$$\omega_k^{'} = \omega_k + \omega_k \frac{(\gamma - 1)\left(\pi - \omega_k\right)}{\pi} \tag{7.105}$$

When altering the line positions it is important to avoid forming unintentional resonances by narrowing the gaps between lines that were previously separated. This problem may be obviated either by moving the entire set of LSPs or providing some checks to the adjustment process. In the former case, movement of lines 1 and 10 closer to angular frequencies of 0 and p may also induce an unintentional resonance. Equation 7.105, designed for upward shifting, progressively limits the degree of formant shift as a frequency of π is neared. A similar method may be applied to downward shifting. Adjusting lines in this way alters the frequency relationship between any underlying formants, and therefore will tend to degrade the quality of encoded speech [McLoughlin, 2008].

Appendix **??** presents the fundamental theorem of palindromic polynomials, which forms the basis of the line spectral pair (LSP) representation.

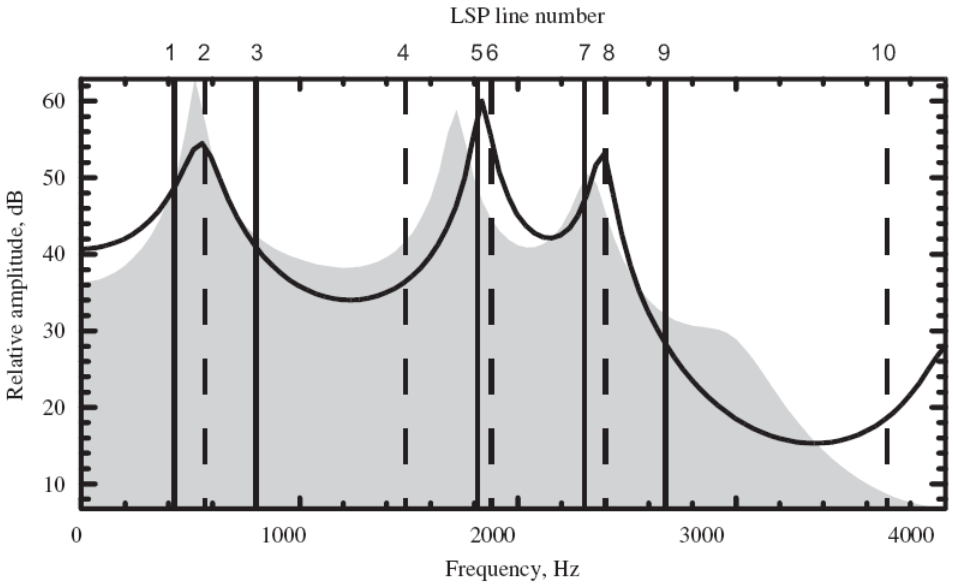

Figure 7.6: The altered set of LSPs, and the resulting LPC spectrum are plotted over the original spectrum, which is shown as a gray area. After McLoughlin [McLoughlin, 2008].

7.7.2 Conversion from Linear Prediction to Cepstral Based Representations

We need to estimate the spectral envelope and manipulate it (interpolation of parameters). For such, we will study the most reliable estimation method and the representation most suited to the problem at hand, interpolation of the spectral envelopes. Since the estimation and manipulation are independent stages of the morphing process, we use both linear prediction and cepstral based representations. There are different possible ways of converting from linear prediction to cepstral representations. I will present recursive and direct analytical relations, followed by an indirect method that uses the power spectrum as intermediate representation between them and that is approximate.

7.7.2.1 Recursive Relations

Recursive relations between cepstrum and predictor coefficients have long been known [Markel and Gray, 1976]. If we let equation (7.92) be an inverse filter polynomial [Markel and Gray, 1976] of order M whose roots are inside the unit circle, the set $\{a_m\}$ are the prediction coefficients. Then, $1/A(z)$ is a stable all-pole filter whose cepstrum coefficients can be expressed as

$$\log\left[\frac{1}{A\left(z\right)}\right] = \sum_{n=1}^{\infty} c_n z^{-n} \tag{7.106}$$

The well-known recursive relation between the a_m and c_n is obtained by differentiating equation (7.106) with respect to z^{-1} and equating equal powers of z^{-1}, yielding [Schroeder, 1980]

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \tag{7.107}$$

The inverse recursive relation is given by

$$a_n = -c_n + \frac{1}{n}\sum_{k=1}^{n-1} -kc_k a_{n-k} \tag{7.108}$$

For some purposes, knowledge of direct relations between these two sets of important parameters characterizing sources and signals is desirable.

7.7.2.2 Direct Relations

A direct (nonrecursive) relation can be obtained by applying a formula [Schröeder, 1999] for the division of two power series to the ratio $-A'(z)/A(z)$ obtained after differentiating the left side of equation 7.106. This gives

$$c_n = \frac{1}{n}(-1)^n \begin{vmatrix} a_1 & 1 & 0 & \cdots 0 \\ 2a_2 & a_1 & 1 & 0 \ldots 0 \\ \vdots & & & \\ na_n & a_{n-1} & \cdots & a_1 \end{vmatrix} \tag{7.109}$$

Unfortunately, this determinant is somewhat unwieldy. An alternative direct form is, therefore, desirable and can be derived as follows. From equations 7.92 and 7.106 we have

$$\ln\left(1 + \sum_{m=1}^{M} a_m z^{-m}\right) = -\sum_{n=1}^{\infty} c_n z^{-n} \tag{7.110}$$

Using the well known power series expansion for $\ln(1+x)$ yields

$$\sum_{k=0}^{\infty} \frac{1}{k}\left(-\sum_{m=1}^{M} a_m z^{-m}\right)^k = -\sum_{n=1}^{\infty} c_n z^{-n} \tag{7.111}$$

or alternatively

$$\sum_{k=0}^{\infty} \frac{1}{k} k! \sum_{n=k}^{\infty} z^{-n} \sum \frac{(-a_1)^{m_1} \cdots (-a_M)^{m_M}}{m_1! \cdots m_M!} = -\sum_{n=1}^{\infty} c_n z^{-n} \tag{7.112}$$

where the third sum has to be taken over all

$$m_1 + 2m_2 + \cdots + Mm_M = n \tag{7.113}$$

and

$$m_1 + m_2 + \cdots + m_M = k \tag{7.114}$$

Because k is summed over all positive integers, the condition 7.114 can be dropped if k in equation 7.112 is replaced by $m_1 + m_2 + \cdots + m_M$.

Equating equal powers of z^{-1} in equation 7.112 then yields the desired *direct* relation between cepstrum and predictor coefficients

$$c_n = \sum \frac{(m_1 + m_2 + \cdots + m_M - 1)!}{m_1! \cdots m_M!} (-a_1)^{m_1} \cdots (-a_M)^{m_M} \tag{7.115}$$

where the sum is to be taken over all m_r that fulfill equation 7.113. But what does the restriction 7.113 on the sum in equation 7.115 mean? In order to understand that, let us assume that $n = 4$

and $M \geq 4$. Then equation 7.113 can be satisfied by the following five choices of m_r, given by table 7.1.

m_1	m_2	m_3	m_4
4	0	0	0
2	1	0	0
1	0	1	0
0	2	0	0
0	0	0	1

Table 7.1: The table lists the values of m.

In addition, all m_r with $r > 4$ must equal zero.

Since m_r is multiplied by r in equation 7.113, we can also say that the different m_r are "counted" r times in adding up to n. In other words, each row of table 7.1 corresponds precisely to one *decomposition* of n into positive integers

number of	1's	2's	3's	4's	
	4	0	0	0	(1+1+1+1=4)
	2	1	0	0	(1+1+2=4)
	1	0	1	0	(1+3=4)
	0	2	0	0	(2+2=4)
	0	0	0	1	(4=4)

Table 7.2: Decomposition of n into positive integers. The table lists the values of m.

Thus, the number of terms in equation 7.115 equals the number of partitions $P(n)$ of n into positive integers not exceeding M. The generating function for $P(n)$ is

$$\prod_{n=1}^{M} (1 - x^n)^{-1} \tag{7.116}$$

a result that can be verified by expanding each term of the product into a geometric series.

The restricted partitions $P(n)$ are related to the unrestricted partitions $p(n)$ by the formula

$$P(n) = p(n) - \sum_{i=0}^{n-M-1} p(i) \tag{7.117}$$

where $p(0)$ is defined to equal 1 and the empty sum is considered to be zero, i.e., for $n \leq M$, $P(n) = p(n)$. Equation 7.117 is obtained by observing that for $m = M + 1$, $P(n) = p(n) - 1$ and by complete induction.

From equations 7.92 and 7.106 we have

$$\sum_{m=0}^{M} a_m z^{-m} = \exp\left[-\sum_{k=1}^{\infty} c_k z^{-k}\right] \tag{7.118}$$

Now we are ready to derive the inverse direct relation, that is, direct computation of predictor coefficients from the cepstrum. Expanding the exponential function into a power series results in

$$\sum_{m=0}^{M} a_m z^{-m} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\sum_{k=1}^{\infty} c_k z^{-k}\right)^n \tag{7.119}$$

Evaluation of the n^{th} power yields

$$\sum_{m=0}^{M} a_m z^{-m} = \sum_{n=0}^{\infty} \sum_{m=n}^{\infty} \sum \frac{(-c_1)^{k_1} \cdots (-c_M)^{k_M}}{k_1! \cdots k_M!} \tag{7.120}$$

where the third sum has to be taken over

$$k_1 + 2k_2 + \cdots + nk_n = m \tag{7.121}$$

and

$$k_1 + k_2 + \cdots + k_n = n \tag{7.122}$$

Because of the sum over n in equation 7.120, the side condition equation 7.122 is obviated. Equating equal powers of z^{-1} gives the desired direct relation for the predictor coefficients in terms of the cepstrum

$$a_n = \sum \frac{(-c_1)^{k_1} \cdots (-c_n)^{k_n}}{k_1! \cdots k_n!} \tag{7.123}$$

where the sum is to be taken over all k_r subject to the condition 7.121.

7.7.2.3 Spectral Power Density Method

The Wiener–Khinchin theorem states that the power spectral density $S_{xx}(\omega)$ of a wide-sense-stationary random process is the Fourier transform of the corresponding autocorrelation function. For the continuous case this gives

$$S_{xx}(\omega) = \mathcal{F}\{r_{xx}(\tau)\} = \int_{-\infty}^{\infty} r_{xx}(\tau) e^{-j2\pi f\tau} d\tau \tag{7.124}$$

where

$$r_{xx}(\tau) = \mathrm{E}[x(t) x^*(t-\tau)] \tag{7.125}$$

is the autocorrelation function defined in terms of statistical expectation, and where $S_{xx}(\omega)$ is the power spectral density of the function $x(t)$. Note that the autocorrelation function is defined in terms of the expected value E of a product, and that the Fourier transform of $x(t)$, does not exist in general, because stationary random functions are not square integrable (power signals). The asterisk denotes complex conjugate, and can be omitted if the random process is real-valued. For the discrete case the formulation is similar

$$S_{xx}(\omega) = \sum_{k=-\infty}^{\infty} r_{xx}[k] e^{-j2\pi kf} \tag{7.126}$$

where

$$r_{xx}[k] = \mathrm{E}\big[x[n] x^*[n-k]\big] \tag{7.127}$$

and where $S_{xx}(f)$ is the power spectral density of the function with discrete values $x(n)$. Being a sampled and discrete-time sequence, the spectral density is periodic in the frequency domain.

Following from the definition of convolution of two signals $x(n)$ and $h(n)$

$$y(n) = \sum_{n=-\infty}^{\infty} x(m)\, y(n-m) = x(n) * h(n) \tag{7.128}$$

we can easily show that for a real signal $x(n)$

$$\mathcal{F}\{r_{xx}(\tau)\} = \mathcal{F}\{x(n) * \bar{x}(-n)\} \stackrel{\circ}{=} X(\omega)\,\bar{X}(-\omega) = |X(\omega)|^2 = S_{xx}(\omega) \tag{7.129}$$

which uses the convolution theorem where indicated by $\stackrel{\circ}{=}$. Equation 7.129 provides a link between the Fourier spectrum of a signal and its autocorrelation, which leads us directly to our main result, the conversion between linear prediction and cepstral coefficients via calculation of the power spectrum. Applying the Levinson-Durbin recursion (7.2.1.8), we can recover the linear prediction coefficients $\{a_n\}$ from the power spectrum. The relation between cepstral coefficients and power spectrum is a direct consequence of the definition of the real cepstrum given by equation 7.75. Since this method is computationally efficient due to utilization of the DFT to calculate the power spectrum and we can use a standard implementation of the Levinson-Durbin recursion, we have chosen this method over the direct relation in equations 7.115 or 7.123, which require complicated conditions on the coefficients. Also, we verified empirically that this method is much more stable than the recursive relation given by equation 7.107, which seems to present convergence problems when truncated.

7.7.3 Quantization Properties

In digital signal processing, quantization is the process of approximating a continuous range of values by a finite set of discrete symbols or integer values. Although the sets of parameters given above provide equivalent information about the linear predictor, their properties under quantization are different. For the purpose of quantization, two desirable properties for a parameter set are:

1. filter stability upon quantization and
2. a natural ordering of the parameters.

Property 1) means that the poles of $H(z)$ continue to be inside the unit circle even after parameter quantization. By natural ordering of the parameters, we mean that the parameters exhibit an inherent ordering, e.g., the predictor coefficients are ordered as $a_1, a_2, \cdots a_p$. If a_1 and a_2 are interchanged then $H(z)$ is no longer the same in general, thus illustrating the existence of an ordering. The poles of $H(z)$, on the other hand, are not naturally ordered since interchanging the values of any two poles does not change the filter. When an ordering is present, a statistical study on the distribution of individual parameters can be used to develop better encoding schemes. Only the poles and the reflection coefficients insure stability upon quantization, while all the sets of parameters except the poles possess a natural ordering. Thus only the reflection coefficients possess both of these properties.

In the experimental investigation of the spectral and cepstral parameters, it was found that the quantization properties of these parameters are generally superior to those of the impulse responses and autocorrelation coefficients. The spectral parameters often yield results comparable to those obtained by quantizing the reflection coefficients. However, for the cases when the spectrum consists of one or more very sharp peaks (narrow bandwidths), the effects of quantizing the spectral coefficients often cause certain regions in the reconstructed spectrum (as described in the previous section) to become negative, which leads to instability of the computed filter.

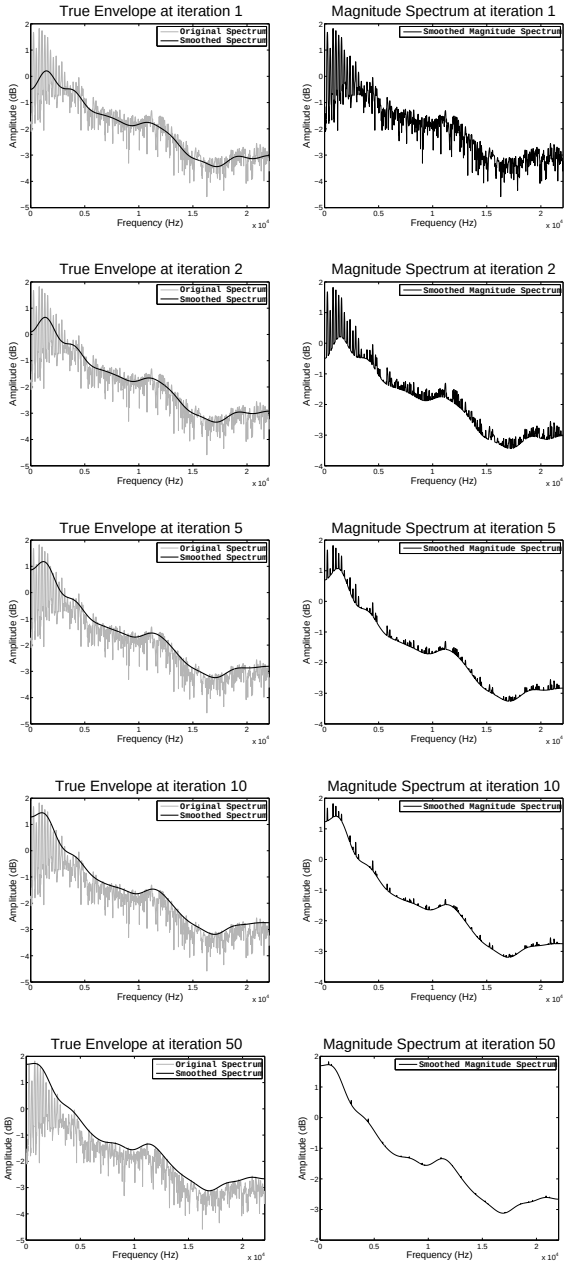

Figure 7.4: "True Envelope" estimation. The figures illustrates the "true envelope" estimation at different iterations i, each corresponding to a row. On the left-hand side we see the original magnitude spectrum $\log|X(k)|$ and the smoothed spectrum $C_i(k)$ at the indicated iteration. On the right-hand side, we see the smoothed magnitude spectrum corresponding to $A_i(k)$ used in the next iteration.

Chapter 8

Temporal Evolution

The aim of this chapter is to present theoretical and technical considerations about the automatic segmentation of musical instrument sounds into perceptually salient temporal segments (or regions). From a theoretical point of view, we need a model of the temporal evolution of musical instrument sounds to guide the segmentation task. In practice, though, only a theoretical model is not enough to obtain robust estimations of the regions defined in the model automatically.

Usually, when we want to automatically detect specific events, we use a detection function that behaves in a particular fashion during the event we want to detect. Thus the automatic detection task becomes simply identifying such behavior and associating it to the event. One example is onset detection, where the event we want to detect is the onset of the sounds (for instance, the beginning of notes played by a musical instrument). There are many possible detection functions, and consequently many different ways of detecting onsets using each type of detection function. The details of onset detection are out of the scope of this work, so I will refer the interested reader to the tutorial review by Bello and colleagues [Bello et al., 2005].

The automatic segmentation task is complex because it requires the detection of several events (onset, end of attack, beginning of release, offset, etc). We simply cannot expect one single detection function (such as the temporal envelope) to contain information about all of the distinct events we want to detect. So we adopted a model that includes spectral information indirectly with the temporal variation of the spectral centroid. The temporal segmentation is done according to the amplitude/centroid trajectory (ACT) model proposed by Hajda [Hajda, 1996] for sustained musical instrument sounds and shown in figure 8.6. Although Hajda proposed a theoretical model to segment sustained musical instrument sounds using spectro-temporal information, there was no explicit recommendation for the automatic identification of the boundaries of the segments. Consequently, I [Caetano and Rodet, 2010a] developed techniques to automatically detect the boundaries of the different events proposed by the ACT model.

The segmentation of musical instrument sounds depends on the correct detection of the boundaries of the regions. Clearly we need a good definition of the regions to be detected in order to be able to estimate them. The first problem we face is that not all instruments contain the same temporal events, so we cannot expect, for example, to be able to estimate the sustain part for a percussive instrument sound. This is where a robust model plays a key role in defining the segments and their boundaries.

The most important aspect to be taken into account is a clear separation of cause and effect. The temporal envelope is merely the description of one of the results of the source-filter interaction. It is fruitless to attempt to detect the boundaries of the events we want to estimate without a proper causal description. We must find the signal level counterparts of the physical events to properly estimate them. The technique we present in this chapter, dubbed ACT segmentation, uses spectro-

temporal cues at the signal level left by the physical/gestural events to correctly segment them.

The difficulty in this approach is that each instrument has its own particularities. Ideally, we search for a model that is robust enough to describe the signal level manifestations of as many types of instruments as possible. We will begin by describing the general model we will consider, namely, the source filter model, and then the physical characteristics of the events we aim to describe. Section 8.4 presents the amplitude/centroid trajectory (ACT) model [Hajda, 1996] for specific classes of instruments. Chapter 12 shows how we use the ACT model to obtain significant estimates of the boundaries from spectro-temporal traces left by the physical gestures.

8.1 Different Regions

Musical instruments are mechanical systems that by themselves are at equilibrium. They need an external source of energy input to produce sound. In general terms, all acoustic musical instruments have one (or more) method for applying mechanical energy to the system, herein termed the excitation method. Pianos have keys connected to hammers that strike a set of tuned strings. Violins have strings that are bowed or plucked. Clarinets have a mouthpiece with a single reed that, when blown, creates a vibrating column of air. Different modes of excitation will generally lead to perceptually different attacks.

The connection between intensity (dynamics), frequency and temporal envelope is far less obvious. Hartmann [Hartmann, 1978] reports on the effects of the amplitude envelope on the pitch of sinusoidal tones. Grey [Grey and Gordon, 1978] and Risset [Risset and Mathews, 1969] have investigated musical instrument sounds independently and concluded that each partial has a particular temporal envelope. They have also discovered that each partial has a slightly different onset, so they termed this phenomenon onset asynchrony. Higher partials tend to start later and this would be an important perceptual cue to group them together into a single percept.

Before the onset, the instrument is in a state of equilibrium. As with all mechanical systems, there is a certain amount of resistance or inertia that keeps the instrument from vibrating on its own. Performers must overcome that inertia before their instrument will sound properly. The more energy a performer uses, the faster the resistance is overcome and the faster the instrument reaches its steady state vibration. For example, different fingerings on a wind instrument produce different lengths of air columns - longer columns mean more mass to vibrate. We know that large masses have more inertia to overcome, but also have more momentum once they are in motion. Thus, low notes have a longer attack and a longer release than high ones.

This section defines the physical/gestural events that generate/define each perceptually different region of the temporal evolution of musical instrument sounds and, more importantly, the signal-level manifestations of the physical events.

8.1.1 Attack

The attack is perhaps the only event that is present in all sounds independent of the mode of excitation. The attack corresponds to the initial excitation of the instrument. The beginning of the attack is perhaps best characterized by the transition between no event and event (or more properly background noise and event for recordings, i.e., signals). This is usually termed onset.

The end of the attack is more difficult to define since it depends on the physical gesture. Notably, transients occur until a permanent resonance mode is attained. For some instruments, we can make a clear distinction between the end of the attack and the beginning of the resonance. The time period when a hammer touches the piano strings would be the attack and the moment the standing wave pattern establishes itself in the string marks the beginning of the resonance mode. For a bowed string it is similar. From the moment when the bow first touches the string

(onset) until the string enters a resonance regime with the bow we can devise two physically and perceptually distinct events. The end of the attack happens before the resonance. For a tube (blown instruments) the situation is similar.

8.1.2 Sustain

The sustain part usually corresponds to the region where the system (musical instrument) is constantly exited with external energy. It is usually defined in terms of approximately constant amplitude. Perceptually, though, it is not reasonable to expect the region where a standing wave vibration pattern manifests as spectrally constant resonances (similar spectral shape) to be sufficiently described solely by the amplitude. Therefore we suggest that constant excitation instruments (bowed and blown, among others), where the energy and the spectral information remain roughly constant, present a sustained part.

8.1.3 Decay

The decay supposedly corresponds to a decrease of energy after the attack during which the permanent excitation regime (such as a standing wave vibration pattern) is already established. This is the region where the amplitude evolution of a percussive instrument sound constantly decays due to losses and strays from constant excitation patterns (blown/bowed strings), where energy is repeatedly input to the instrument during a period of time.

When we look closely, the decay remarkably contains standing wave patterns, even though the amplitude is decreasing. In plucked strings, for example, there is a clear spectral pattern that remains constant throughout and that is perceptually important. The decay contrasts with the amplitude evolution of a blown or bowed instrument, whose standing wave vibration pattern coincides with a more or less constant amplitude.

8.1.4 Release

The release phase admits several interpretations, and its definition has not been consistent among authors. On the one hand, it can refer to the release of the excitation, such that the release segment is the interval between the time instant where the energy ceases to be supplied and the vibrations dying out (offset). This definition is common for sustained sounds, but not always used for non-sustained (percussive) sounds. In the latter case release would be equivalent to decay. On the other hand, it can refer to an intentional interruption of the vibration by the player. Most notably, in stringed keyboard instruments, this corresponds to the release of a key, which causes the damper to stop string vibrations. To avoid confusion between these very different physical events, the following convention will be used. The term release will correspond to the release of the excitation in sustained instruments, while interruption will refer to the case of intentional interruption of vibration in non-sustained instruments.[1]

8.2 The Helmholtz Model

We are looking for a model that allows us to automatically detect perceptually salient temporal events such as attack, decay, sustain, and release of musical instrument sounds. The automatic

[1] Note that, while rare, it is also possible for the release phase of a sustained instrument to be followed by an interruption phase, such as when a violinist intentionally interrupts the vibrations of the strings after having stopped supplying energy to them by bowing.

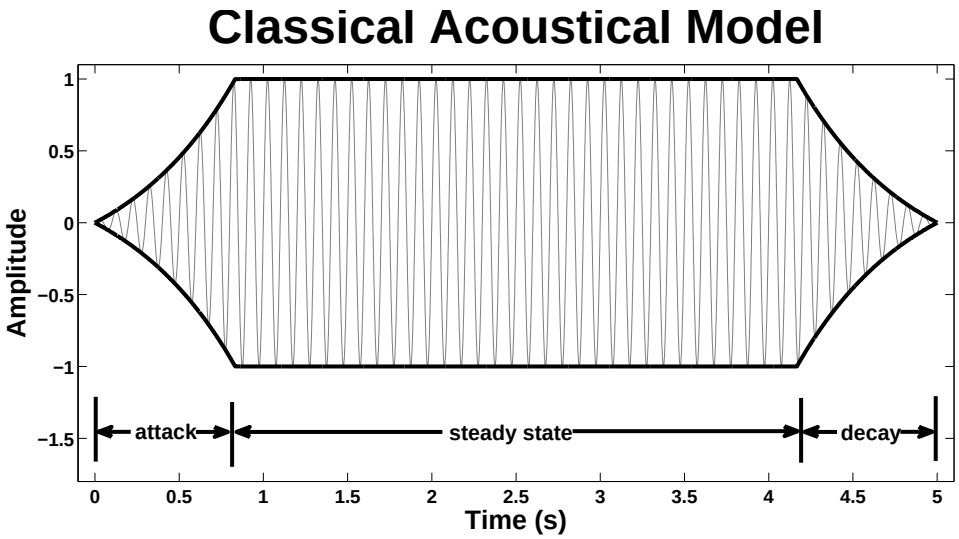

Figure 8.1: The Helmholtz model of the temporal evolution of acoustic musical instrument sounds. Helmholtz defined that the amplitude envelope can be divided into attack, steady state and decay.

detection of these regions usually consists in defining the beginning and end of each event, and here this task is called temporal segmentation.

Usually, these regions have a natural temporal progression (the attack always comes first, for example) and they are successive, such that the boundaries coincide. In chapter 12 the method developed to automatically segment musical instrument sounds and how to use the markers from the automatic segmentation in the temporal alignment part of the morphing process will be described. At this point, we are looking for the signal-level counterparts of the events we want to detect. In other words, we want to determine what we need to measure in the signal in order to detect perceptually relevant events such as attack, sustain and release.

Historically, Helmholtz was the first to propose the segmentation of isolated acoustic musical instrument sounds according to their temporal evolution [Helmholtz, 1885]. Helmholtz characterized what he called musical tone as a waveform that follows an amplitude envelope that consists of the attack, the steady state and the decay, as shown in figure 8.1. During the attack, the amplitude increases from zero to its peak value. In the steady state portion the amplitude is constant and finally decreases back to zero during the decay. Helmholtz concluded that sounds that evoke the sensation of pitch possess fixed waveforms that do not change in the course of the tone, apart from the amplitude envelope, whose temporal evolution has great impact on the perception of the tone, according to him. We should notice that this model only takes into account temporal cues provided by the temporal envelope to define perceptually salient features such as the attack, steady state and decay.

The classic Helmholtz model led to the development of some segmentation techniques that only take temporal cues into account [Jensen, 1999, Peeters, 2004]. Notably, these methods rely on the estimation of the amplitude (or energy, which is amplitude squared) envelope and use it as detection function to estimate the boundaries of the regions defined by the model.

However, the classical Helmholtz model breaks down when we examine musical instrument sounds on a small scale. When the harmonic content of sound is examined with the STFT over small time periods, we discover that, contrary to the Helmholtz model, a sound's spectrum changes profoundly over time. During the attack portion of a sound, harmonic content may change rapidly

and unpredictably. This phenomenon is called the initial transient. During the release, upper partials tend to disappear more quickly before the entire sounds fades away. While the sustain portion of the sound, when it exists, is certainly more stable than the attack or decay, it is hardly as static as Helmholtz would suggest.

Clearly, the basic premise of the classical Helmholtz model - a static spectral envelope with a fixed temporal envelope evolution is by no means an accurate and robust characterization of a wide range of acoustic musical instrument sounds. All these facts suggest that, in order to better understand the temporal evolution of sounds, we need a model that accounts for spectro-temporal changes.

The vast majority of research in sound perception has focused either on the acoustic properties of musical instruments [Risset and Wessel, 1982] or on the perception of sounds as unveiled by psycho-acoustic experiments [McAdams et al., 2005]. The challenge we face today is to find the link between the two in order to be able to manipulate the sounds in a more perceptually meaningful way. A classical example is Risset's discovery that brassy trumpet sounds present a broader spectrum

8.3 The Classical Attack-Decay-Sustain-Release (ADSR) Model

Robert Moog is usually associated with the ADSR envelope model shown in figure 8.2. Moog used the ADSR model in his synthesizer and it quickly became the standard way to describe the amplitude envelope generator functions [Pinch and Trocco, 2002]. However, as early as 1938 (25 years before the first Moog synthesizer), the Hammond Novachord used a 7-position switch to select different ADS (attack-decay-sustain) values, and also had a footpedal to control the release time, which created a sort of pseudo-ADSR envelope controller. It wasn't until Vladimir Ussachevsky, the head of the Columbia-Princeton Electronic Music Center, started working with Bob Moog in 1965, and suggested to Moog that he use an ADSR envelope that it became part of synthesizer history.

Even though the ADSR model was developed to emulate the characteristic time varying amplitude changes of acoustic sounds, it usually does not describe well the temporal evolution of most musical instrument sounds. However, most segmentation techniques [Jensen, 1999, Peeters, 2004] rely on the detection of these events/regions based solely on the use of the temporal envelope. Particularly, the attack is notoriously thought as being dependent on the rise time of the amplitude envelope [Luce and Clark, 1965] and some authors use it as its definition [Bello et al., 2005]. Therefore, we will present some amplitude envelope estimation techniques usually used in the detection of some of these events.

Finally, we present two previously proposed techniques to automatically segment individual musical instrument sounds based solely on the amplitude envelopes, namely derivatives [Jensen, 1999, Skowronek and McKinney, 2006] and efforts [Peeters, 2004]. Both methods try to detect the inflection points of the amplitude envelope based on the assumption that the amplitude envelope changes correspond to the boundaries of the regions we are looking for. Notably, these models define the attack as the rise time of the amplitude envelope, like other authors [Bello et al., 2005, Luce and Clark, 1965].

8.3.1 The Attack-Decay-and-Sustain-Release (AD&SR) Model

Skowronek [Skowronek and McKinney, 2006] proposed a segmentation method based on their attack-decay-and-sustain-release (AD-&-SR) model. They obtain an approximation of the amplitude envelope and use it together with its first derivative to estimate the boundaries of the three

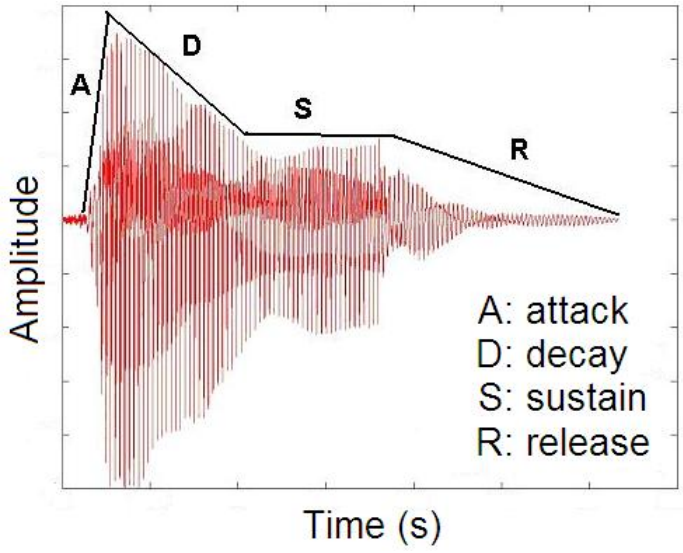

Figure 8.2: ADSR model applied to a wind instrument to explain its temporal evolution.

regions defined as start of attack (soa) and end of attack (eoa); and start of release (sor) and end of release (eor), as exemplified in Figure 8.3.

The first step of determining the desired A-D&S-R approximation is to determine the phases' start and end points. This approach is similar to that proposed by Jensen [Jensen, 1999], which consists of a three stage process illustrated in figure 8.3.

First we compute a heavily smoothed envelope and determine the desired start and end points. Secondly we adjust these points step by step using less and less smoothed versions of the envelope until the unsmoothed version is reached. Jensen's procedure of detecting the time instances from the heavily smoothed envelope has been developed for single harmonic components of instrument sounds. Hence, he uses the temporal envelope of single partials to automatically detect the boundaries.

In this approach, we compute the first derivative of the smoothed envelope and use different derivative thresholds in order to find good candidates for the desired start and end points as follows:

1. The algorithm searches for the steepest point (derivative criterion) having a reasonable value (envelope criterion) and claims this as the middle of attack phase (moa). Starting from this moa point, the algorithm goes backward until certain derivative and envelope criteria are fulfilled and defines this point as start of attack phase (soa). Then starting from moa again, the algorithm goes forward and uses another derivative and envelope criterion for finding the end of attack phase (eoa)

2. The algorithm looks for the start and end points of the release phase (sor, eor) in a similar way, this time starting with the identification of the middle of release (mor) and using negative derivative criteria.

3. Finally the Decay/Sustain phase is defined as the period beginning at the end of attack (eoa) and ending at the start of release (sor).

This gives start and end points of the three phases for the smoothed envelope.

In the second stage the algorithm uses an iterative procedure to perform the adjustment of the found time instances to the unsmoothed case. Step by step a less smoothed version of the envelope

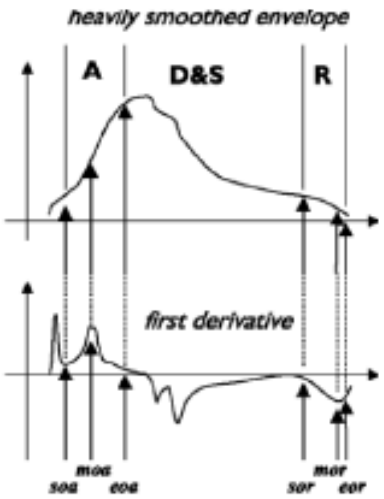

Figure 8.3: Attack Decay and Sustain Release model of the temporal evolution of musical instrument sounds. The figure shows the temporal envelope (top) and its first derivative (bottom), used in the detection of the boundaries of the regions defined by the model. After Skowronek [Skowronek and McKinney, 2006]

is computed and the time instances (soa, eoa, sor, eor) are adjusted using a certain time and level criterion: The new candidate must not be too far away from the former time instance and its new envelope value not too far from the former envelope value.

Once the above mentioned start and end points were found, the three-phase approximation of the signal envelope is adjusted according to the parametric description of the envelope proposed by Jensen [Jensen, 1999]

$$AE(x) = v_0 + (v_1 - v_0)\left(1 - (1-x)^n\right)^{1/n} \tag{8.1}$$

The boundary conditions v_0 and v_1 are the envelope values for the start and end points of the phase. The variable x is the time normalized between zero and one ($t_{start} \rightarrow x = 0$ and $t_{end} \rightarrow x = 1$). The scalar parameter n determines the curve form. If n is equal to 1, then the curve form is linear, if n is smaller than 1, then the curve form has an exponential characteristic; and if n is greater than 1, then the curve form is logarithmic. The optimal curve form parameter n_{opt} is found by minimizing the least-square error between the resulting curve form and the envelope.

In summary the algorithm provides a three-phase parametric description of the envelope with 11 parameters: 4 time instances (soa, eoa, sor, eor), 4 level values (env(soa), env(eoa), env(sor), env(eor)) and 3 curve shape parameters (one for each phase: nA, nD&S, nR).

The use of the derivative as detection function assumes implicitly that the boundaries of the regions we are looking for (soa, eoa, sor, and eor) correspond to the inflection points of the temporal envelope. This is a drawback of this method because naturally sounds do not follow such simplified model. If we consider a sound that presents tremolo, clearly not every modulation in the amplitude of the sound will correspond to spectral changes (even though vibrato most certainly does correspond to spectral changes and is usually accompanied by amplitude modulation as well). This method is particularly sensitive to ripples in the temporal envelope and it depends heavily on the temporal envelope estimation technique. In other words, it is not very robust.

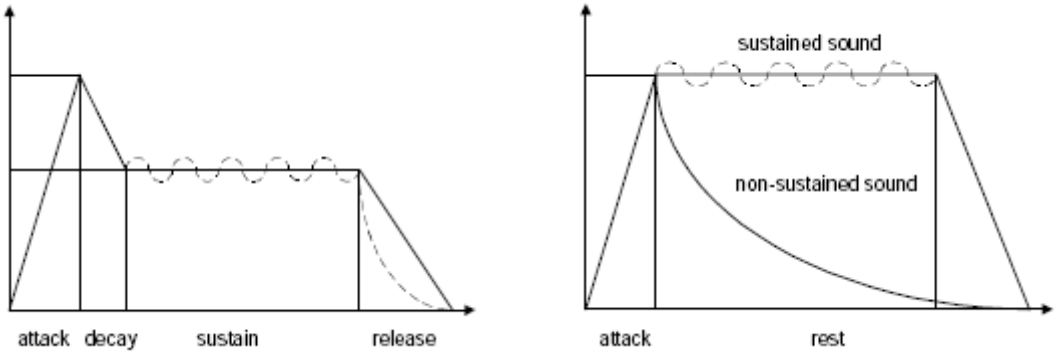

Figure 8.4: Attack Rest model of temporal evolution. After Peeters [Peeters, 2004]

8.3.2 The Attack-Rest (AR) Model for Percussive Sounds

Peeters [Peeters, 2004] adopted two different models of the temporal evolution of musical sounds, corresponding to sustained and percussive excitation, shown respectively on the left and right of figure 8.4. He proposes to use this model to segment the sounds into two regions, the attack and the rest (because the attack is the only region that is present in both sustained and percussive sounds.) The segmentation function he uses is the RMS envelope, and the method detects the beginning and end of the attack, defined as the rise time at the beginning of the sound.

The left-hand side of figure 8.5 shows the fixed threshold method and the right-hand side shows method of efforts, introduced by Peeters [Peeters, 2004]. The fixed threshold method simply defines the beginning of the attack as the point where the RMS envelope reaches 20% of its peak value during the initial rise, and the end of the attack is defined as the point corresponding to 90% of it, following Luce and Clark [Luce and Clark, 1965].

These two criteria are not very realistic because they do not take context such as background noise into account and because not every musical instrument sound follows such simple temporal evolution. Therefore using fixed thresholds does not give very accurate or robust estimations of the start and end of attack. The method of efforts tries to fix some of the problems of the fixed thresholds by determining a way of finding adaptive thresholds tailored for each individual sound. It relies on the division of the initial rise into efforts, as explained in the next paragraph and shown in figure 8.5.

First we divide the slope corresponding to the rise time into N equal intervals according to the amplitude incremental values (called thresholds). The start of attack and end of attack are estimated according to the slope of the rise region. So we calculate a piecewise measure of the slope for each threshold jump by measuring how long it takes to go from one threshold to the next (called efforts). The selected threshold is the one whose value is smaller than M times the mean threshold for both the start and end of the attack. Peeters recommends using $M = 3$. Although this method is more robust than the previously presented, it still yields results that do not accurately capture the attack because it uses strictly temporal information. We will show in the next section why methods that rely solely on the amplitude fail to segment sounds into perceptually meaningful events because they use restricted information.

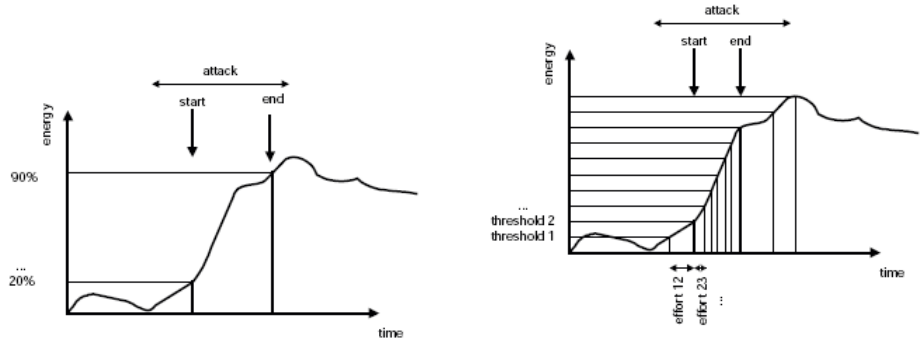

Figure 8.5: Method of efforts. After Peeters [Peeters, 2004]

8.4 The Amplitude/Centroid Trajectory (ACT) Model

Hajda [Hajda, 1996] proposed a segmentation model he dubbed the amplitude/centroid trajectory (ACT) that relies on both the amplitude envelope and temporal evolution of the spectral centroid shown in figure 8.6. The spectral centroid is calculated as follows

$$C(t) = \frac{\sum_{b=1}^{M} f_b(t)\, a_b(t)}{\sum_{b=1}^{M} a_b(t)} \tag{8.2}$$

Here $C(t)$ is the time-varying spectral centroid, $f_b(t)$ is the frequency in Hz and $a_b(t)$ is the amplitude of frequency band b up to the M^{th} band computed. In this model, the spectral centroid gives information about the excitation indirectly. The sudden transition characteristic of the onset reflects as a brief broadening and narrowing of the spectrum, causing the centroid to drop until the steady state resonance establishes itself, bringing the centroid up again to a somewhat steady value. For continuous excitations characteristic of sustained instruments, the release is the moment when the player stops supplying energy to the instrument. This reflects a new drop in the centroid because the higher partials tend to fade before the lower ones, until the sound/note fades away, characterizing the offset. Figure 8.6 depicts the regions (letters) and boundaries (numbers) of the ACT model for sustained sounds. In the figure, BN stands for background noise, A for attack, T for transition, S is sustain, and R is release. The boundaries are the onset (1), end of attack (2), begin of sustain (3), begin of release (4) and offset (5). Using this model, Hajda defines the attack as that part of the signal from onset during which the amplitude increases and the centroid decreases. Pre-attack noise is indicated by more or less uncorrelated fluctuations of both amplitude and centroid. According to the model, the attack ends when the centroid slope changes direction. A new segment, the attack/steady state transition, is defined as that segment immediately following the attack during which the amplitude continues to increase and the centroid increases overall. The sustain begins when the amplitude has achieved a local maximum; during this segment, the amplitude and centroid vary in a more or less monotonic fashion [Hajda, 1996]. The release (or interruption) begins when both the amplitude and centroid decrease.

In this work, the ACT model was chosen to be used in the automatic segmentation task because it outperformed the others. Chapter 12 will explain in detail how to automatically detect the boundaries of the regions defined in the model. Chapter 12 also compares the results of the automatic detection obtained with the ACT method to those of a baseline method, which will be Peeters' AR method.

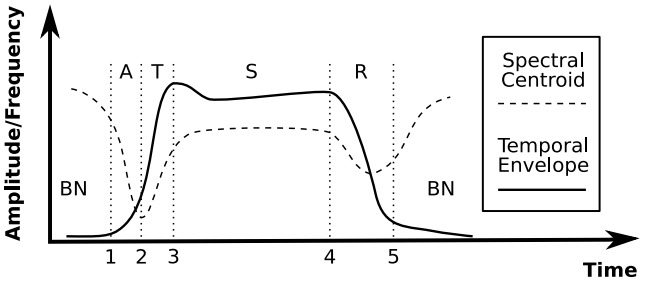

Figure 8.6: The amplitude/centroid trajectory (ACT) model.

Chapter 9

Temporal Envelope Estimation

An audible sine tone may have thousands of cycles per second, depending on its frequency in Hertz, but we perceive it as a steady sound. That is because its amplitude is a constant. By contrast, sinusoidal signals that are heard as fluctuating have amplitudes that change in time. So the ear distinguishes two different time scales that are perceived differently. One scale is the rapid variation perceived as frequency. The other time scale is the slower variation perceived as amplitude modulation and is directly related to the changes in the temporal envelope [Hartmann, 1998].

The temporal envelope is an extremely important factor in the perception of sounds because the fluctuations in amplitude are perceived differently than pitch. Hartmann [Hartmann, 1998] states that there is little difference, if any, between the concept of a time-varying (modulated) amplitude and the temporal (amplitude) envelope. It is important, however, to make a distinction between the temporal envelope and the temporal variation of energy when we consider the Fourier decomposition of signals into a set of time-varying partials.

For each isolated partial (a sinusoidal signal), the temporal variations in amplitude are detected by the ear neglecting phase information [Plomp, 1966]. For a signal resulting from a combination of partials, however, the influence of the phase of each partial on the total temporal evolution of amplitude is not neglectable. Plomp et al. [Plomp and J.M.Steeneken, 1969] investigated if the differences of phase (and therefore global temporal amplitude of the composite signal) are perceptually relevant for "complex tones", that is, signals composed of partials, and concluded that phase information is mostly neglected.

When confronted with partials that present different amplitude modulations, the brain uses the modulations of the total energy that reaches the ear rather than the global time-varying amplitude of the sound. Therefore, for musical instrument sounds (composed of a set of quasi-harmonic time-varying partials), we can distinguish between the temporal amplitude envelope and the temporal energy envelope.

This chapter presents the classic temporal envelope estimation techniques found in the literature, some of which detect the amplitude, others the energy envelope. The amplitude envelope estimation techniques that will be presented are the classical low-pass filtering (LPF), root-mean squared (RMS) energy, and analytic signal amplitude demodulation, as well as frequency-domain linear prediction (FDLP) and the "true amplitude envelope" (TAE) method [Caetano and Rodet, 2011a], developed in the context of this work.

9.1 Early Temporal Envelope Estimation Techniques

An early attempt [Schloss, 1985] consisted of a piece-wise linear approximation of the waveform. The amplitude envelope is created by finding and connecting the peaks of the waveform in a window that moves through the data. Jensen [Jensen, 1999] proposed a method that fits curve shape approximations to model the amplitude envelope of the partials of an additive model of instrument sounds. Later, Skowronek [Skowronek and McKinney, 2006] applied it to approximate the global amplitude envelope.

9.2 Low-Pass Filtering

Low-pass filtering (LPF) is the most straightforward way of obtaining a smooth signal that follows the amplitude evolution of the original waveform. It is based on a classical amplitude demodulation envelope follower technique [Bello et al., 2005], that low-pass filters a half-wave (hwr) or full-wave rectified (fwr) version of an amplitude modulated (AM) signal. The principle of amplitude modulation (AM) is that the amplitude changes of the signal carry the information we seek.

There are many possible filter designs with different characteristics and the choice affects the quality of the final envelope. For instance, Jensen [Jensen, 1999] proposes convolving the waveform with a Gaussian window function, resulting in a suboptimal estimation. When using a finite impulse response (FIR) filter, we should consider the ear's "integration time" because FIR filters shorter than the ear's integration time are perceptually instantaneous [Smith, 2011]. Also, the cut-off frequency of the filter has a major impact on the result. High cut-off frequencies will likely produce an amplitude envelope with ripples. On the other hand, very low cut-off frequencies give temporal envelopes that are smoother but also less responsive to sudden amplitude changes, which can be an issue when estimating the amplitude envelope of percussive sounds.

It is always possible to use low-pass filtering to calculate the energy envelope instead of the amplitude envelope of a signal $x(t)$. We need to low-pass filter the instantaneous energy $x^2(t)$ instead of the signal $x(t)$ and then take the square root of the envelope. The filter lag then becomes the temporal window that can be adjusted to account for the "ear's integration time." We will see that this is equivalent to estimating the root-mean square (RMS) energy envelope.

9.3 Root-Mean Square

The root-mean square (RMS) energy envelope is perhaps the most popular [Tzanetakis and Cook, 2002, Hajda, 1996] method for estimating the temporal evolution of the signal energy. The RMS energy envelope is based on the root mean square energy calculation and can be easily adapted to obtain an estimate of the temporal envelope by simply applying it with a sliding rectangular window, as shown in equation 9.1

$$RMS(t) = \sqrt{\frac{1}{T}\sum_{i=1}^{T} x_i^2(t)} \tag{9.1}$$

where $x_i(t)$ is the i^{th} local sample of the signal centered around t as seen through the window, t is the number of samples the analysis window moves, and T is the window length.

The RMS is a special case of the generalized mean with exponent $p = 2$ and as such, also functions as a sort of moving average, low-pass filter that smooths out the instantaneous energy $x^2(t)$ of the signal $x(t)$. The analysis step t imposes a trade-off between the temporal sampling

rate of the envelope and how much information it represents. Small values of t react sooner to sudden changes in amplitude, while presenting ripple in more steady regions and larger values smooth out the ripples but tend to lag behind abrupt energy changes.

9.4 Analytic Expression

We know from Fourier's theorem that any signal $x(t)$, periodic or not, with a finite number of components can be written as a sum of cosine waves

$$x(t) = \sum_{n=1}^{N} A_n \cos(\omega_n t + \phi_n) \tag{9.2}$$

where the amplitudes A_n are positive real numbers, and there are no restrictions on the frequencies ω_n that are included in the sum. In chapter 18, Hartmann [Hartmann, 1998] derives an analytic expression for the temporal amplitude envelope curve which can be used to understand the limitations of the use of the analytic signal (based on the Hilbert transform) to estimate the temporal amplitude envelope of musical instrument sounds. To find the temporal envelope of $x(t)$, we begin by extracting a sinusoidal character by writing each frequency ω_n as the sum of a characteristic frequency $\bar{\omega}$ plus a deviation δ_n as follows $\omega_n = \bar{\omega} + \delta_n$. Even though for practical purposes of determining the temporal envelope $\bar{\omega}$ can be computed as the average frequency present in the signal's spectrum, the actual value of $\bar{\omega}$ is not important.

We begin by rewriting equation 9.2 as

$$x(t) = \sum_{n=1}^{N} A_n \cos(\delta_n t + \phi_n) \cos\bar{\omega}t - \sum_{n=1}^{N} A_n \sin(\delta_n t + \phi_n) \sin\bar{\omega}t = R(t)\cos\bar{\omega}t - I(t)\sin\bar{\omega}t \tag{9.3}$$

where $R(t) = \sum_{n=1}^{N} A_n \cos(\delta_n t + \phi_n)$ and $I(t) = \sum_{n=1}^{N} A_n \sin(\delta_n t + \phi_n)$.

We should notice that although the division of the time dependence into factors that oscillate at frequency $\bar{\omega}$ and the factors $R(t)$ and $I(t)$ is arbitrary, the point is that we suppose that $R(t)$ and $I(t)$ vary much more slowly than $\bar{\omega}$. Over one cycle of the oscillation at frequency $\bar{\omega}$ the functions $R(t)$ and $I(t)$ can be considered as approximately constant.

Equation 9.3 for $x(t)$ has an explicit oscillation at frequency $\bar{\omega}$, given by a sine term plus a cosine term. This can be written as a single cosine

$$x(t) = E(t)\cos[\bar{\omega}t + \Phi(t)] \tag{9.4}$$

where $E(t)$ is a non-negative time-varying amplitude called the temporal (amplitude) envelope, and $\Phi(t)$ is a phase. The connection between $E(t)$ and functions $R(t)$ and $I(t)$ is given by

$$E(t) = \sqrt{R^2(t) + I(^2t)} = \sqrt{\left[\sum_{n=1}^{N} A_n \cos(\delta_n t + \phi_n)\right]^2 + \left[\sum_{n=1}^{N} A_n \sin(\delta_n t + \phi_n)\right]^2} \tag{9.5}$$

Equation 9.4 shows that any signal can be written as a cosine wave having a frequency of $\bar{\omega}$, so long as it also has a time-dependent phase $\Phi(t)$ and a time-dependent amplitude $E(t)$. At this point we should notice that only because we can always write a signal in the form of equation 9.4, it does not mean that it is always useful to do so. Equation 9.4 explicitly expresses the signal $x(t)$ in terms of an oscillation at a characteristic frequency $\bar{\omega}$, and this representation is useful if the

frequency $\bar{\omega}$ is representative of the signal as a whole. Usually, when the signal consists of a narrow band of frequencies centered at the vicinity of $\bar{\omega}$, this representation is useful. In the case of a narrow band, $E(t)$ and $\Phi(t)$ vary slowly compared to the characteristic frequency $\bar{\omega}$ because they are derived from $R(t)$ and $I(t)$ which are slow varying by construction. This, in turn, is because the range of deviations δ_n is small.

9.4.1 Analytic Signal

The above derivation was done using only real functions. A more elegant way of deriving the same result is by means of the analytic signal $\tilde{x}(t)$. The analytic signal can be obtained by replacing the cosines in equation 9.2 by a complex exponential of the form

$$\tilde{x}(t) = \sum_{n=1}^{N} A_n e^{i(\omega_n t + \phi_n)} \tag{9.6}$$

Unlike the real signal $x(t)$, the analytic signal $\tilde{x}(t)$ is complex. From Euler's theorem, we know that each cosine in the real signal $x(t)$ contains a term $e^{|i(\omega_n t+\phi_n)|}$ and a term $e^{|-i(\omega_n t+\phi_n)|}$, in other words, both positive and negative frequencies. The analytic signal $\tilde{x}(t)$ is what we get when we omit the negative frequency terms and multiply by two. Before we do that in the next section, we will show that the temporal (amplitude) envelope is given by the absolute value of the analytic signal.

In order to show that the temporal (amplitude) envelope is given by the absolute value of the analytic signal, as expressed below in equation 9.7, we need to take a few steps back.

$$E(t) = |\tilde{x}(t)| \tag{9.7}$$

From equation 9.5, we can write $R(t) = E(t)\cos\Phi(t)$ and $I(t) = E(t)\sin\Phi(t)$. Euler's formula, in turn, allows us to simplify it even further by writing

$$R(t) + iI(t) = E(t)\, e^{i\Phi(t)} = \sum_{n=1}^{N} A_n e^{i(\delta_n t+\phi_n)} \tag{9.8}$$

where the right hand side of equation is obtained by simple substitution of the cosine and sine expressions for $R(t)$ and $I(t)$. Then, using equations 9.5 and 9.8we readily see that

$$E(t) = \left| \sum_{n=1}^{N} A_n e^{i(\delta_n t+\phi_n)} \right| \tag{9.9}$$

A clever mathematical trick [Hartmann, 1998] makes the final step possible. We begin by writing $\left|e^{i\bar{\omega}t}\right| = 1$, such that equation 9.9 can be rewritten as

$$E(t) = \left|e^{i\bar{\omega}t}\right| \left| \sum_{n=1}^{N} A_n e^{i(\delta_n t+\phi_n)} \right| = \left| e^{i\bar{\omega}t} \sum_{n=1}^{N} A_n e^{i(\delta_n t+\phi_n)} \right| = \left| \sum_{n=1}^{N} A_n e^{i(\omega_n t+\phi_n)} \right| = |\tilde{x}(t)| \tag{9.10}$$

Even though equation 9.10 can be useful in theory to calculate the temporal (amplitude) envelope of a time-domain signal $\tilde{x}(t)$, it still depends on the decomposition of $\tilde{x}(t)$ in terms of a sum of sinusoids. The Hilbert transform is a practical way of obtaining the analytic signal representation of a real signal $x(t)$ by means of the signal $x(t)$ only.

9.4.2 Hilbert Transform

In the previous section we saw that the analytic signal is the result of taking the Fourier transform of a signal, eliminating the negative frequency terms, and calculating the inverse Fourier transform to go back to the time domain. In this section we will do just that only to discover that this procedure leads to the Hilber transform, part of a signal processing technique for amplitude demodulation [Potamianos and Maragos, 1994].

We begin with the real signal $x(t)$ as the inverse Fourier transform of function $X(\omega)$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} d\omega \tag{9.11}$$

The integral goes over all frequencies, positive and negative. The analytic signal $\tilde{x}(t)$ is the same, except that the negative frequencies are excluded and it is multiplied by two.

$$\tilde{x}(t) = \frac{1}{2\pi} \int_{0}^{\infty} X(\omega) e^{i\omega t} d\omega \tag{9.12}$$

We can rewrite equation 9.12 using the unit step function $U(\omega)$, defined here as

$$U(\omega) = \begin{cases} U(\omega) = 0, & \omega < 0 \\ U(\omega) = 1, & \omega > 0 \\ U(0) = 1/2 \end{cases} \tag{9.13}$$

so equation 9.12 becomes

$$\tilde{x}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} U(\omega) X(\omega) e^{i\omega t} d\omega \tag{9.14}$$

Equation 9.14 can be interpreted as the inverse Fourier transform of the product between $u(\omega)$ and $X(\omega)$, which becomes the convolution of the inverse Fourier transforms in the time domain. We already know that the inverse Fourier transform of $X(\omega)$ is $x(t)$, so we define $u(t)$ as the inverse Fourier transform of $U(\omega)$. Thus,

$$\tilde{x}(t) = 2 \int_{-\infty}^{\infty} x(\tau) u(t-\tau) d\tau \tag{9.15}$$

The inverse Fourier transform of $U(\omega)$ is given by

$$u(t) = \frac{1}{2}\delta(t) + \frac{i}{2\pi}\frac{1}{t} \tag{9.16}$$

such that equation 9.15 becomes

$$\tilde{x}(t) = 2 \int_{-\infty}^{\infty} x(\tau) \left[\frac{1}{2}\delta(t-\tau) + \frac{i}{2\pi}\frac{1}{(t-\tau)}\right] d\tau = x(t) + i\mathcal{H}\{x(t)\} \tag{9.17}$$

where $\mathcal{H}\{x(t)\}$ is the Hilbert transform of $x(t)$, defined as

$$\mathcal{H}\{x(t)\} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{(t-\tau)}d\tau = x(t) * \frac{1}{\pi t} \tag{9.18}$$

where * stands for convolution. Using equation 9.18, we can define the analytic signal $\tilde{x}(t)$ as

$$\tilde{x}(t) = x(t) + j\mathcal{H}\{x(t)\} = E(t)\exp[j\Phi(t)] = \sqrt{x^2(t) + [\mathcal{H}\{x(t)\}]^2}\exp\left[j\arctan\left(\frac{\mathcal{H}\{x(t)\}}{x(t)}\right)\right] \tag{9.19}$$

Equation 9.19 provides us with an expression for the temporal amplitude envelope $E(t)$ that depends only on the time domain, without any reference to frequency. In those cases where the Hilbert transform can be found from the integral in equation 9.18, the temporal amplitude envelope can be found as the absolute value of the analytic signal, as defined in equation 9.19. However, the integral in equation 9.18 converges slowly and there are practical cases when it cannot be used. A more reliable way of computing the Hilbert transform of a signal is by means of equation 9.14, using the inverse Fourier transform to compute the analytic signal $\tilde{x}(t)$.

The analytic signal is useful for envelope detection since its modulus $E(t)$ and time derivative of the phase $\Phi(t)$ can serve as estimates for the amplitude envelope and instantaneous frequency of $x(t)$ under certain conditions. Notably, if the Hilbert transform of $x(t)$ is equal to its quadrature signal [Potamianos and Maragos, 1994], then the estimates are equal to the actual information signals. Synthetic (i.e., AM) signals can be constructed to have this property, but there is no reason to expect that acoustic musical instrument sounds also present it. A more realistic condition is verified when we are dealing with narrowband signals, which is rarely the case for musical instrument sounds. The Hilbert transform can be effectively used to extract the amplitude envelope of individual partials if applied to each frequency bin of the STFT, but when applied to the whole signal it is equivalent to trying to demodulate several AM signals at the same time. However, the absolute value of the analytic signal representation is always positive, so we use it as half-wave rectifier in this work.

9.4.3 Temporal Envelope Power

The temporal envelope power is the long-term average value of $E^2(t)$, defined analogously to the familiar concept of signal power, which is the long-term average value of $x^2(t)$. Because the power in the Hilbert transform of a signal is equal to the power in the signal itself, we see from equation 9.19 that the temporal envelope power is twice the signal following directly from $E^2(t) = x^2(t) + [\mathcal{H}\{x(t)\}]^2 = 2x^2(t)$.

9.5 Frequency-Domain Linear Prediction

Traditional linear prediction [Makhoul, 1975] estimates the spectral envelope from the time-domain signal. The idea behind FDLP [Athineos and Ellis, 2003] is to exploit time-frequency duality to extract the temporal amplitude envelope by applying linear prediction to a spectral representation. In particular, the used spectral representation is the discrete cosine transform (DCT), since it is real-valued. The envelope peaks, whose number and width are determined by the model order, will now be their frequency domain counterparts, the rectified waveform peaks. Thus, the model order has to be adjusted with respect to the temporal structure of the signal, and not to the formant structure of the spectrum.

9.5.1 Discrete Cosine Transform (DCT)

The discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. The distinction between a DCT and a DFT is that a DCT implies different boundary conditions than the DFT. The DCT and its inverse, the IDCT are usually defined as

$$\tilde{X}(k) = \sum_{n=0}^{N-1} x(n) cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], k = 0, ..., N-1 \tag{9.20}$$

$$\tilde{X}(k) = \frac{1}{2}x(0) + \sum_{n=1}^{N-1} x(n) cos\left[\frac{\pi}{N}n\left(k+\frac{1}{2}\right)\right], k = 0, ..., N-1 \tag{9.21}$$

Equation 9.20 is known as the forward discrete cosine transform, and equation 9.21 is known as the inverse discrete cosine transform. The DCT presents the important properties of decorrelation and energy compactation.

9.6 True Amplitude Envelope (TAE)

Ideally, the amplitude envelope should be a curve that outlines the waveform, following its general shape without representing information about the harmonic structure. One of the most challenging aspects of this problem is that we are looking for a curve that is smooth during rather stable regions of the waveform, while being able to react to sudden changes (such as percussive onsets) when they occur. In the context of this work [Caetano and Rodet, 2010a, Caetano and Rodet, 2011a], a temporal envelope estimation technique named true amplitude envelope (TAE) was developed to meet the above mentioned requirements.

TAE uses a dual of "true envelope" estimation, explained in chapter 7, in the time domain. The time domain signal is subjected to the algorithm instead of the Fourier spectrum. In this way, the amplitude envelope is expected to match the amplitude peaks corresponding to the period of the waveform more closely than the previously introduced methods. The idea behind TAE is to mimic the structure of the spectrum with the time-domain signal to be able to apply the true envelope method directly.

The basic steps to estimate the TAE are as follows. First we obtain a rectified version of the waveform (so that that are no negative amplitudes), next we zero-pad the rectified waveform to nearest power of two (thus mimicking the DFT), and then we finally add a time-reversed version of the zero-padded rectified waveform to represent the negative frequencies. Before estimation, we still need to exponentiate the amplitudes because true envelope supposes that we are fitting a smooth curve to the log magnitude spectrum. The result is illustrated in figure 9.1. The last step is the application of the true envelope estimation technique to obtain the true amplitude envelope (TAE), represented as a solid line outlining the rectified waveform.

It is important to notice that the peaks of the waveform do not carry the same information as the spectral peaks. Each peak of the spectrum corresponds to a partial, such that for quasi-harmonic spectra the separation between spectral peaks is given by the value of the fundmental frequency f_0. On the other hand, in only one period, the peaks of the half-wave rectified waveform generally contain information about all the frequencies contained in that signal (depending on their phases). Therefore, the time-domain counterpart of the near optimal order selection must take into account only the period T of the waveform, instead of all rectified peaks. The optimal order is now directly proportional to the fundamental frequency of the waveform, instead of inversely proportional when using true envelope in the spectral domain because the separation of the spectral peaks Δf is now

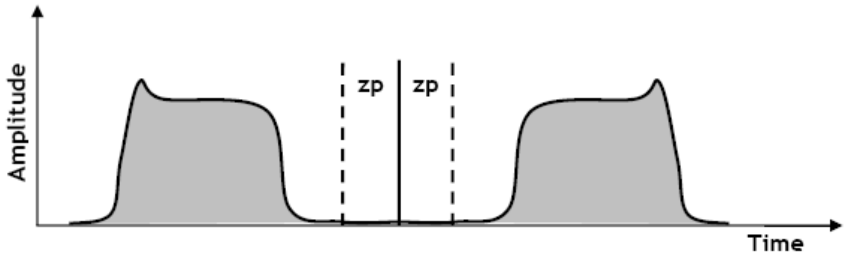

Figure 9.1: Symmetrical waveform used as input for the true amplitude envelope estimation technique. The figure shows the full-wave rectified and zero padded (zp) version of the waveform with its time-reversed counterpart used in the true amplitude envelope estimation method.

represented by ΔT and given by the period of the signal T_0 for a half-wave rectified waveform (hwr) as equation 9.22 shows

$$\hat{O} = \frac{F_s}{2\Delta T} = \alpha \frac{F_s}{T_0}, \alpha = 0.5 \tag{9.22}$$

A full-wave rectified (fwr) version would present twice as many main peaks, requiring half T_0, or $\alpha = 1$. At this point we should remember that the absolute value of the analytic signal $|\tilde{x}(t)|$ gives a half-wave rectified version of signal $x(t)$, while the absolute value of the signal $|x(t)|$ is a full-wave rectified version of $x(t)$.

9.7 Comparison of Amplitude Envelope Estimation Techniques

This section presents a comparison of temporal envelope estimation techniques. The techniques compared are low-pass filtering (LPF), frequency-domain linear prediction (FDLP), root-mean square (RMS) and true amplitude envelope (TAE). Figure 9.2 shows a comparison of the temporal envelope estimation techniques using the rectified version of the waveforms of two different musical instrument sounds.

9.8 Model Conversion

At this point it should be clear that converting between cepstral based and linear prediction based parametric representations of the temporal envelope is strictly equivalent to the conversion operation in the frequency domain. In other words, we can use the same techniques to extract and convert parameters of spectral envelopes for the temporal envelope, because they are similar.

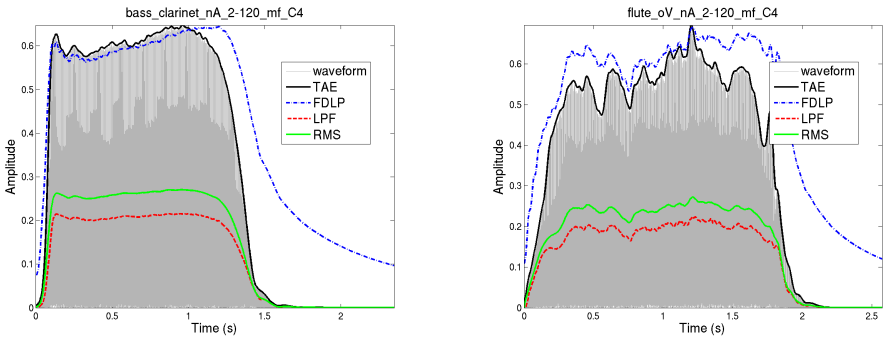

Figure 9.2: Amplitude envelope estimation techniques. The figure shows the waveform and the true amplitude envelope (TAE), frequency domain linear prediction (FDLP), low-pass filtering (LPF), and root-mean square (RMS).

Part III

Morphing Musical Instrument Sounds Guided by Sonic Features

Chapter 10

Vienna Sound Database

This chapter is dedicated to presenting the sound material used in this thesis. The aim of this work is morphing isolated quasi-harmonic acoustic musical instrument sounds across timbre dimensions. Naturally, the focus on musical instruments restricts the possible choice of sounds used in this work. For example, environmental or vocal sounds are out of the scope of the investigation proposed. Moreover, the emphasis on timbre requires a set of musical instrument sounds equalized across other perceptual dimensions, such as pitch, loudness, duration, spatial localization, among others. Finally, quasi-harmonic sounds exclude most percussive instruments.

The type of transformation investigated, cyclostationary morphing, also influenced the choice of sound material. Dynamic morphs would probably demand longer sounds because the transformation happens along the course of the sound. A cyclostationary morph, on the other hand, takes listening to multiple versions of the sound. Thus shorter durations are favored to spare the listener and to avoid memory effects when evaluating the transformation (the listener already forgot the first sound when listening to the last one).

Therefore, a musical instrument sound database that meets such specific needs is necessary. Among the most popular choices are the RWC database (`http://staff.aist.go.jp/m.goto/RWC-MDB/`), Electronic music studios from the university of Iowa (`http://theremin.music.uiowa.edu/`), Ircam solo instruments (`http://www.zikinf.com/news/ircam-solo-instruments-218`), McGill university master samples (`http://www.music.mcgill.ca/resources/mums/html/index.htm`), and Vienna symphonic library (`http://www.vsl.co.at/en/65/71/84/1349.vsl`). These databases were originally conceived with different purposes, such that the instruments, pitch range and available dynamics vary across databases.

All the sounds used in this thesis are from the Vienna symphonic library, which is generally considered very high quality. The samples were played by professional musicians and recorded in controlled conditions to be used in (sample based) synthesizers. The Vienna sound database allows to choose the sounds with differences due mainly to spectral envelope (color as defined by Slawson) and attack times. The sounds are used in all figures presented in chapters 11, 12, 13 and in the listening tests presented in chapter 14.

10.1 Vienna Sound Database

The Vienna sound database contains samples from most musical instruments commonly found in an orchestra recorded under controlled conditions and played by professional instrumentalists. There are woodwind, brass, plucked and bowed string instrument samples in the database covering the normal pitch range of each instrument. The isolated notes are usually played in 7 different

dynamics (pp, p, mp, m, mf, f, ff).

To avoid timbral differences related to pitch and dynamics, the sounds chosen were played with the same pitch and dynamics. The pitches chosen (C3 and C4) were merely the ones for which the most instruments intersect. I used figure 10.1 as a guide. The dynamics chosen was always "forte" for all instruments used in the experiments. For some instruments, notably in the brass family, the timbre changes a great deal when we move up the dynamics scale. Musical instruments in the brass family are commonly described as "brassy" when played "fortissimo", and the characteristics of the sounds change very clearly.

We can find sound recordings with or without vibrato for most instruments in the Vienna sound database. Since vibrato is sometimes considered as a separate timbre dimension, sometimes considered a mere effect, I decided to avoid sound recordings with vibrato whenever possible. In fact, only the strings do not have recordings without vibrato in the Vienna sound database.

One interesting factor to take into consideration is the different attacks for each instrument. Winds have slow, normal and staccato recordings for each pitch and and dynamics. Strings are bowed (normal or staccato) or plucked. Whenever possible, I tried to avoid too many variations, so I selected normal attack. But for some parts of the model, it proved valuable to explore these differences. Notably, in chapter 12 about temporal alignment, I used some of the attack variants to test the robustness of attack time estimation under these changes. The instruments used in the experiments are listed next.

Woodwinds	Brass	Strings
Bass Clarinet	Bass Trombone	Double Bass
Bassoon	Bass Trumpet	Cello
Clarinet Bb	Cimbasso	Viola
English Horn	Contrabass Tuba	Violin
Flute	French Horn	
Oboe	Tenor Trumpet	
	Trumpet C	
	Tuba	

Table 10.1: Vienna sound database. The table lists the instruments used in this work by instrument family.

In the next pages the instrument sounds used in the experiments will be presented by family. The waveform and spectrogram of each instrument sound will be shown.

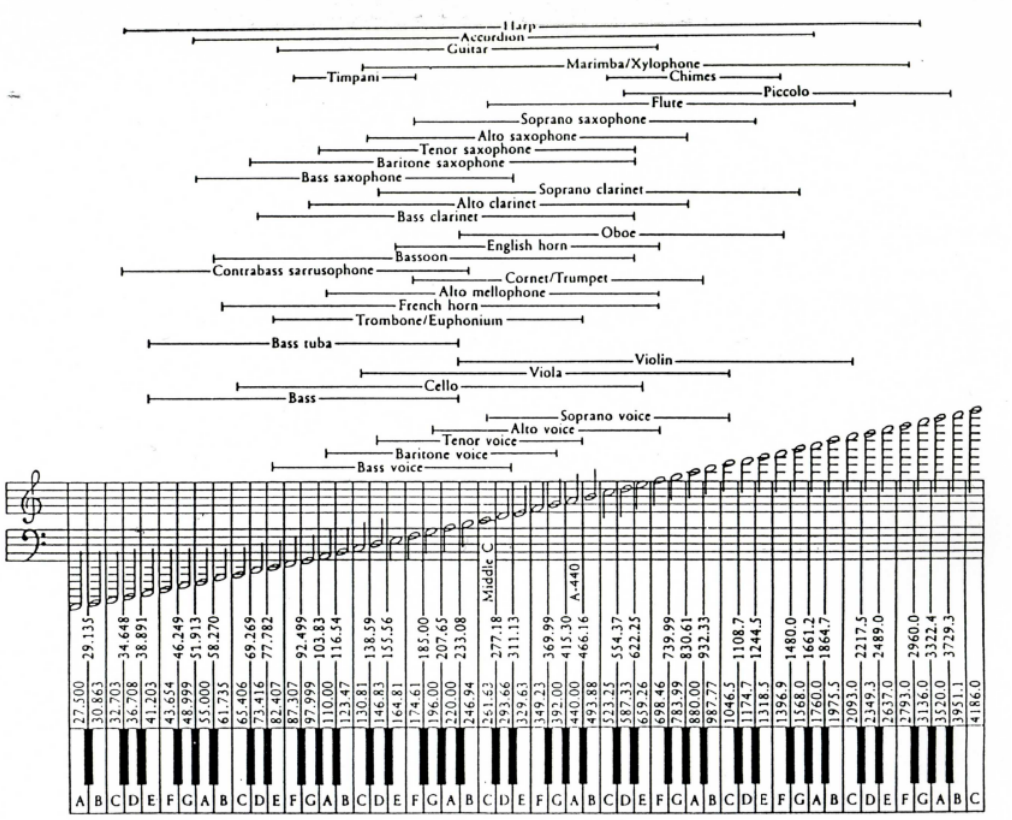

Figure 10.1: Pitch range for the musical instruments normally found in an orchestra.

10.1.1 Woodwinds

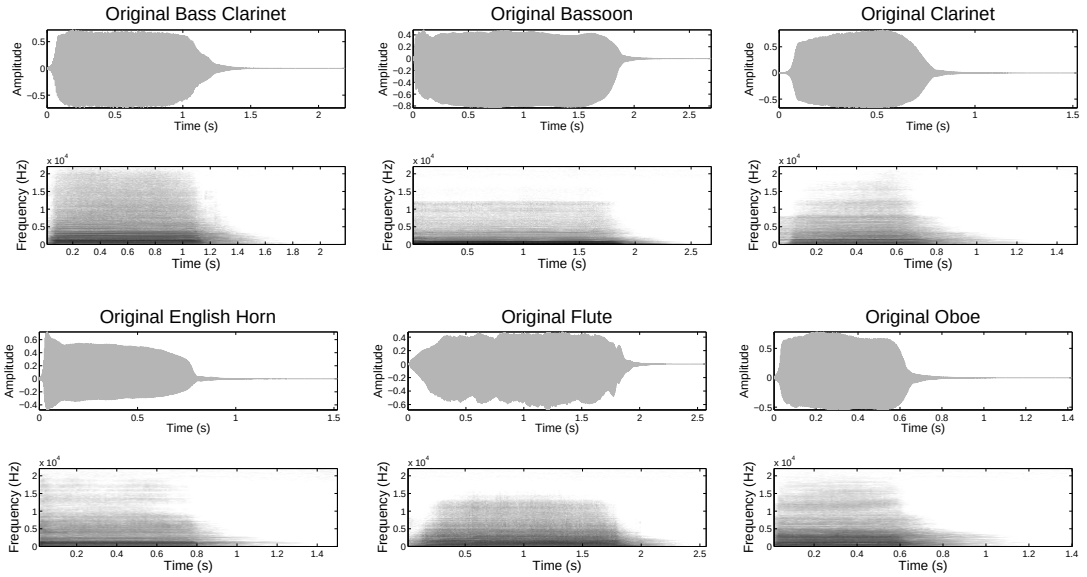

Figure 10.2: Waveform and spectrogram representation of the woodwind instrument sound recordings used in the experiments.

10.1.2 Brass

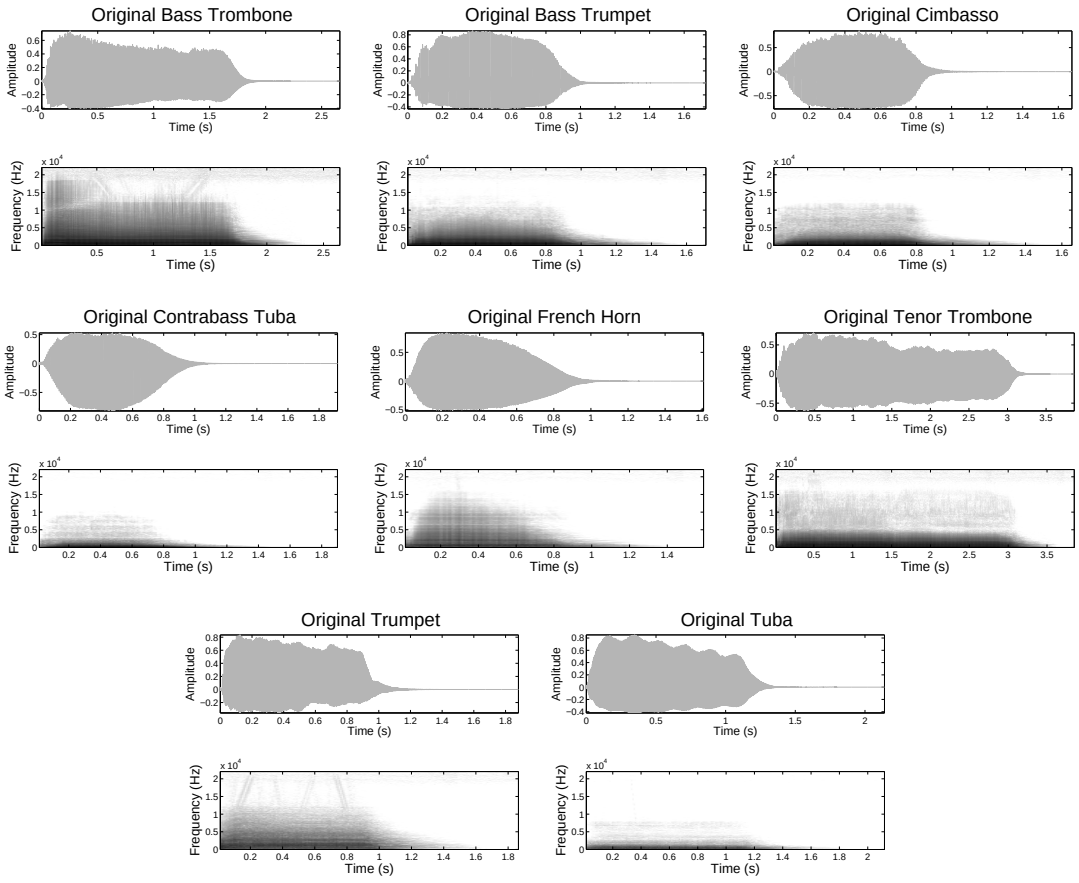

Figure 10.3: Waveform and spectrogram representation of the brass instrument sound recordings used in the experiments.

10.1.3 Strings

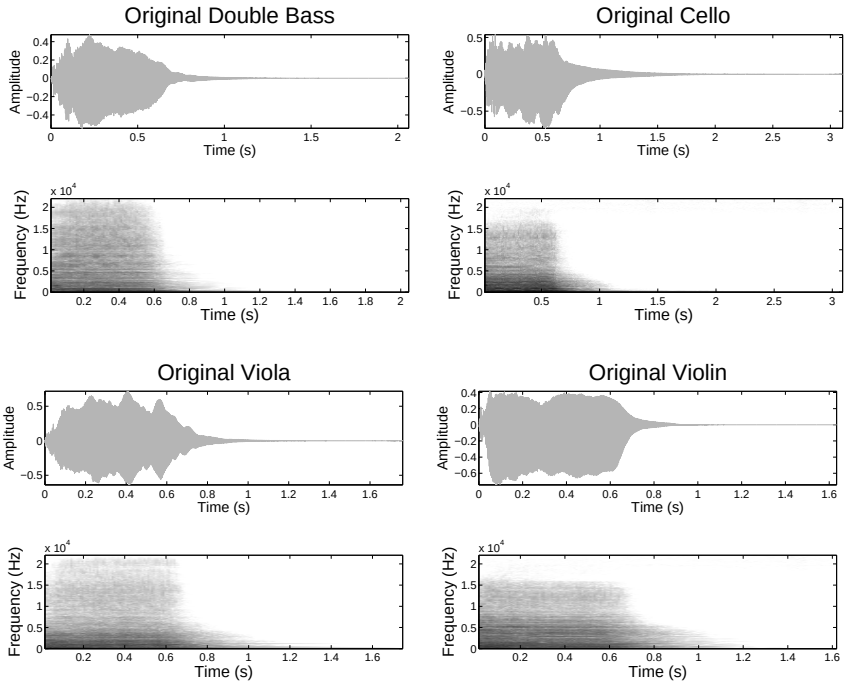

Figure 10.4: Waveform and spectrogram representation of the string instrument sound recordings used in the experiments.

Chapter 11

Overview of the Method

This is a key chapter because, on the one hand, it presents an overview of the sound morphing algorithm proposed in this work step by step; on the other hand, it will show practical results concerning the implementation of the source-filter model used in this work. The source-filter model was presented from a theoretical point of view in the second part of this text, that is, chapters 6 through 9. In this chapter, we will see examples of the estimation of source and filter from a temporal and spectral framework.

First of all, I will present a step by step overview of the sound morphing algorithm developed, as shown in figure 11.1. The algorithm can be subdivided into three parts, temporal processing, spectral processing and morphing procedure. Each part will be presented separately. In the temporal processing part, we will see the results of the automatic temporal segmentation in chapter 12. In the spectral processing part, first some examples of the sinusoidal plus residual decomposition will be shown, then some examples of the spectral (source and filter) modeling. Finally, the morphing procedure is briefly introduced. The morphing procedure coresponds to spectral envelope and temporal envelope morphing. Chapter 12 presents the temporal alignment procedure algorithmically. Chapter 13, in turn, is entirely dedicated to spectral envelope morphing. One very important aspect of the sound morphing procedure when applied to muscal instrument sounds is the spectral morphing, so we will see in chapter 14 the results of the evaluation adopted.

11.1 Sound Morphing Algorithm

Figure 11.1 depicts the general steps applied in the morphing scheme in the actual order in which I apply them. First of all, we should notice that there are three distinct parts, a temporal processing stage, followed by a spectral modeling stage, and finally the morphing procedure is applied. The blocks with a dark background represent sound signals, i.e., waveforms. The temporal processing blocks present a light gray background and represent steps where we act directly on the waveform. Finally, the blocks with a white background represent modeling and morphing of spectral features done on each frame of the source-filter model representation.

The global temporal features log attack time, transition time, sustain (or steady state) time, release time, and temporal centroid are used in the temporal processing stage, as will be explained later in chapter 12. The temporal centroid guides the temporal envelope morphing stage.

On the spectral domain, the spectral shape features guide not only the spectral morphing of each frame, but also their temporal variation with time. The spectral shape features are spectral centroid, spectral spread, spectral skewness, and spectral kurtosis, and their variation when morphing the filter part (spectral envelopes) of each frame of the model will be discussed in detail in

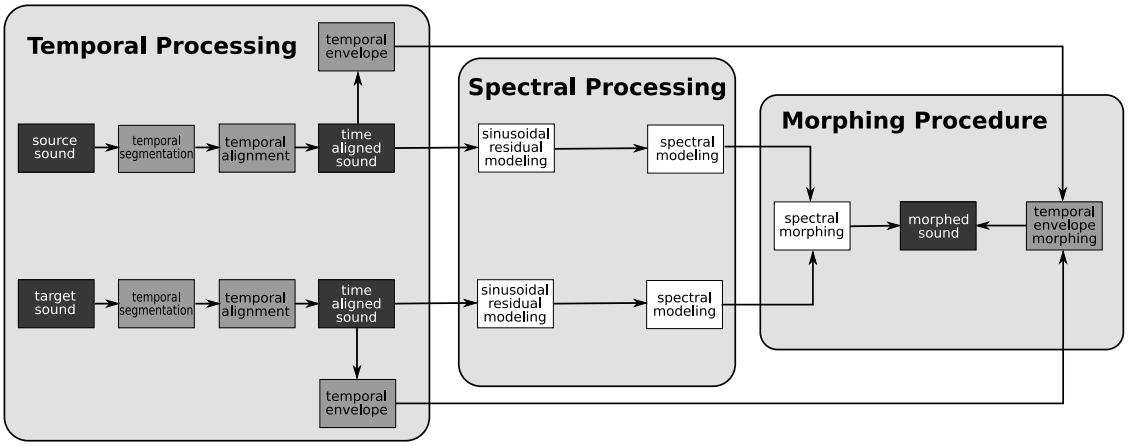

Figure 11.1: Depiction of the general steps of the musical instrument sound morphing algorithm. There are three distinct parts, temporal processing, spectral processing, and morphing procedure. The blocks represent temporal and spectral feature extraction and processing.

chapter 13.

At last, the morphing algorithm uses the morphed temporal envelope to modulate the spectral frames of the sinusoidal and residual components. Chapter 14 briefly discusses which representation of the temporal envelope leads to a smoother morphing. The criteria used and the discussion are similar to their spectral-level counterparts.

11.2 Temporal Processing

The temporal processing part consists of the temporal segmentation step, followed by temporal alignment. The temporal alignment step can be considered as a global temporal morphing procedure or pre-processing for the spectral modeling step. On the one hand, the global temporal morphing view derives from the intermediate duration of the regions after temporal alignment. The global temporal features of the sounds being morphed should also be perceptually intermediate. On the other hand, time-aligning the sounds helps guarantee spectral smoothness in the morphed sound. That is, it avoids the combination of attack transients with steady sustain frequency information, etc.

11.2.1 Temporal Segmentation

The results of the temporal segmentation are crucial to the rest of the morphing procedure because the temporal alignment uses the results from the temporal segmentation step. The temporal alignment should guarantee not only correspondence between the number of spectral frames, but also make sure that attack frames are matched with attack frames, etc. The results of the automatic temporal segmentation method developed in this work will be presented in chapter 12, together with the temporal alignment procedure.

11.2.2 Temporal Alignment

The temporal alignment procedure can be considered mere pre-processing or an actual part of the morphing procedure (strictly in the time domain). An important consequence of the tempo-

ral alignment procedure is that each segment ends up having an intermediate duration. This is especially important for the attack time, which is a very important perceptual cue for musical instrument sounds. Even though the temporal centroid has been shown to be correlated with the concept of percussiveness, the log attack time usually explains better the first dimension of timbre spaces obtained with MDS [Caclin et al., 2005, McAdams et al., 2005]. However, we will not neglect the relative importance of the temporal centroid. The temporal morphing procedure, which relies on the temporal envelope estimation, uses the value of the temporal centroid to guide the results.

11.2.3 Temporal Envelope Estimation

Even though the temporal envelope has already been estimated before in other parts of the algorithm, it is the result of the estimation of the temporal envelope of the time-aligned sounds that is morphed. The temporal envelopes will be morphed and used to shape the temporal evolution of the morphed sounds, by modulating the morphed spectral frames. To account for the temporal variation of the energy, the temporal envelope is estimated using RMS [Caetano and Rodet, 2010a].

11.3 Spectral Processing

After the temporal processing part, both sounds have not only the same number of frames (one-by-one correspondence), but also the corresponding regions (attack, transient, sustain, and release) have intermediate lengths and are properly aligned. The spectral modeling part is represented in figure 11.2. The first step is to separate the time-aligned sound into a harmonic sinusoidal and noise residual using a sinusoidal model. The sinusoidal component is modeled as a set of partials for each frame. Each partial is modeled as a sinusoid whose amplitude values vary slowly when compared to its frequency, such that each partial is described by the partial frequency, its corresponding amplitude and phase values. The noise residual is simply the result of the subtraction of the sinusoidal part from the original signal.

In the section about the spectral modeling step, we will see a comparison between the spectral representation of the traditional sinusoidal model with the implementation of the source-filter model adopted in this thesis. We should consider two important aspects in the spectral modeling part, accuracy of representation and ease of manipulation. Ideally, the model should represent the original sound accurately and allow independent and coherent manipulation of different parts of the model. One way to test the accuracy of the representation is to resynthesize a sound from the parameters of the model representation and compare it with the original. An accurate model should lead to a sound that is perceptually identical to the original sound (or at least close enough depending on the intended application). On the other hand, the ease of manipulation is essential when performing sound transformations. If a representation has too many independent parameters, it becomes cumbersome to manipulate all of them coherently to obtain a certain result. Perceptually speaking, a coherent manipulation of the amplitude values of a spectral representation would change the values of the amplitudes in such a fashion that the balance of the distribution of spectral energy is changed, rather than the amplitudes of isolated partials independently from the energetic context; i.e. the amplitudes of the partials around it.

One final aspect to take into consideration is the spectro-temporal variations, also called spectral flux or fluctuations [Caclin et al., 2005, Grey and Gordon, 1977, McAdams et al., 2005, Krumhansl, 1989]. Time is such an inherent dimension in the perception of sounds that Smalley states that "spectrum is perceived through time, and time is perceived as spectral motion [Smalley, 1986]." Like mentioned earlier in the text, a sound and its time-reversed version are rarely perceived as identical (or even similar), even though the spectral contents are exactly the

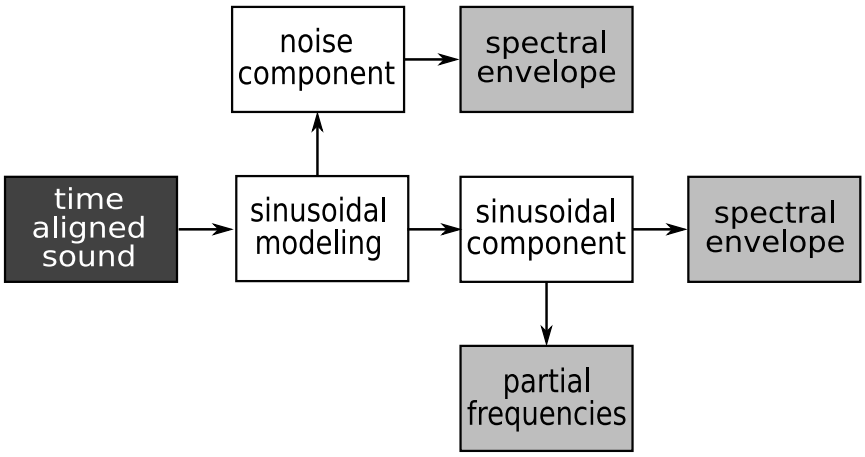

Figure 11.2: Spectral modeling. The time-aligned sound is represented with a dark background, the sinusoidal modeling blocks have a white background, and the blocks where features are extracted have a light gray background.

same. In this chapter, we will see the temporal representation of the source-filter model and the temporal variation of the spectral shape features for the sounds of acoustic musical instruments.

11.3.1 Sinusoidal plus Residual Decomposition

For musical instrument sounds, the sinusoidal component contains most of the acoustic energy present in the signal because musical instruments are designed to have very steady and clear modes of vibration. To exemplify, this section will show the result of the sinusoidal plus residual decomposition for musical instrument sounds of each family considered in this work, namely, woodwinds, brass and strings. Figure 11.3 illustrates the sinusoidal plus residual decomposition for musical instrument sounds. In figure 11.3, we see the waveform at the top and the spectrogram at the bottom of the original sound on the left-hand side, the sinusoidal representation in the middle, and the noise residual on the right-hand side. Figure 11.3 shows that most of the acoustic energy corresponds to the sinusoidal component for musical instrument sounds, as stated earlier.

11.3.2 Spectral Modeling

This section discusses the technical aspects of the model presented in chapter 6. The aim of this section is to specify what technique we used to estimate each part of the model and justify why whenever necessary. The spectral modeling step comprises modeling the sinusoidal and the residual components independently. In this work, the source-filter model is used for both the sinusoidal and the residual components. The source is represented by sinusoidal partials for the sinusoidal component and white noise for the noise residual. The filter is represented by a spectral envelope curve for both sinusoidal and residual components. The implementation of the source-filter model will be presented separately for the sinusoidal and residual components.

11.3.2.1 Source-Filter Modeling of the Sinusoidal Component

Like explained earlier in chapter 4, the result of the sinusoidal analysis for each spectral frame is a set of values that describe the partials contained in that frame. Each partial has a partial number,

amplitude, frequency and phase value associated. For example, one frame where N partials are detected would contain the following.

partial number	amplitude	frequency	phase
1	a_1	f_1	ϕ_1
2	a_2	f_2	ϕ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	a_N	f_N	ϕ_N

(11.1)

where the first column contains the partial number, a_n is the amplitude of the n^{th} partial, f_n the frequency value, and ϕ_n the phase. The source-filter representation replaces the amplitude values estimated with the sinusoidal model with the values given by the spectral envelope curve. Thus the next section presents a comparison of the representation of the amplitudes of partials between the sinusoidal and the source-filter model.

Like explained in section 6.2, the source-filter model can be used to represent the result of sinusoidal analysis. The amplitudes of the sinusoids are represented with a spectral envelope model, while the frequencies of the partials are considered the discrete frequencies at which we sample the amplitude information contained in the spectral envelope curve. The sound model developed in this thesis stores the frequencies of the partials and uses them later to obtain the amplitudes of the partials from the morphed spectral envelope. We retrieve the amplitudes of the partials from their frequency values using the property that sinusoids are the eigenfunctions of linear shift-invariant systems. Therefore, the filter must be converted to an LSI representation (e.g., LPC) before this operation.

11.3.2.2 Spectral Envelope Estimation

In order to obtain the best possible estimation of the spectral envelope, we chose to use true envelope [Röbel and Rodet, 2005] for the extraction of the spectral envelope of every frame of the source-filter model. The result of true envelope estimation is a set of cepstral coefficients representing the estimated spectral envelope curve. Next, this cepstral based representation is converted to a linear prediction based representation using the spectral power density method explained in chapter 7. The conversion needs to be done upon resynthesis to retrieve the amplitudes of the partials from the LPC representation of the filter.

The spectral power density method is used because it outperformed the other methods empirically. For speech, there are studies on the quality of the conversion [Villavicencio et al., 2006], so the same method was adopted for musical instrument sounds. But first, let us investigate how accurate the spectral envelope estimation is.

Figure 11.4 shows the original spectrum and the corresponding partials (spectral peaks selected by the peak-picking algorithm of the sinusoidal analysis). At the bottom, we see the spectral envelope curve (estimated with "true envelope") representing the amplitude of the partials, and the partials represent the frequency values at which we sample the spectral envelope curve. We should notice that the amplitudes of the partials shown at the bottom of figure 11.4 as vertical lines are from the sinusoidal analysis. That is, these are the values estimated from the original spectrum, so we can compare them with the equivalent amplitudes given by the spectral envelope curve.

At this point it should be clear that both representations retain essentially the same information (amplitude and frequency of partials peaks) in different ways. The spectral envelope curve is only an interpolation of the amplitudes of the partials using a cepstral model. The sinusoidal representation has a more accurate representation of the amplitudes of the partials, while the source-filter model representation presents small in the values of the amplitudes inherited from the spectral envelope

estimation procedure. On the other hand, the source-filter representation is very flexible when we want to transform source and filter independently.

The reader should remember that the spectral shape features are a measure of the balance of the distribution of spectral energy, and this is one of the reasons why they are correlated to the way we hear sounds. In other words, even though the ear (seen as a spectral analyzer) only measures the distribution of spectral energy (what we hear), the brain interprets this information in terms of balance of this distribution (how we hear). For example, when we use the sinusoidal representation to perform spectral transformations, we can change the amplitude of only one of the partials, while changing the value of only one parameter of the spectral envelope model used to represent the filter usually has a more distributed impact on the amplitudes of the partials. The particular impact of changing the parameters of different spectral envelope representations depends on the nature (cepstral, linear prediction, etc) of the spectral envelope model and what information each parametric representation encodes. Chapter 13 is entirely dedicated to the problem of interpolation of the parameters of different spectral envelope representations.

The sinusoidal representation is accurate for the amplitudes of the partials, but it is not very intuitive to manipulate coherently because there are too many degrees of freedom. The source-filter model represents the amplitudes of the partials less accurately because the estimation of the spectral envelope curve presents errors, but it is more intuitive to manipulate coherently.

11.3.2.3 Partials Frequency Values

The values of the partials are essential in the resynthesis step because they represent the frequencies at which we will "sample" the morphed spectral envelope that represents the filter part of the morphed sound. The filter is estimated with a spectral envelope technique (in the cepstral domain) and then converted to linear prediction coefficients (LPC). The LPC representation is necessary upon resynthesis because the sinusoids used to represent the partials are the eigenfunctions of linear shift-invariant systems. In other words, we can simply "sample" the LPC representation of the filter part of the source-filter model and we obtain the corresponding amplitudes. In mathematical terms, the sinusoidal component $s_s(t)$ is expressed as a sum of sinusoids

$$s_s(t) = \sum_{k=0}^{K(t)} A_k(t) \sin(2\pi f_k t + \psi_k) \tag{11.2}$$

where A_k is the amplitude of the k^{th} partial, and f_k its frequency in Hertz. Then, if we express the LPC representation of the filter by $H_s(\omega)$, we can obtain the amplitudes of the partials A_k from their frequency values f_k and $H_s(\omega)$ as follows

$$\sum_{k=0}^{K(t)} A_k(t) = H_s\left(\sum_{k=0}^{K(t)} 2\pi f_k t + \psi_k\right) \tag{11.3}$$

Thus the signal $s_s(t)$ can be represented as

$$s_s(t) = \sum_{k=0}^{K(t)} s_k(t) H_s(2\pi f_k t + \psi_k) \tag{11.4}$$

where $s_k(t) = \sin(2\pi f_k t + \psi_k)$ is a slowly varying sinusoid with frequency f_k in Hertz.

11.3.2.4 Source-Filter Modeling of the Noise Residual

As explained earlier in chapter 4, the noise residual is obtained by subtraction of the synthetic signal from the original sound. This procedure of subtracting the waveform in the time domain is valid because the sinusoidal component preserves the waveform by estimating the phases of each partial. The filter part of the noise residual is modeled with a spectral envelope like for the sinusoidal component. The only difference between them is the representation of the source. In the sinusoidal component, the source is supposed sinusoidal, like in equation 11.2, and in the noise residual component the source is represented by white noise.

This model supposes that there is information present in the spectrum of a quasi-harmonic musical instrument sound that is not well represented with sinusoids. The same can be said for speech [Stylianou, 2008]. The representation of the noise residual is simple, we suppose that the noise residual can also be explained by the source-filter model, but this time the source is white noise instead of sinusoidal tracks. So the noise residual is modeled as white noise filtered by the spectral envelope estimated from the noise residual signal. For the residual component, this work uses linear prediction to estimate the spectral envelope curve because we are looking for a curve that fits the statistical properties of the noise residual, rather than the peaks of the spectrum [Makhoul, 1975].

11.3.2.5 Spectral Shape Descriptors

Now we are ready for the spectro-temporal view using the source-filter representation. Figure 11.5 shows the sinusoidal model and the source-filter model representations side by side to allow us to compare them. The source-filter representation shows the temporal variation of the frequency of the partials (representing the source for the sinusoidal component) at the top and the temporal variation of the spectral envelope envelope (the filter) at the bottom.

The spectral envelope representation shows the same information as the spectrogram (higher amplitudes are darker). One important difference is the scale of the frequency information represented. While the spectrogram shows the spectrum for each frame (with information about the fundamental frequency), the spectro-temporal view of the filter shows the spectral envelope curves (only distribution of spectral energy).

The spectro-temporal view of the source-filter model allows us to see the temporal variation of the spectral shape features for each musical instrument. Figure 11.6 shows the waveform and the value of each spectral shape descriptor calculated on each frame of the spectral representation. The spectral shape features are calculated using the perceptually related model explained in chapter 5.

The first thing we should notice about figure 11.6 is how the spectral shape features are correlated to one another. The spectral centroid and spread are positively correlated, and so are the spectral skewness and kurtosis. We should also notice that the lower order moments (spectral centroid and spread) are negatively correlated with the higher order moments (skewness and kurtosis). The interpretation of the spectral shape features from a perceptually related point of view as measures of the balance of the distribution of spectral energy now gives us insight into the spectro-temporal evolution of musical instrument sounds. Interestingly, the segmentation model adopted for the automatic segmentation of musical instrument sounds (explained in chapter 12) uses the information conveyed by the spectral shape features (specifically the centroid) to aid in detecting the segments.

11.4 Morphing Procedure

After the temporal and spectral modeling steps, we are finally ready for the morphing procedure, which comprises the interpolation of the spectral representation of each frame and the temporal

envelope morphing. When we use the source-filter model, we interpolate the parameters of the representation of the source (frequencies of the partials) and of the filter (parameters of the spectral envelope model). Each one of these procedures will be explained separately below.

At last, the frames of the morphed sound are finally modulated by the morphed temporal envelope. The temporal envelope morphing procedure is guided by the value of the temporal centroid, the temporal counterpart of the spectral centroid.

11.4.1 Spectral Morphing

The spectral morphing process consists of two steps, spectral envelope morphing and interpolation of the values of the frequencies of the partials frame by frame. Each of these will be considered separately next.

11.4.1.1 Spectral Envelope Morphing

The reader must remember that the spectral envelope is associated with musical instrument identification. So morphing the spectral envelope is essential when we want to obtain sounds that would correspond to hybrid musical instruments. Spectral envelope morphing is such a perceptually important step in morphing musical instrument sounds that it will be considered separately in chapter 13, entirely dedicated to the subject. For now, suffice it to say that the balance of the distribution of spectral energy is a key aspect of the spectral envelope morphing procedure. The morphed spectral envelope should have an intermediate balance of distribution of spectral energy to be perceived as intermediate with respect to the spectral shape. The spectral shape features adopted in this work as guides for the spectral envelope morphing procedure are measures of the balance of distribution of spectral energy. Chapter 13 explores in depth the several intricacies of spectral envelope morphing.

11.4.1.2 Interpolation of Partials Frequency Values

For musical instrument sounds, expressivity has an impact on the source part of the source-filter model. For example, vibrato usually leads to a modulation in the frequency values of the partials. For this reason, it is very important to interpolate the frequencies of the partials too. If one of the sounds to be morphed presents vibrato, but the other one does not, we want this particular feature to gradually change when morphing from one sound to the other in a cyclostationary fashion (like explained in chapter 3).

The amplitudes of the partials are represented by the spectral envelope, such that the interpolated frequency values only represent the frequency values at which we will "sample" the response of the morphed spectral envelope, which models the filter of the hybrid musical instrument

The partial number is used to make sure that the principle of correspondence will hold. In other words, we interpolate partials that have the same partial number. The interpolation of the values of the frequencies of the partials is based on the interval between the frequency values measured in cents. An interval ς between frequency f_{11} and frequency f_{12} can be expressed in cents the following way

$$\varsigma = 1200 \log_2 \left(\frac{f_{n1}}{f_{n2}} \right) \tag{11.5}$$

where f_{n1} represents the frequency value of the n^{th} partial of the first sound, and f_{n2} the frequency value of the n^{th} partial of the second sound. Then we interpolate the interval ς in cents rather than the frequency values f_{n1} and f_{n2} directly. First we define the frequency f_α with the aid of equation 11.5 as a fraction α of the interval ς in cents. That is,

$$f_{n\alpha} = f_{n1}2^{\frac{\alpha\varsigma}{1200}} = f_{n1}2^{\alpha \log_2\left(\frac{f_{n2}}{f_{n1}}\right)} \tag{11.6}$$

We use the value of $f_{n\alpha}$ as the n^{th} partial interpolated frequency. Naturally, a frequency $f_{n[1-\alpha]}$ can be analogously defined for the fraction $[1-\alpha]$, but we should notice that it gives the same value as $f_{n\alpha}$ when both f_{n1} and f_{n2} were detected. That is, when both sounds present the n^{th} partial for the frame being morphed.

This brings us to an important point when morphing the values of the frequencies of the partials, namely, correspondence. When we strictly follow the correspondence principle adopted in chapter 2, we can only interpolate partials that have a matching partial number. That is, if the first sound has N_1 partials and the second N_2, we only interpolate until the N_1^{th} partial number when $N_1 < N_2$, and until the N_2^{th} otherwise.

Nevertheless, we can use a clever trick to "interpolate" between frequencies of partials that were not detected for sounds whose spectrum fits well the quasi-harmonicity model. In other words, when $N_1 < N_2$ and the partial number $n > N_1$, we can replace the frequency value f_{n1} by a harmonic estimate based on the fundamental frequency f_{11} and the harmonic number n as $f_{n1} \simeq nf_{11}$.

However, we should notice that this substitution can only be used when both sounds are quasi-harmonic because it implicitly supposes so. The substitution does not work well when one of the sounds is slightly inharmonic, that is, when the values of the frequencies of the partials of one of the sounds being interpolated present a slight harmonic deviation (such as piano sounds, for example). When we have an ihnarmonic set of values of partial frequencies we must only interpolate the intervals in cents between pair of partials that were detected. Alternatively, we can use a model of the inharmonicity to predict the frequencies of upper partials that were not detected and therefore do not have a match.

11.4.1.3 Phase Values

The phase values are not interpolated because the phase of the morphed sound does not necessarily correspond to the interpolation between the phase values of source and target sounds being morphed. In fact, the phase values are simply discarded when morphing. The phase of the morphed sound is simply reconstructed integrating the values of the frequencies of the partials at the synthesis stage using equation 4.2. In this work, the difference between the integrated phase value and phase values calculated as an interpolation of the estimated phases is neglected.

11.4.2 Temporal Envelope Morphing

Finally, the temporal envelope estimated from the time aligned sinusoidal and residual components is morphed and used to modulate the frames of the morphed sound upon resynthesis. The estimation of the temporal envelope is done using RMS. The temporal envelope curve obtained can be converted to its cepstral or even linear prediction representation and morphed using the same techniques for spectral envelope morphing.

In this work, temporal envelope morphing is guided by the value of the temporal centroid, which is correlated with the "percussiveness" of (musical instrument) sounds [Skowronek and McKinney, 2006]. The results of the temporal envelope morphing procedure will be shown and evaluated in chapter 14.

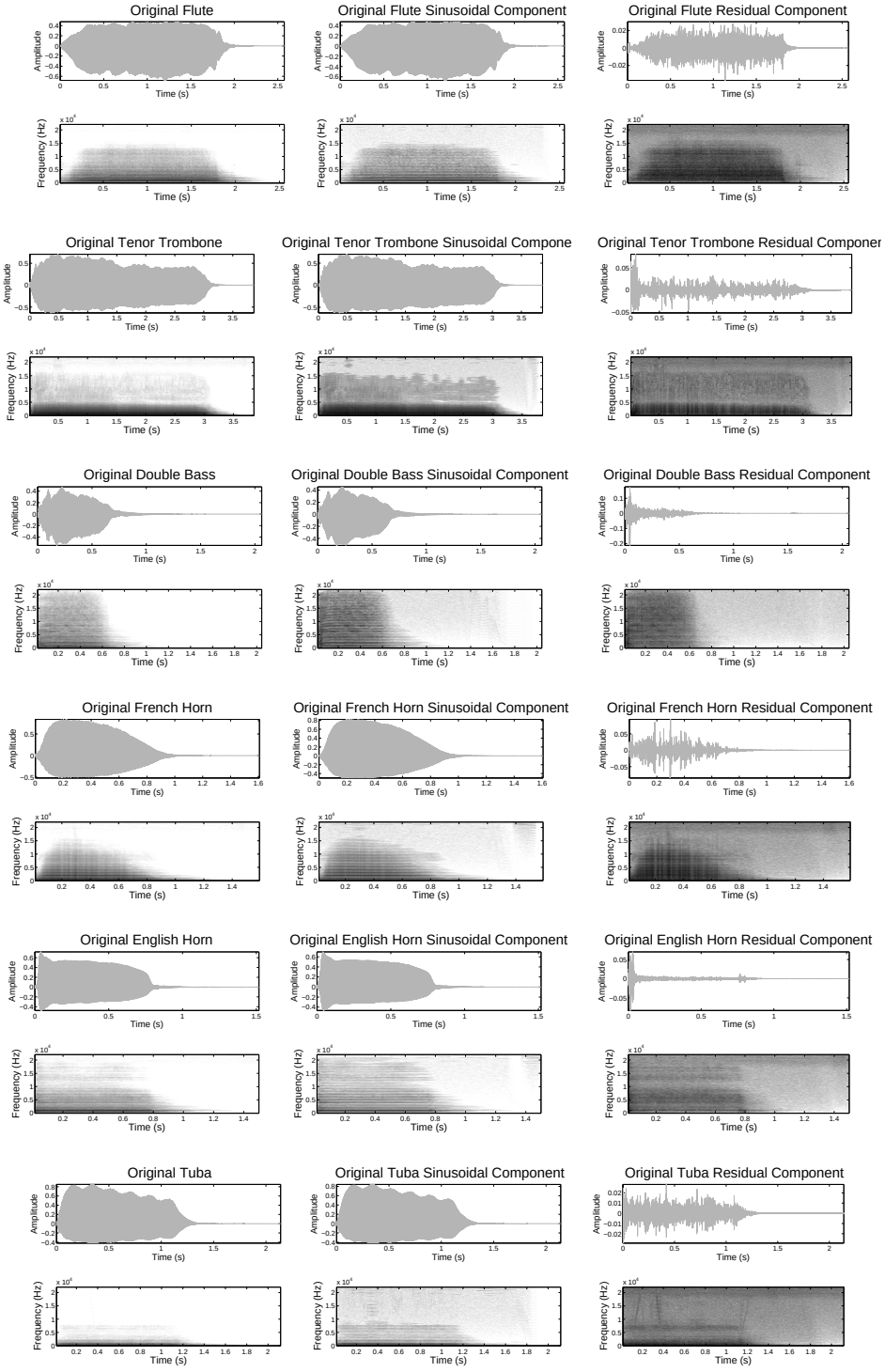

Figure 11.3: Sinusoidal plus residual decomposition. The figure shows the original sound (on the left) with the corresponding sinusoidal component (in the middle) and the residual component (on the right).

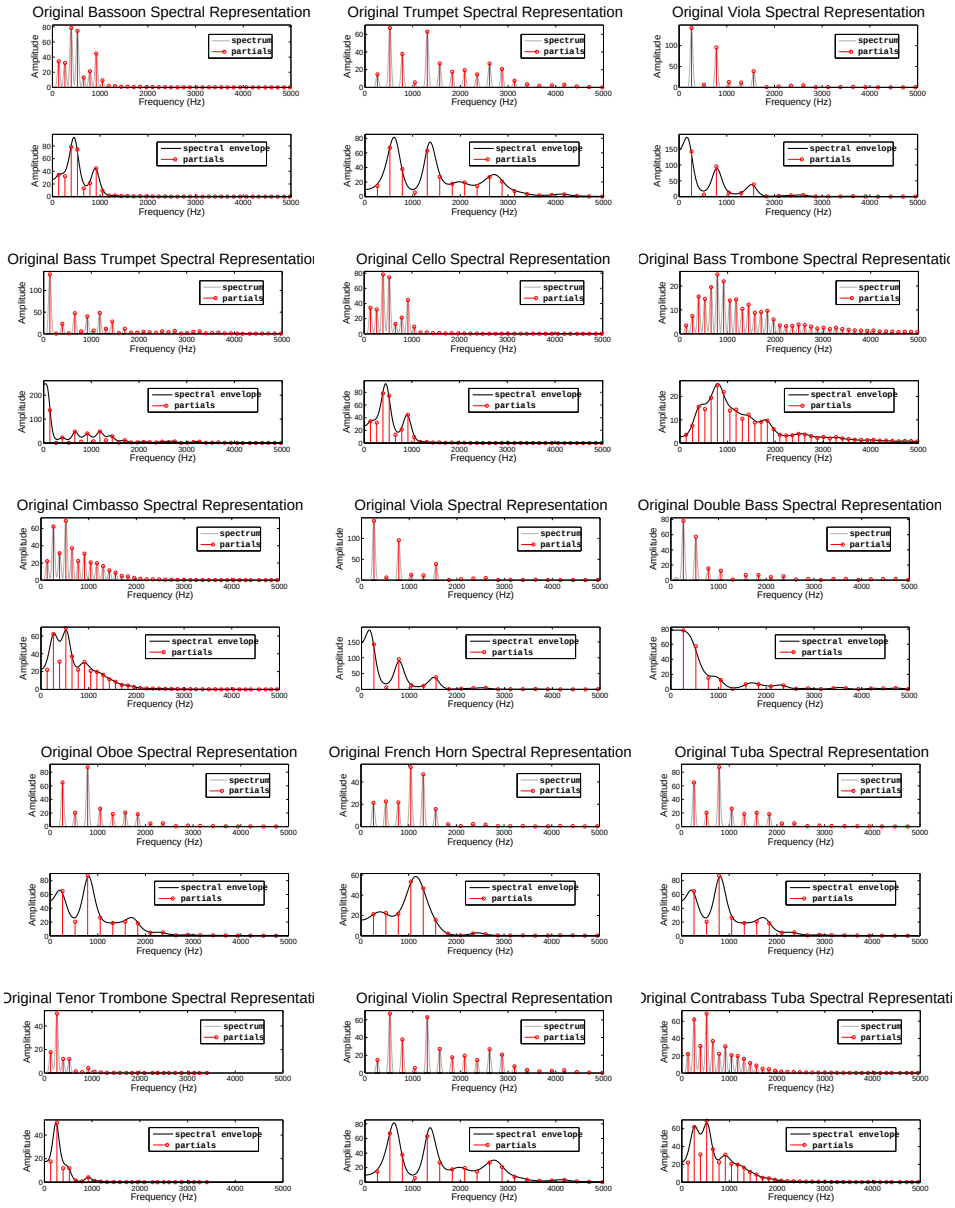

Figure 11.4: Spectral view of the source-filter model. Each figure shows the traditional sinusoidal representation at the top and the source-filter representation at the bottom for one frame.

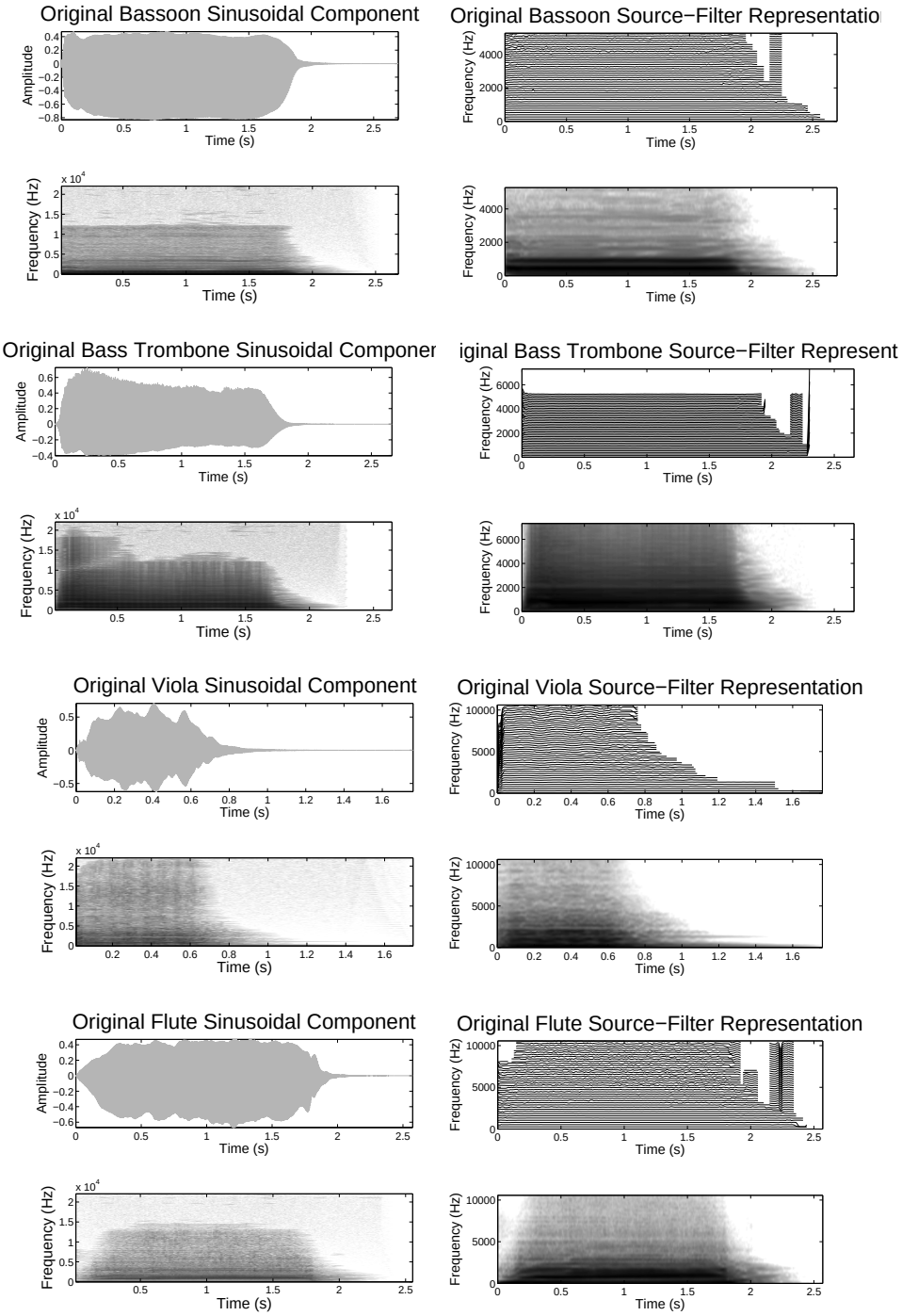

Figure 11.5: Comparison between the spectro-temporal view of the sinusoidal and source-filter representations. The left-hand side shows the sinusoidal spectro-temporal representation (or spectrogram) at the bottom. The right-hand side shows the spectro-temporal representation of the source at the top and the filter at the bottom. The source is represented as the temporal variation of the frequencies of the partials, while for the filter the higher amplitudes are darker, like the spectrogram.

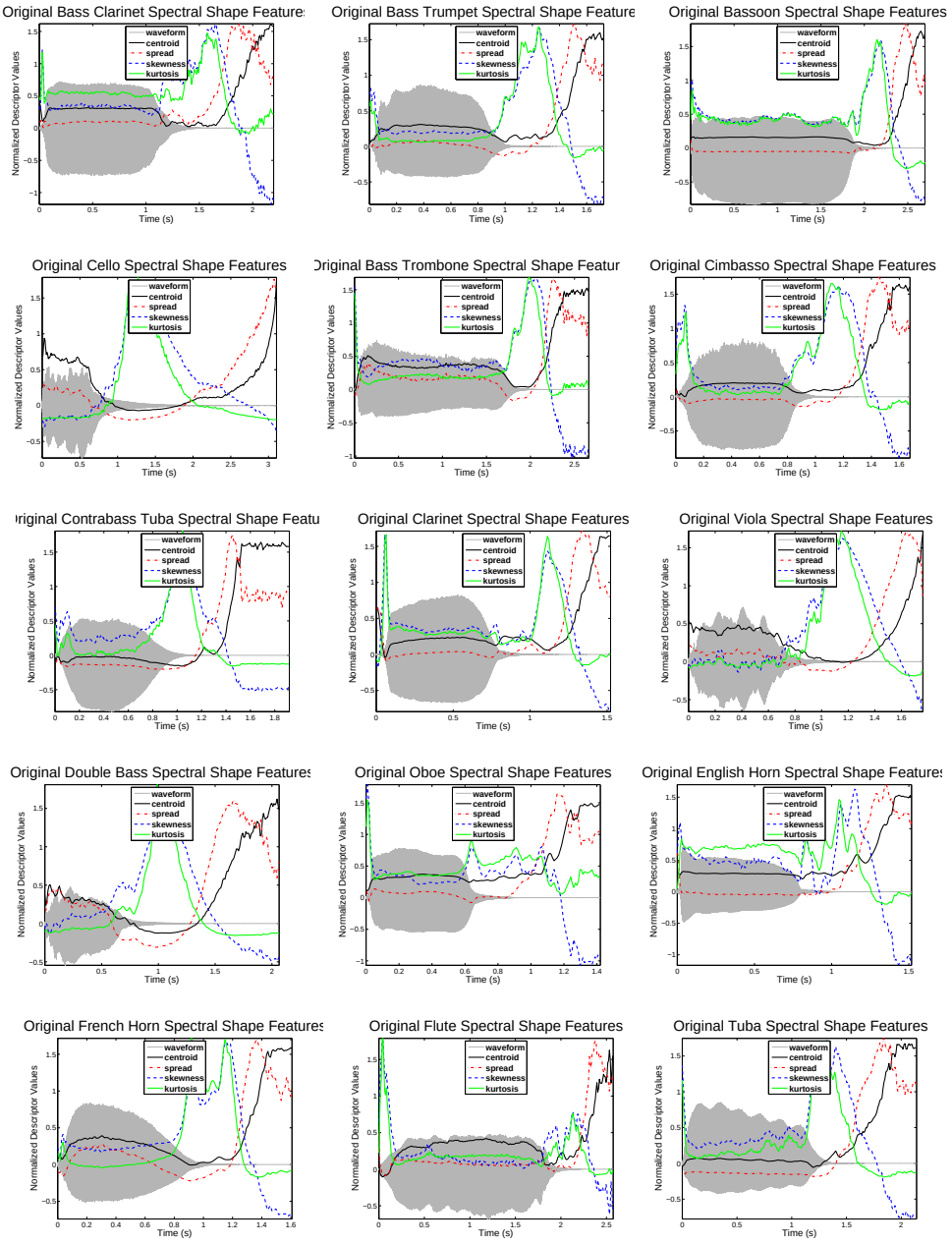

Figure 11.6: Temporal variation of spectral shape features. The figure shows the temporal variation of the spectral centroid, spread, skewness, and kurtosis. Notice how the temporal variation of spectral shape features reveals that they are correlated. It is very important to keep this correlation in the morphed sounds to obtain perceptually meaningful results.

Chapter 12

Temporal Alignment

This chapter addresses the first important step in the morphing algorithm developed in this work, namely, the temporal alignment of perceptually different regions. Chapter 8 explains that the attack, for example, is characterized by fast transients and the sustain part is much more stable. We cannot expect to attain good results if we combine a sound that has a long attack with another sound with a short one regardless of their temporal differences. The region where attack transients are combined with more stable partials will not sound natural.

To achieve a more perceptually seamless morph, we need to temporally align these regions so that their boundaries coincide. To do this, we need a model to correctly identify these regions and their boundaries. Here we propose to use the automatic temporal segmentation based on the ACT model presented in chapter 8 and originally proposed by Hajda [Hajda, 1996]. The temporal alignment of two sounds depends on the previous segmentation.

However, Hajda does not propose an algorithm to automatically segment musical instrument sounds using the model. The automatic segmentation task, which consists in automatically detecting the boundaries of the regions defined by the model, is an important part of the temporal alignment procedure. This thesis proposes an algorithmic procedure [Caetano and Rodet, 2010a] to automate this task using the ACT model. This chapter presents the automatic segmentation method developed, followed by the temporal alignment procedure.

However, any method that provides good estimates of the boundaries of the segments can be used in the temporal alignment procedure. Naturally, annotating the sounds by hand is a possibility that renders excellent results (they are usually considered the ground truth, i.e., the reference against which we measure the quality of automatic estimations). Clearly, the drawback is that annotating by hand is a cumbersome task that takes a lot of effort. We are looking for the tradeoff between automatization and precision in the results.

The attack portion is one of the most salient perceptual regions of musical instrument sounds, so the results of the attack-time detection method I developed are compared against a baseline method proposed by Peeters [Peeters, 2004], hereafter called AR model. The main difference between the methods is the segmentation model they use (ACT and AR). The aim of the comparison is to exemplify the impact of the model in the results of the estimation.

There exist other automatic segmentation methods proposed in the literature, presented in chapter 8. However, this chapter only compares the results of the segmentation using the ACT model against the AR model because preliminary experiments using the AD&SR model and the method of derivatives presented in chapter 8 suggested that this method is too sensitive to the variations of the temporal envelope.

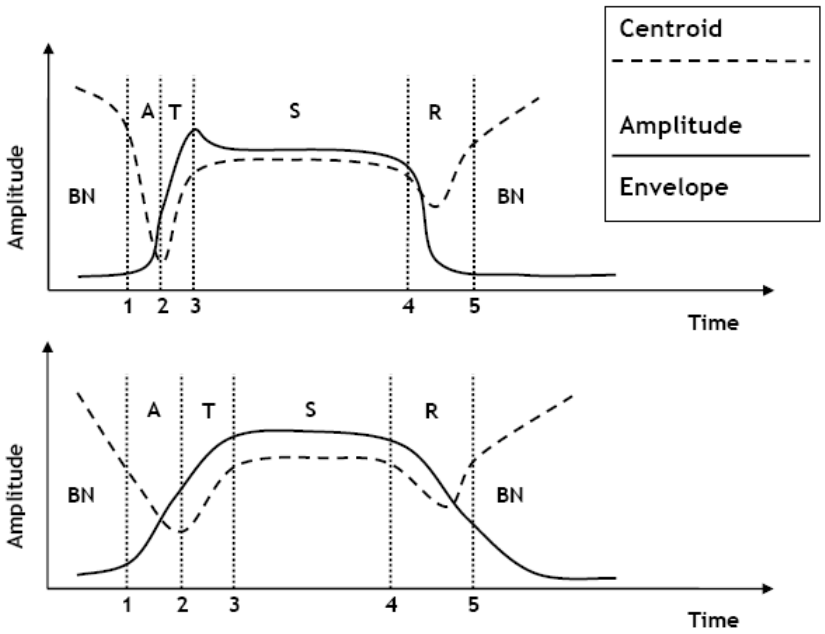

Figure 12.1: Temporal alignment. The figure represents two sounds segmented according to the ACT model. The temporal alignment operation consists of time-stretching or compressing the regions (letters) in order to align the boundaries (numbers) in time.

12.1 Temporal Segmentation

In this work, the temporal segmentation task consists in automatically estimating the boundaries of four perceptually important regions that occur in sustained musical instrument sounds, namely, the attack, the transition, the sustain or steady-state, and the release, defined in chapter 8. The result of each estimation, called marker, is a time value.

The automatic temporal segmentation technique developed here uses the ACT model shown in figure 12.1 and it can be summarized by the block diagram shown in figure 12.2. However, each block involves more operations than explicitly said. For example, according to the block diagram in figure 12.2, the automatic segmentation technique depends solely on the calculation of spectral centroid and temporal envelope. In fact, the calculation of the temporal variation of the spectral centroid itself requires some steps.

The same can be said for the estimation of the temporal envelope, which notably depends on the estimation of the fundamental frequency to define the order of the cepstral model used. The steps involved in the estimation of the temporal envelope were presented in chapter 9, while the calculation of the temporal variation of the spectral centroid was presented in chapter 11.

12.1.1 Automatic Segmentation

The ACT model defines the boundaries of the segments with the aid of changes in the temporal envelope and temporal variation of the spectral centroid. So naturally, before the automatic segmentation, we need to estimate these two detection functions. The accuracy of the results improves significantly when the temporal envelope describes the overall energy evolution of the sound. Therefore the RMS temporal envelope estimation is used to improve the results of the

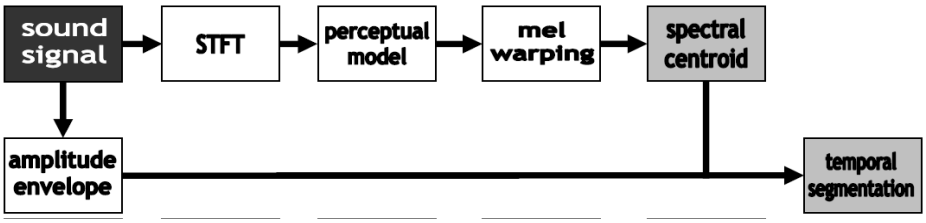

Figure 12.2: Temporal segmentation. The sound that will be segmented is represented by the block with a dark background, the signal processing steps have a white background, and the feature extraction steps have a light gray background.

segmentation task.

Also, the temporal variation of the spectral centroid needs to be a smooth curve so that we can easily and robustly detect the changes that carry important information (minimum and last inflection point). The boundaries to be detected are the following

- onset detection (1);
- end of attack/beginning of transient (2);
- end of transient/beginning of sustain (3);
- end of sustain/beginning of release (4);
- offset (5).

Next, the proposed procedure to automatically detect each boundary for the segmentation task will be explained.

12.1.2 Automatic Detection of Boundaries

In this section, I will describe the automatic detection algorithm that gives the markers (time values corresponding to the boundaries of the regions) in the actual order in which they are calculated, as follows

1. Detect the onset (1);
2. Going backwards in time, detect the offset (5) as the first frame with the same amplitude as the onset (this step uses the temporal envelope);
3. Detect the beginning of sustain (3) using the method of efforts with $M = 3$ (this step uses the temporal envelope);
4. Going backwards in time, detect the beginning of release (4) using the method of efforts with $M = 5$ (this step uses the temporal envelope);
5. Detect the end of attack (2) as the minimum value of the centroid between (1) and (3) (this step uses the spectral centroid).

The first important event to be detected is the onset of the sound. An onset detection algorithm that detects transients as variations of phase [Röbel, 2003] is used. Naturally, other onset detection algorithms could have been used. The chosen approach revealed to be very robust and accurate.

The next boundary to be detected is the beginning of the sustain. The method of adaptive efforts proposed by Peeters [Peeters, 2004] is used in this step. Then, the end of the attack is detected as the minimum of the centroid between the onset and the beginning of the sustain.

The method of adaptive efforts is originally used to detect the first inflection point of the temporal envelope (corresponding to the end of the attack according to the original AR model). This thesis proposed [Caetano and Rodet, 2010a] to adapt it to detect the beginning of the release from the temporal envelope too by simply inverting the logic behind the beginning of sustain detection.

Finally, the offset is defined as the first point after the beginning of the release where the temporal envelope attains the same value as the onset.

The result of the automatic segmentation method developed in the context of this work are the five time markers used in the temporal alignment procedure. The next section shows examples of the automatic segmentation and compares them against the baseline method [Peeters, 2004] that uses the AR model.

12.1.3 Examples of Automatic Segmentation

The automatic segmentation algorithm developed in the course of this thesis [Caetano and Rodet, 2010a] will be illustrated next in figures 12.3 to 12.9. The results of the automatic segmentation will be presented separately for each family of instruments considered (selected from the Vienna database and briefly presented in chapter 10), woodwinds, brass, and bowed strings.

Each figure shows the results for a specific instrument family. The baseline method against which I compare my results is the attack-rest (AR) model proposed by Peeters [Peeters, 2004]. Naturally, annotations by hand would ideally be considered the "ground truth" segmentation. The difficulty in this case is that the segmentation depends on the annotator, and we would need to collect data from several people and include statistical analysis in the results. If we take into account that the segmentation is not the aim of this work, but just a means, then the statistical relevance of the annotations is out of the scope of this work. As a compromise, I decided to include the spectrogram in the figures and let the readers judge by themselves by visual inspection.

Each figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of both segmentation methods.

The markers (numbers) for the ACT segmentation method are shown as solid vertical lines, while the AR markers are shown as dashed vertical lines with the corresponding time values on top. The purpose of the figures is to show that, in general, the ACT method outperforms the AR. The ACT method is expected to give more accurate and robust estimations in general because it uses spectro-temporal information, rather than only temporal information like the AR method.

However, there are some specific cases where, for the same instrument, the estimates given by the ACT method are visually poor, although rarely outperformed by AR. Each figure illustrates one case, aiming to show that the accuracy and robustness of the estimation actually depends on the waveform, and not on the instrument. The conclusion, naturally, is that we would need a specific model of the instrument to increase accuracy and guarantee robustness of estimation for notes played by the same instrument.

Originally, the ACT model was proposed for sustained (nonpercussive) sounds only. This work [Caetano and Rodet, 2010a] tested the model on any type of excitation (percussive and sustained), using woodwind, bowed and plucked strings. The conclusion was that in general the model does not apply to nonsustained (percussive) sounds.

12.1.3.1 Woodwinds

In this section, the results of the automatic segmentation method using both the ACT and AR models are compared for the following woodwinds: bass clarinet, bassoon, clarinet, english horn, flute, and oboe. In general, the result of the automatic segmentation is satisfactory when applied to woodwinds because the ACT model applies well to this instrumental family.

12.1.3.2 Brass

In this second part we will see the comparison of the automatic segmentation for the following brass instruments: bass trumpet, cimbasso, contrabass tuba, french horn, tenor trombone, trumpet, and tuba. Once again the result of the automatic segmentation is satisfactory when applied to brass instruments, even though it was empirically determined that the minimum of the centroid at the end of the attack is not so clear for this instrumental family, which could render a poor automatic estimation.

12.1.3.3 Strings

Finally, in the third part, the results of the automatic segmentation is compared for the following strings: cello, double bass, viola, and violin. In this case, empirical results indicate that the ACT model does not describe very accurately the temporal evolution of bowed strings. Even though the result of the automatic segmentation is satisfactory, we verified that the behavior of the spectral centroid does not correspond well with the predicted by the model, notably presenting more than one valley or a very shallow one, hard to detect automatically.

The robustness of the automatic segmentation method using the ACT model is yet to be tested. However, for all the sounds tested the results seem to better suit the characteristics of each segment when compared to the baseline AR method.

12.2 Temporal Alignment

Once both sounds are segmented and the boundaries of A, T, S and R are estimated, the temporal alignment process is simple, as represented in figure 12.1. For each sound, the length of each region (labeled with letters) is measured by computing the time difference using the markers (numbers). The length of the attack is represented in logarithmic scale as defined in equation 5.2, where at_1 and at_2 correspond respectively to the markers (1) and (2) shown in figure 12.1. For the other regions the representation is linear, as expressed below.

$$\begin{cases} lat = \log A = \log\left(at_2 - at_1\right) = \log\left[(2) - (1)\right] \\ T = \left[(3) - (2)\right] \\ S = \left[(4) - (3)\right] \\ R = \left[(5) - (4)\right] \end{cases} \tag{12.1}$$

where lat is the log attack time, and A is the linear attack time. The other letters correspond directly to the ACT model. The next step is to calculate the length of each segment in the morphed sound. The attack time is perceived logarithmically [Caclin et al., 2005, Grey and Gordon, 1977, Krimphoff et al., 1994, Krumhansl, 1989, McAdams et al., 2005], so we should interpolate its logarithm and retrieve the corresponding linear value A_{12}.

12.2.1 Interpolation of Lengths

Equation 12.2 shows how to obtain the length of the segments in the morphed sound by interpolation. The interpolated lengths are represented by a letter that indicates the segment and subscripts indicating both sounds. For example S_{12} stands for the sustain of the morphed sound, obtained by interpolation between S_1 and S_2, the sustain of the sounds used in the morph.

$$\begin{cases} \log A_{12} = \alpha \log A_1 + [1-\alpha] \log A_2 \\ T_{12} = \alpha T_1 + [1-\alpha] T_2 \\ S_{12} = \alpha S_1 + [1-\alpha] S_2 \\ R_{12} = \alpha R_1 + [1-\alpha] R_2 \end{cases} \tag{12.2}$$

In equation 12.2, S_1 represents the length of the sustain region for the first sound and S_2 for the second sound, and S_{12} is the calculated length of the sustain that the morphed sound should have. In order to guarantee that the hybrid sound will present attack transients during A_{12}, stable partials during S_{12}, etc, we align these segments in time for the original sounds by time-stretching/compressing each region by the appropriate factors, calculated as explained in the next section.

12.2.2 Calculate Time-Stretch Factors

The stretch/compress factors are then calculated according to the following

$$\begin{cases} \nu_{A_1} = \frac{A_1}{A_{12}}, & \nu_{A_2} = \frac{A_2}{A_{12}} \\ \nu_{T_1} = \frac{T_1}{T_{12}}, & \nu_{T_2} = \frac{T_2}{T_{12}} \\ \nu_{S_1} = \frac{S_1}{S_{12}}, & \nu_{S_2} = \frac{S_2}{S_{12}} \\ \nu_{R_1} = \frac{R_1}{R_{12}}, & \nu_{R_2} = \frac{R}{R_{12}} \end{cases} \tag{12.3}$$

where ν represents the time stretch/compress factor and the subscript follows the same convention as in the previous section. Factors $\nu > 1$ represent temporal stretching, while factors $0 < \nu < 1$ represent temporal compression of the corresponding region.

12.2.3 Temporal Alignment

Finally, the temporal alignment is done by simply applying the temporal stretch/compress factors ν calculated for each sound to each corresponding region. Notice that when $\nu_{S_1} > 1$, it follows that $0 < \nu_{S_2} < 1$, because of how they are calculated. This simply means that the length of the corresponding segment S_{12} is intermediate between S_1 and S_2.

After the temporal alignment operation, the boundaries of the segments for both source and target sounds will be in different time positions (indicated by a prime, such that (3)' is the position is the beginning of the sustain after temporal alignment) that can be calculated according to the following

$$\begin{cases} (1)'_1 = (1)_1, & (1)'_2 = (1)_2 \\ (2)'_1 = \nu_{A_1} (2)_1, & (2)'_2 = \nu_{A_2} (2)_2 \\ (3)'_1 = \nu_{T_1} (2)'_1, & (3)'_2 = \nu_{T_2} (2)'_2 \\ (4)'_1 = \nu_{S_1} (3)'_1, & (4)'_2 = \nu_{S_2} (3)'_2 \\ (5)'_1 = \nu_{R_1} (4)'_1, & (5)'_2 = \nu_{R_2} (4)'_2 \end{cases} \tag{12.4}$$

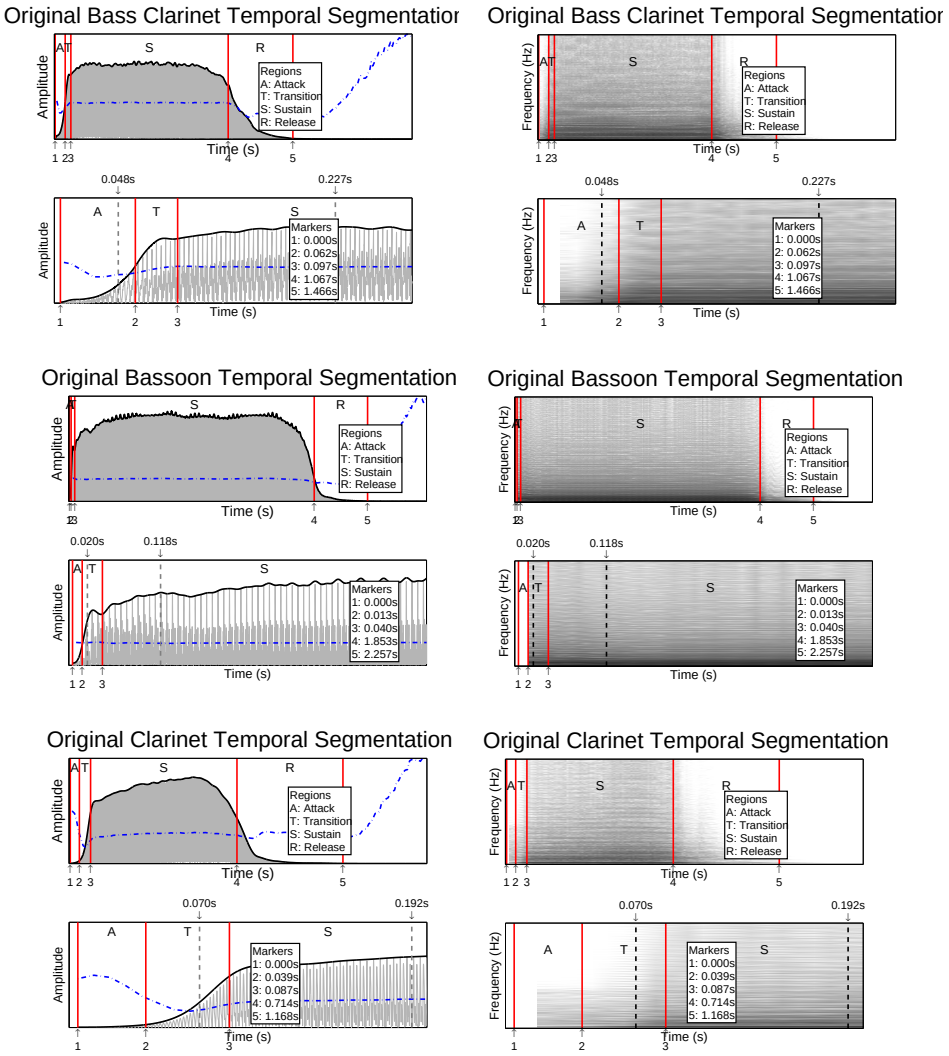

Figure 12.3: Automatic segmentation for the woodwinds. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

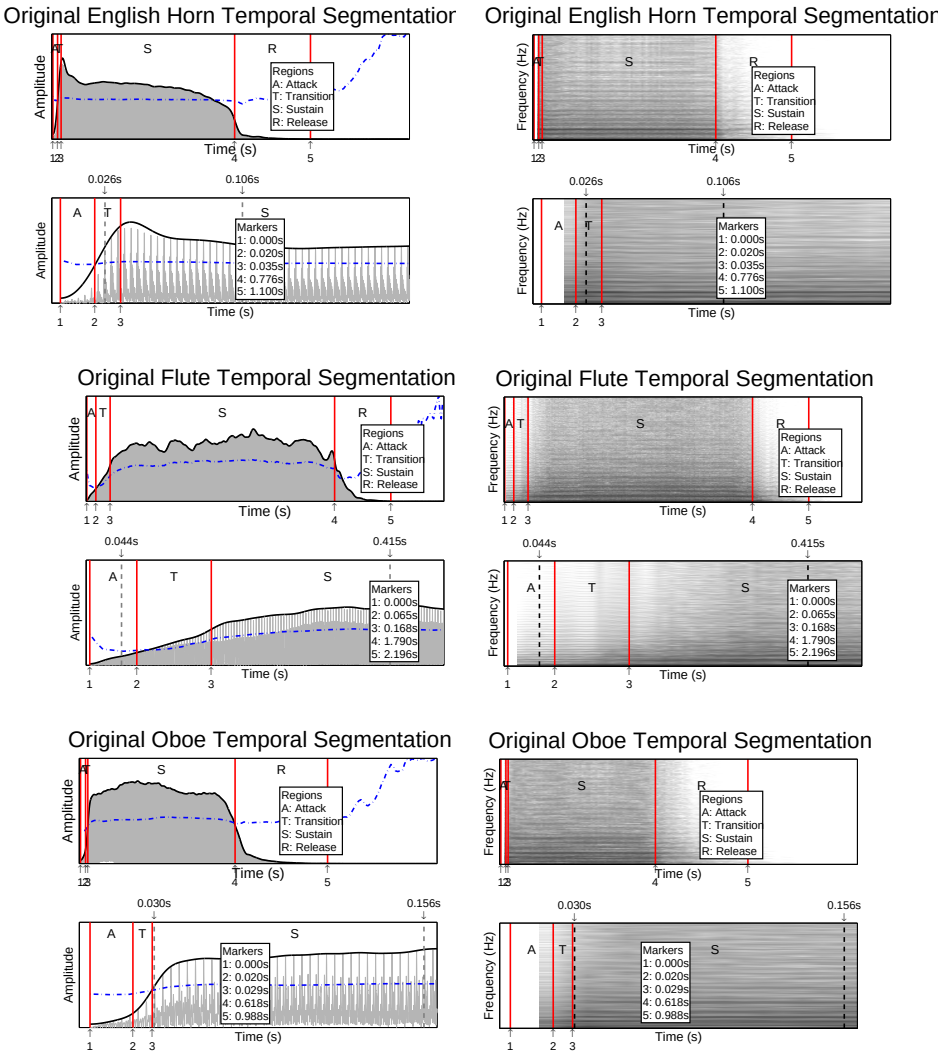

Figure 12.4: Automatic segmentation for the woodwinds. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

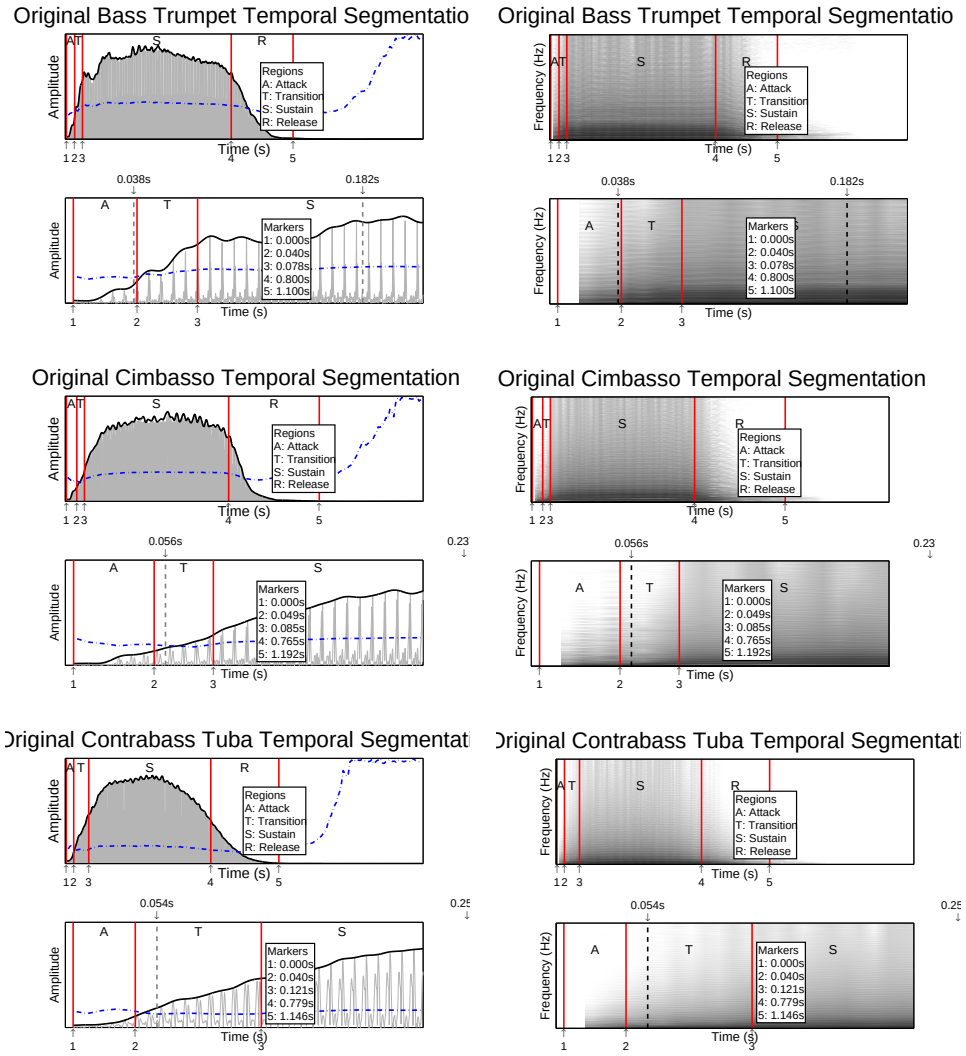

Figure 12.5: Automatic segmentation for brass. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

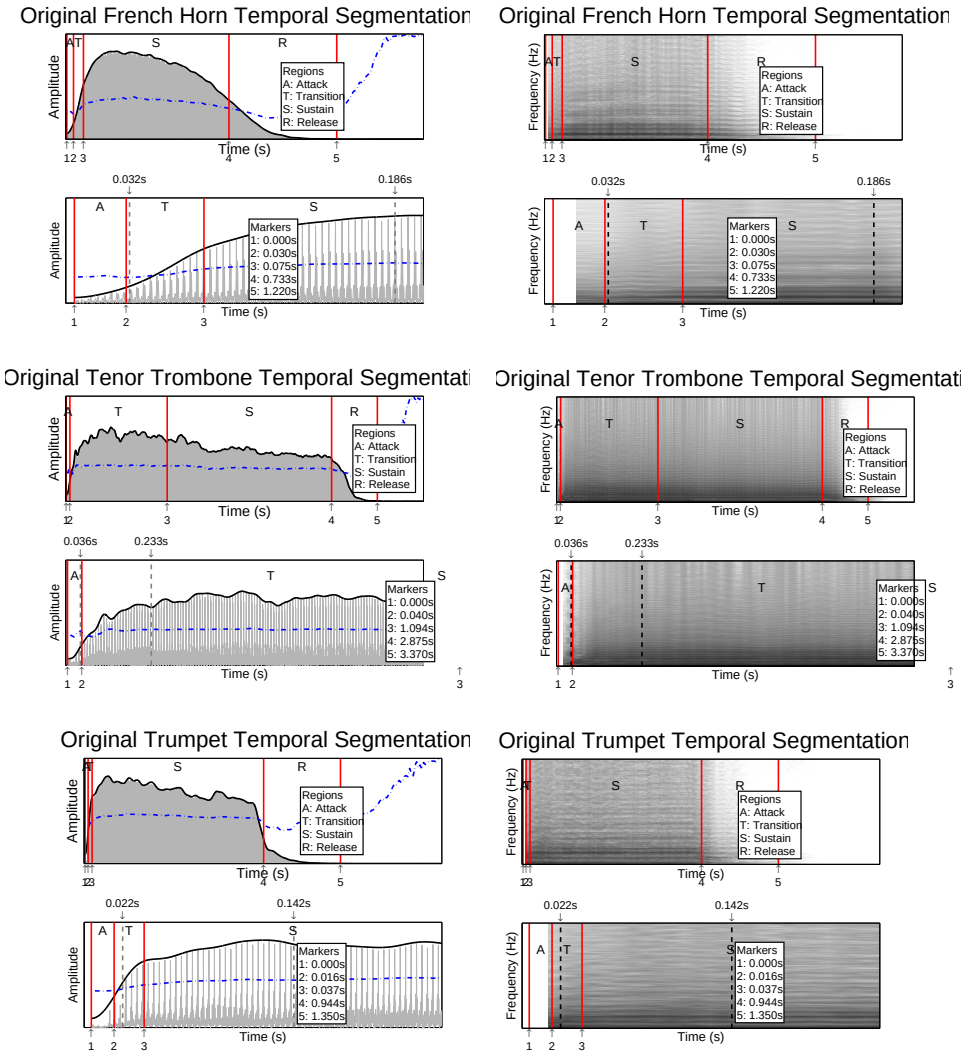

Figure 12.6: Automatic segmentation for brass. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

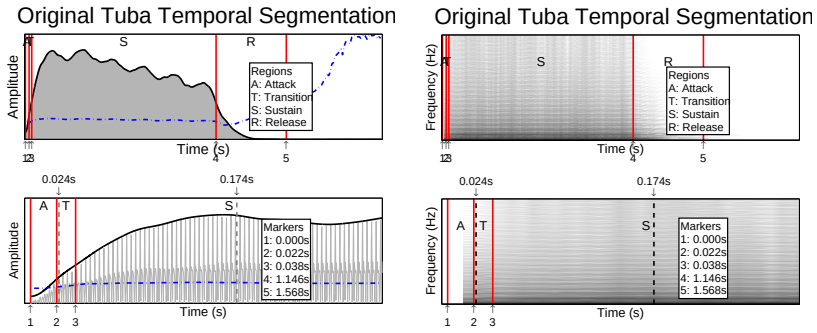

Figure 12.7: Automatic segmentation for brass. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

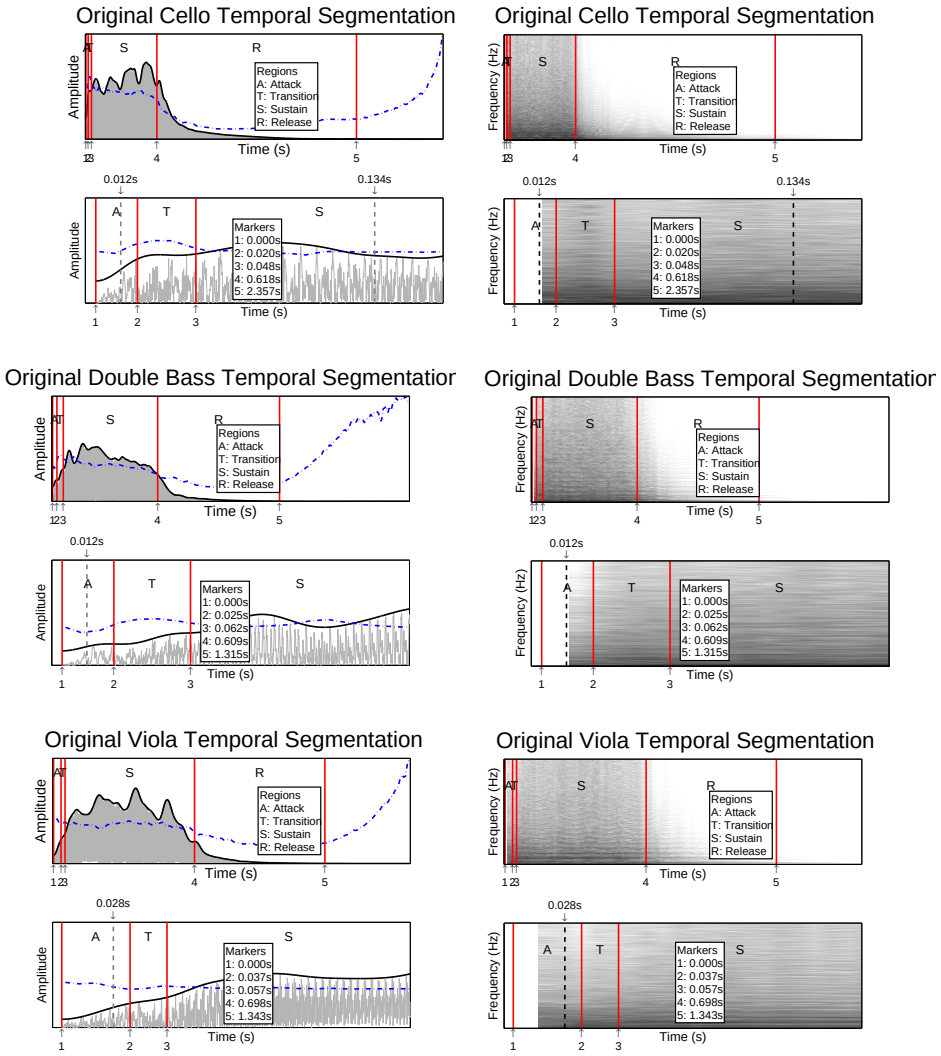

Figure 12.8: Automatic segmentation for strings. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

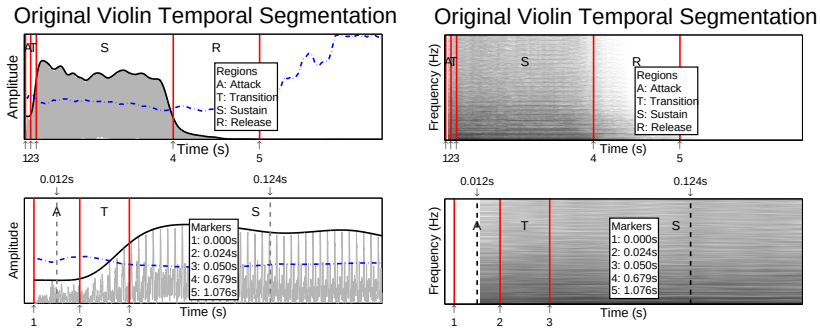

Figure 12.9: Automatic segmentation for strings. The figure compares the results of the automatic segmentation with the ACT and AR methods. The figure shows the markers (numbers) as solid vertical lines, and the segments (letters) against the waveform (left-hand side) and the spectrogram (right-hand side). The top of the figures show the whole duration of each sound to provide a global view of the segmentation results. At the bottom of the figures we see a zoom of the attack (A) and transient (T) segments with the purpose of comparing the results of the ACT method (solid line) with the AR method (dashed line).

Chapter 13

Spectral Envelope Morphing

In this chapter, we will consider the problem of morphing spectral envelopes. Spectral envelope morphing involves the combination of two (or more) spectral envelopes to produce another one with intermediate characteristics. The application dictates which characteristics we should take into consideration, and therefore how we should perform the combination of spectral envelopes. For example, in concatenative synthesis, it is usually necessary to even out the edges between units to make the transitions smoother. The techniques applied depend on the sound material (speech units, etc) and on the type of transition that we are trying to achieve. When morphing between musical instrument sounds, we usually want the result to sound like a hybrid between the instruments used in the morph. This requirement imposes certain restrictions on the resultant morphed spectral envelopes that make some spectral envelope morphing techniques more appropriate than others.

The spectral envelope is one of the most important characteristics of musical instrument sounds. Perceptually, the spectral envelope is related to musical instrument (sound source) recognition and timbre perception. Helmholtz [Helmholtz, 1885] was among the first to investigate the relationship between the relative amplitudes of the partials and timbre perception for musical instrument tones (pitched sounds or notes). In the source-filter model, the spectral envelope models the filter, as explained in chapter 6. The filter is associated with the subset of attributes of timbre that Slawson dubbed sound color [Slawson, 1985]. The peaks of the spectral envelope, also called formants, are the frequency regions of higher energy in the spectrum. The formant peaks reflect the result of the interaction between the excitation and the natural modes of vibration of the body of the instrument. These natural modes of vibration depend mostly on the material of the resonant cavity and its shape, and can be considered a sort of signature of the instrument

But we know that one instrument can present timbral variations (like the different registers of the clarinet or "brassy" trumpet sounds). These timbral differences are manifested in many spectro-temporal features, the balance of spectral energy being among the most important. The spectral shape features (spectral centroid, spread, skewness and kurtosis) are a measure of the balance of spectral energy. Some verbal attributes used by musicians and composers to refer to specific qualities of musical instrument sounds, such as bright, have an acoustic correlate, the spectral centroid. For example, a brassy trumpet sound, usually considered brighter than softer ones, has a higher spectral centroid.

Therefore, when morphing between musical instrument sounds, we need to take into account not only the formants, but also the balance of spectral energy. On the one hand, we can keep track of the formants if we study the behavior of the spectral envelope peaks. On the other hand, the balance of energy can be measured by the values of the spectral shape features.

The aim of this chapter is to investigate the problem of morphing spectral envelopes from both perspectives, behavior of the formant peaks and variation of spectral shape features. The

spectral envelope morphing techniques proposed in the literature will be evaluated from these perspectives considered above using the morphing evaluation criteria adopted in this work, namely, correspondence, intermediateness, and smoothness. Ideally, we are looking for a spectral envelope morphing technique that presents both a desireable behavior of spectral peaks and linear variation of spectral shape features at the same time. The aim is to select the spectral envelope morphing technique that leads to the optimal variation of spectral shape features under the constraints of intermediateness and smoothness. The concept of optimal in the sense of variation of spectral shape features will be discussed, leading to the evaluation of spectral envelope morphing presented in chapter 14.

13.1 Morphing Spectral Envelopes

Before presenting the approach developed in this thesis to morph spectral envelopes, let us consider the previous solutions. The problem of morphing spectral representations of sounds has been studied before in the context of sound morphing [Slaney et al., 1996, Ezzat et al., 2005, Moorer, 1978]. Notably, Ezzat [Ezzat et al., 2005] analyzes soberly the problem of interpolating spectral envelopes and argues that any reasonable solution should account for proper formant shifting between source and target.

Slaney [Slaney et al., 1996] proposes to morph spectral envelopes by cross-fading between the the MFCCs that represent each spectral envelope. The conclusion is that the method should be improved with more perceptually related representations of the spectral contents. Pfitzinger [Pfitzinger, 2004] uses dynamic frequency warping (DFW), a frequency domain counterpart of the widely known dynamic time warping algorithm, in a spectral smoothing approach applied to concatenative speech synthesis. Ezzat [Ezzat et al., 2005] studied the use of DFW to morph spectral envelopes in the context of musical sounds. The above mentioned approaches focus on the behavior of the spectral peaks under the transformation, but they rarely consider the balance of spectral energy.

Only recently did we start to take the variations of spectral shape features such as the spectral centroid into consideration [Williams and Brookes, 2007, Williams and Brookes, 2009, Caetano and Rodet, 2009, Caetano and Rodet, 2010c, Hatch, 2004], and the result is the addition of another step in the process, feature extraction. We note that, while the choice of features is definitely important, most research efforts concerning morphing guided by the values of features have concentrated on the challenging problem of feature guided transformations [Verfaille et al., 2006].

More specifically, the most important aspect of spectral envelope morphing is the resultant transition between the spectral envelope curves. There are two different things to take into consideration when analyzing the transition between spectral envelope curves, the behavior of the formant peaks, and the balance of spectral energy. The formant peaks are associated with sound color [Slawson, 1985], while the balance of spectral energy, which can be measured with the spectral shape features, is correlated with timbral qualities [Caclin et al., 2005, Krumhansl, 1989, Krimphoff et al., 1994, McAdams et al., 2006, McAdams et al., 2005]. Ideally, when morphing between musical instrument sounds, we want intermediate representations of the spectral envelopes that lead to spectral envelope curves with peaks in positions that would correspond to a hybrid instrument between the two, and that would also have an intermediate balance of spectral energy. In the next section, we will consider both criteria, first separately and then at the same time.

We will consider two possible transitions between spectral envelope curves concerning the spectral peaks, namely, spectral peak shifting and spectral peak rise and wane. One way of describing the peaks of the spectral envelope curve is in terms of their center frequency and bandwidth or magnitude. The center frequencies of the spectral peaks are the spectral regions of high energy, while their magnitudes and bandwidths reflect the energy loss (usually due to damping). Spectral

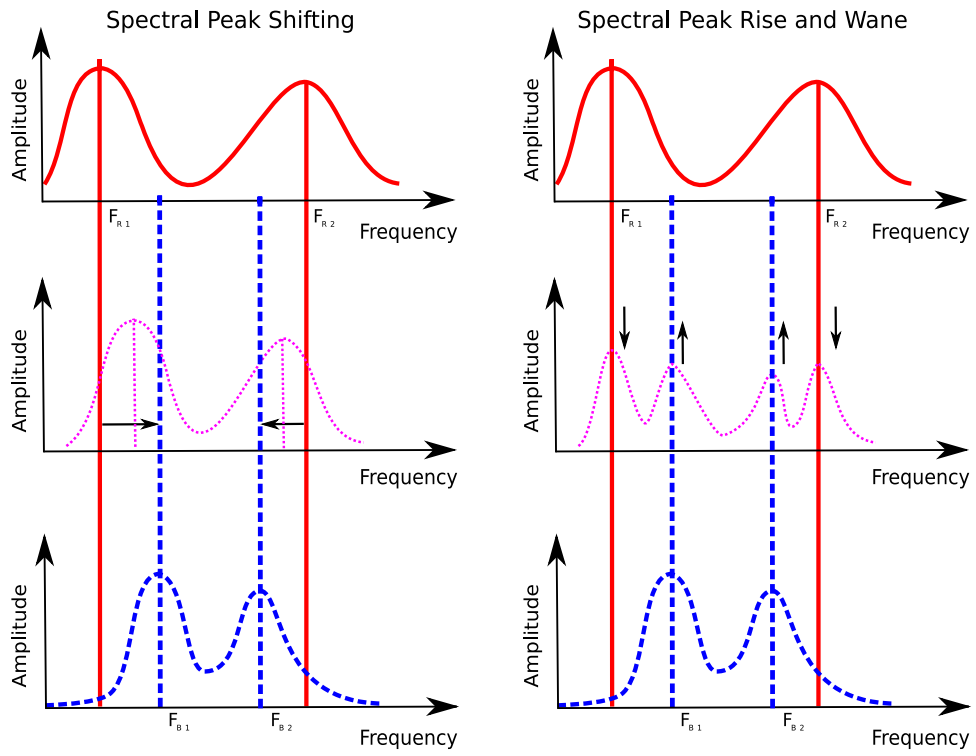

Figure 13.1: Hybrid spectral envelopes. The figure shows two spectral envelopes with peaks in different absolute positions in the frequency axis. When producing hybrid spectral envelopes that correspond to a gradual transformation between those two, on the lefthand side of the figure we imagine a one-to-one Correspondence between the spectral peaks, whose center frequencies simply shift in frequency. On the righthand side we imagine that the spectral peaks simply appear and disappear remaining in the same absolute positions in frequency.

peak shifting and spectral peak rise and wane describe the behavior of the spectral peaks regarding their center frequencies and magnitudes.

13.1.1 Spectral Peak Shifting

Just like the name suggests, when this type of transition occurs, the center frequencies of the (formant) peaks shift in frequency. This is illustrated on the left-hand side of figure 13.1, which shows two spectral envelopes with two (formant) peaks each. When producing morphed spectral envelopes that correspond to a gradual transformation between those two, we imagine a one-to-one correspondence between the (formant) peaks, and consequently their center frequencies simply shift in frequency. We should notice that the morphed spectral envelope curves corresponding to intermediate steps of this transformation keep the same number of peaks as the original curves.

13.1.2 Spectral Peak Rise and Wane

Another possible transition between two spectral envelope curves concerning the spectral peaks in which the spectral peaks simply appear and disappear (rise and wane) is shown in figure 13.1. In this transition, the center of the formant peaks do not shift in frequency, instead, they remain fixed during the transition while their amplitudes increase or decrease (depending on the direction of transformation). Notice that, in this type of transition, the morphed spectral envelope curves corresponding to intermediate steps do not keep the same number of formant peaks as the original curves, because the intermediate spectral envelopes would have twice as many peaks as them, as illustrated on the right-hand side of figure 13.1.

Naturally, both transitions proposed above rely on an underlying correspondence between the spectral envelope curves used in the transformation in terms of the number of formant peaks. We have supposed for both possibilities that each formant peak on a spectral envelope curve has a partner on the other one. This leads us to the following question, when there is no peak to peak correspondence, is it still possible to propose reasonable solutions to the problem? And more importantly, supposing that there exist solutions, what is their perceptual quality or how do they sound? Let us first consider the problem of (formant) peak correspondence under spectral envelope morphing transitions. Then, later in chapter 14 we will finally investigate the much more important question in the context of this work on how they sound when applied in morphed sounds.

13.1.3 No Spectral Peak Correspondence

Figure 13.2 illustrates the case when one spectral envelope has more peaks than the other used in the transformation. Let us analyze two possible transitions between them inspired by the spectral peak shifting and rise and wane paradigms.

When we imagine a gradual transformation between them under the spectral peak shifting paradigm, the two peaks' center frequencies will shift to the same center frequency and merge into one. This means that all the hybrid spectral envelopes will have two peaks with center frequencies that get closer and closer (or farther and farther apart, depending on the direction of the transformation). On the other hand, when we imagine the same gradual transformation using the rise and wane model, all the morphed spectral envelopes now have three formant peaks.

The problem of correspondence between spectral peaks has been addressed independently by Osaka [Osaka, 1998] for sinusoidal peaks and Laura [Laura and Rodet, 1990] in the context of spectral envelope peaks. Both solutions proposed rely on a description of each spectral peak in terms of center frequency and bandwidth/amplitude, and attack the problem trying to find the best match for each peak in terms of a distance measure between their center frequencies. Algorithmically, the solution involves combinatorics and minimization of the distance measure, which can be very complex and time-consuming to solve.

In this work, I propose an alternative solution to this problem that always guarantees correspondence between the spectral representations. Notice that I will reason in terms of spectral envelopes, but the same arguments apply to plain Fourier spectra. The idea is fairly simple, we can work around the lack of correspondence between the number of spectral peaks, which arises when we describe the spectral peaks individually, if we represent the spectral envelopes using a spectral envelope model instead. In this case, all we have to do to guarantee correspondence is to use the same number of coefficients (i.e., the same model order) for both spectral envelopes used in the transformation. In principle, this solves the correspondence problem elegantly. There is one important consideration, though. Now we are interpolating the parameters of spectral envelope models to obtain the morphed spectral envelopes (and consequently the corresponding morphed spectral envelope curves). So we should study the behavior of the spectral peaks when we change (interpolate) the parameters of different spectral envelope models. Naturally the question about

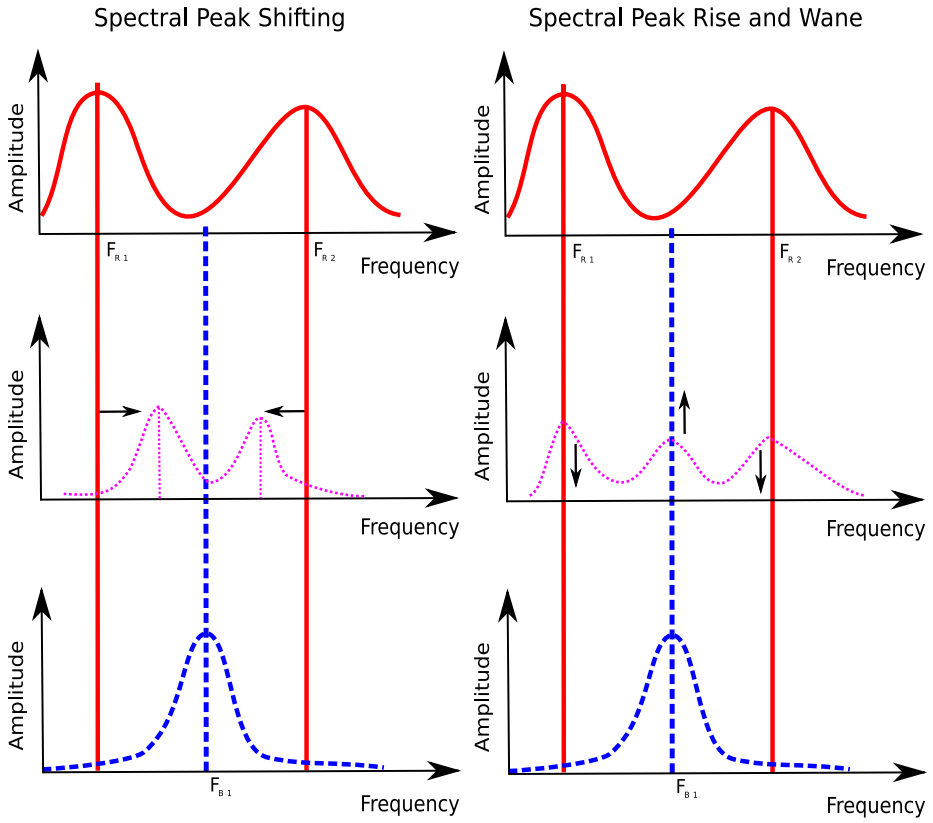

Figure 13.2: No spectral peak correspondence. The figure illustrates both spectral peak shifting and rise and wane spectral envelope morphing paradigms when there is no correspondence between spectral peaks.

the behavior of the spectral peaks remains the same as before. As we will see in the next section, the answer depends on how the spectral envelope model encodes information about the spectral envelope curve.

13.1.4 Spectral Envelope Morphing

Spectral envelope morphing is a spectral envelope transformation technique that takes two (or more) spectral envelope models as input as produces one as output. Spectral envelope morphing can be understood as the convex combination of the parameters of the spectral envelope representation of a given spectrum. Given two vectors of parameters $\boldsymbol{\sigma}_p$ and $\boldsymbol{\sigma}_q$ of a spectral envelope model whose map is S, the convex combination between two spectral envelopes is defined as the convex combination of the parameters of the representation as expressed in equation 2.4 and rewritten below.

$$H_{p,q}(\omega) = S(\boldsymbol{\sigma}_{p,q}) = S(\alpha\boldsymbol{\sigma}_p + [1-\alpha]\boldsymbol{\sigma}_q) \tag{13.1}$$

Naturally this operation is not the only possibility to obtain a spectral envelope as the com-

bination of two. Notably, we can always perform a convex combination of the spectral envelope curves directly instead of the parameters, as expressed in equation 13.2. Naturally the result of the operations defined in equations 13.1 and 13.2 usually is not the same for most spectral envelope representations.

$$H_{p,q}(\omega) \neq \alpha H_p(\omega) + [1-\alpha] H_q(\omega) = \alpha S(\boldsymbol{\sigma}_p) + [1-\alpha] S(\boldsymbol{\sigma}_q) \tag{13.2}$$

In order to exemplify the behavior of the spectral peaks under the interpolation of different spectral envelope representations, I created an artificial example based on an all-pole representation (equivalent to LPC) with four poles each (which guarantees correspondence between them if we use the same model order). Figure 13.3 shows the two original artificial spectral envelope curves. At the top we see the spectral envelope curve resulting from the spectral envelope model with two poles at center frequency $F_1 = 5$KHz and magnitude $r_1 = 0.91$ and two other poles at center frequency $F_2 = 10$KHz and magnitude $r_2 = 0.88$. At the bottom we see the spectral envelope curve resulting from placing all four poles at center frequency $F_3 = 8$KHz with magnitude $r_3 = 0.80$ each.

Figure 13.3 also shows the line spectral frequencies resulting from the conversion from the all-pole filter coefficients. Notice how the line spectral pairs tend to concentrate around the peaks of the spectral envelope curve in both situations, notably when there is only one peak the line spectral frequencies are distributed fairly equally across the spectrum. Now we are going to convert the all-pole representation into cepstral coefficients and compare the behavior of the spectral peaks when we perform the convex combination of the line spectral frequency and cepstral representations of these two spectral envelope curves.

Figure 13.4 shows (from two perspectives) the spectral envelope curves resulting from the interpolation of the line spectral frequencies (LSFs) shown in figure 13.3 (on the left-hand side) and their cepstral coefficient (CC) representation (on the right-hand side). Since we are interested in the gradual transition, we use several values of interpolation factor $\alpha = [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$. At the top of figure 13.4 all curves are shown superimposed, and at the bottom they are shown from the top (as in a topographic map) for each value of α separately.

The purpose of figure 13.4 is to exemplify the behavior of the spectral peaks when we interpolate the values of parameters of different spectral envelope representations. The bottom plot in figure 13.4 shows that the center frequencies of the peaks shift when interpolate LSFs, while the interpolation of CCs leads to the increase/decrease in magnitude of the peaks while their center frequencies do not change value. Remark one important detail about the interpolation of LSFs; it is not only the center frequencies that shift, but also the magnitude values of the peaks also change accordingly. This is due to how LSFs encode information about the spectral peaks. That is, the closer a pair, the higher the magnitude.

Figure 13.5 shows the values of the spectral envelope model parameters as a function of the interpolation factor α for both LSFs and CCs. First of all we should notice that the order of the representations is different. There are only four LSFs, while there are 50 cepstral coefficients. The cepstrum can be seen as information about the rate of change of the Fourier spectrum in different frequency bands with fixed center frequency and bandwidth. Contrary to the CCs, the LSFs do not have a fixed position in frequency, adapting their values relative to one another according to the spectral information they represent. Consequently, when we interpolate the values of the LSFs, the center frequencies shift as the values of the LSFs shift and the magnitudes of the spectral peaks vary accordingly as the LSFs get closer/farther apart.

An important requirement when interpolating LSFs is that they do not cross, otherwise the spectral information they convey would be misinterpreted and the behavior of the spectral envelope curves under interpolation of the parameters would be unpredictable. However, as long as we have

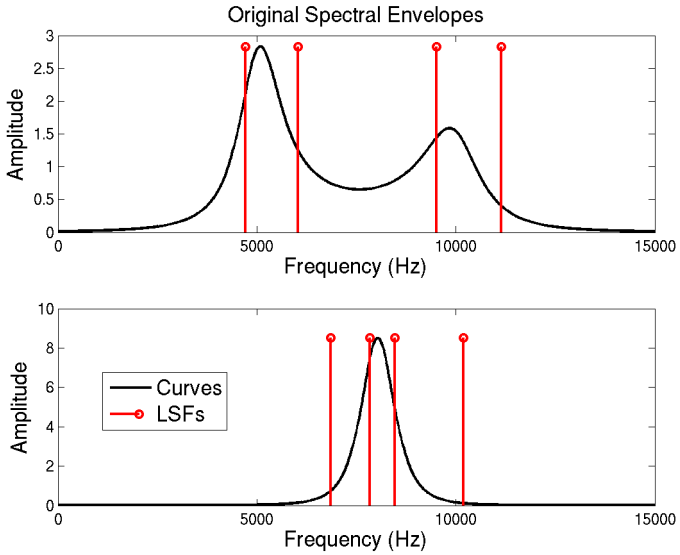

Figure 13.3: Original spectral envelope curves. The figure shows two spectral envelope curves with different number of peaks that will be used to exemplify the type of transition that we obtain (spectral peak shift or rise and wane) when we interpolate one linear prediction based and one cepstral based representation. The goal is to illustrate the behavior of the spectral peaks under the interpolation of different spectral envelope representations.

the same number of LSFs representing both spectral envelopes, their paths will never cross when interpolating them one by one in pairs.

Even though this is just an artificial example, it is indicative of the behavior of the spectral peaks under the interpolation of certain representations of spectral envelope models. Like previously stated, these are not the only possible solutions to the problem of spectral envelope morphing. We can interpolate the parameters of several different spectral envelope representations and verify that they all behave differently under the same conditions. There are two main types of spectral envelope models, namely, those based on linear prediction and those based on cepstral representations. Some representations of spectral envelopes based on linear prediction are, for example, the poles, the filter coefficients, reflection coefficients, line spectral pairs, among others. Cepstral coefficients include the real cepstrum, the complex cepstrum, mel-frequency cepstral coefficients (MFCCs), among other variants.

In general, even though the result of interpolating the parameters of each one of these representations is different, they can be grouped together according to the behavior of the spectral peaks. Linear prediction representations present a general tendency to lead to spectral peak shifting, while cepstral based representations generally result in spectral peak rise and wane. These are not the only possible spectral envelope morphing approaches. Let us not forget that there are techniques that use the spectral envelope curves directly, such as dynamic frequency warping [Pfitzinger, 2004, Ezzat et al., 2005] or simply interpolate the curves directly using equation 13.2. So, the question that remains is which spectral envelope morphing technique is most appropriate. And the answer is that it depends on the problem. We need to define what characteristics are desirable and select the spectral envelope morphing technique that presents them in most cases. For the musical instrument sound morphing problem we set out to investigate, a reasonable proposal seems to be an intermediate balance of spectral energy, measured by the spectral shape features.

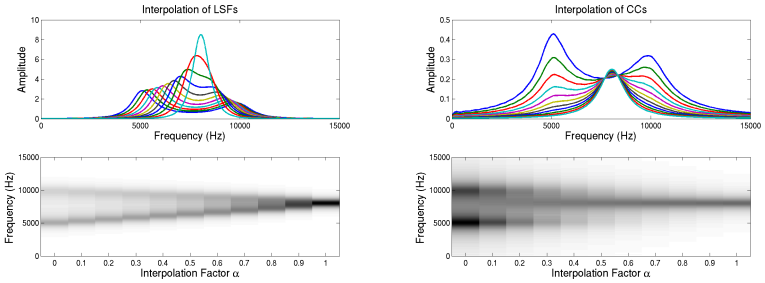

Figure 13.4: Interpolation of line spectral frequency representation.

13.1.5 Balance of Spectral Energy

Now we will consider the application of the spectral envelope morphing techniques presented in the previous section in our problem of interest, morphing musical instrument sounds. The reasoning is that we want a morphed spectral envelope curve that corresponds to a hybrid musical instrument. If we imagine that a hybrid musical instrument would probably have a resonant cavity with an intermediate shape, the morphed spectral envelope curve should reflect that. But even more importantly, the morphed spectral envelope should correspond to a perceptually intermediate sound.

According to the evaluation criteria we adopted, the morphed sounds should be not only perceptually intermediate, but the cyclostationary transformation should be perceived as smoothly as possible. One way of measuring perceptual intermediateness would be to perform a listening test with morphed sounds synthesized with a given method like Osaka [Osaka, 1998] proposes. However, this approach is very time consuming and the results are complex to evaluate because the task of estimating perceptual intermediateness related to morphed spectral envelopes involves many psychological mechanisms that are not well understood yet. An alternative would be to use surrogates of perceptual features as measures, such as the spectral shape features presented earlier, namely, spectral centroid, spread, skewness and kurtosis.

We will focus on the balance of spectral energy as measured by the spectral shape features, and rely on the correlation between their values and musical instrument sound perception. For now we will simply require that the interpolation factor α be used to control the transformation. The principle is fairly simple, we measure the values of the spectral shape features for two sounds we want to morph between, giving feature vectors $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$. Now we require that the values of the spectral shape features be interpolated according to equation 2.2, that is

$$\boldsymbol{\delta}_{1,2} = \alpha\boldsymbol{\delta}_{\mathbf{1}} + [1-\alpha]\,\boldsymbol{\delta}_2 \tag{13.3}$$

Interestingly, the feature values can also be used to evaluate the smoothness of the results when they respect equation 13.3. The criterion we adopt to evaluate both intermediateness and smoothness via the values of the spectral shape features is to calculate a deviation (squared error) between the theoretical value given by equation 13.3 and the value measured on the corresponding morphed spectral envelope curve. For example, we want the values of the spectral shape features of a smooth transformation to be as close to a straight line between $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ as possible when the morphing factor α varies linearly ($\alpha = [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$). This criterion corresponds to requiring a gradual shift of balance of distribution of energy.

Then we adopt a minimum error approach to select the smoothest spectral envelope morphing technique. This is explained in detail in chapter 14. The spectral envelope of morphed musical instrument sounds should reflect both constraints, that is, it should be intermediate and smooth

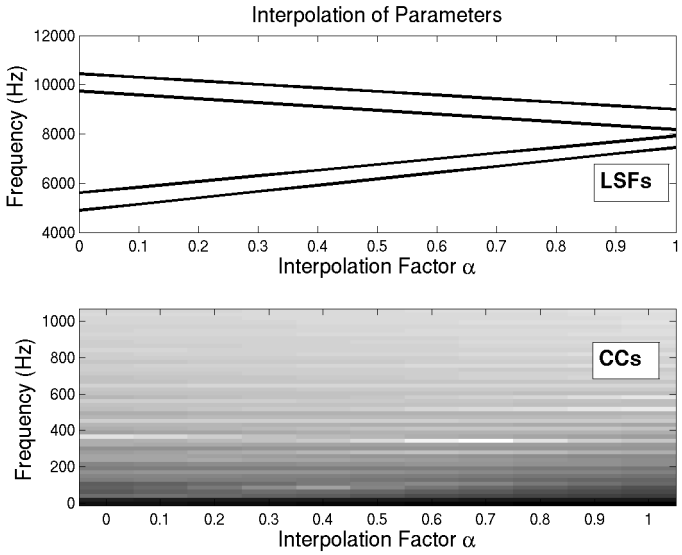

Figure 13.5: Interpolation of spectral envelope parameter values as a function of the interpolation factor α

concerning the spectral peaks and the balance of spectral energy (measured by the spectral shape features.) It is a difficult task to meet either one of these conditions isolated, let alone trying to meet them both together at the same time. Let us see why.

13.2 Target Feature Values

The spectral peak shifting and spectral peak rise and wane paradigms give rise to at least two possible solutions for the morphed spectral envelope curves. Therefore, we need to specify a way of selecting one of these according to some criteria. The balance of spectral energy comes in handy because of its correlation with the perception of sounds. In most models proposed, linear variation of interpolation parameters does not produce perceptually linear morphs [Boccardi and Drioli, 2001, Hikichi, 2001, Hope and Furlong, 1997, Slaney et al., 1996, Tellman et al., 1995]. Thus one interesting requirement to add to spectral envelope morphing is intermediate values of spectral shape features for the morphed spectral envelopes.

This requirement should restrict cases where more than one possible intermediate shape is possible. Nevertheless, we still have to be careful to add enough constraints to narrow down the possible spectral shapes to one single possibility. In other words, specifying only the spectral centroid is not enough to restrict the possible morphed spectral envelopes. Figure 13.6 illustrates the case when there are not enough restrictions to specify one single spectral envelope. If we only take the values of the spectral centroid into consideration, these three spectra would be possible candidates. But because the spectral peaks are in different absolute positions on the frequency axis they would still be perceived differently.

Naturally, when we specify that the spectra in figure 13.6 should have the same spectral centroid and the same spectral spread, then not all three spectra shown in the example satisfy the constraints. The spectral shape features are the statistical moments of the normalized magnitude spectrum, like presented in chapter 5. The question that arises naturally is about the statistical

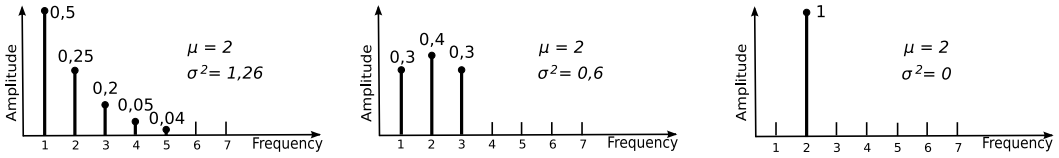

Figure 13.6: Different statistical distributions with the same mean. The figure illustrates the case when spectra have peaks in different absolute positions in frequency but have the same value of spectral centroid.

moments of a given distribution. Is there a series of moments that uniquely defines a distribution by defining its probability density function (PDF)? If there is, how many moments do we need to guarantee that the PDF is unique? Finally, how does it apply to multimodal PDFs? These questions are related to the problem of moments in statistics, which will be briefly reviewed next.

13.3 The Problem of Moments

Any statistical distribution can be characterized by the moments μ_m, which describe the nature of the distribution [Papoulis, 1991]. The problem of moments arises as the result of trying to invert the mapping that takes a measure λ to the sequences of moments μ_m, calculated as

$$\mu_m = \int_{-\infty}^{\infty} p_m(x)\, d\lambda(x) \tag{13.4}$$

for an arbitrary sequence of probability distribution functions p_m. The question appears in probability theory [Kolassa, 2006], asking whether there is a probability measure having specified mean, variance and so on, and whether it is unique.

13.3.1 Statistical Moments

The m^{th} central moment and m^{th} moment about zero of a discrete Probability Distribution Function (PDF) $p(k)$ is defined as

$$\mu_m = E\left[(X - E[X])^m\right] = \sum_k (k-\mu)^m p(k), \forall k \in S \tag{13.5}$$

where the sum is evaluated for all k in S, the sample space of the random variable X, whose Cumulative Distribution Function (CDF) defines *p(k)*. E is the expectation operator, defined as $E[X] = \sum_k kp(k) = \mu$, and μ is the expected value or mean of the PDF *p(k)*. The moments about the origin, on the other hand, are defined as

$$\mu'_m = E\left[(X)^m\right] = \sum_k k^m p(k), \forall k \in S \tag{13.6}$$

and the moment is said to exist if the series is absolutely convergent [Cramér, 1945].

The moment of order zero, or zeroth moment, is simply the sum of *p(k)* for all k, which, for a PDF is always 1. The first moment about the zero is the mean, μ. The second central moment is the variance σ^2, whose square root gives the standard deviation σ of the PDF. Sometimes it is convenient to convert moments about the origin to moments about the mean. The general

equation for converting the m^{th}moment about the origin to the m^{th} moment about the mean is [Papoulis, 1991]

$$\mu_m = \sum_{j=0}^{m} \binom{m}{j} \mu'_j \left(-\mu\right)^{m-j} \tag{13.7}$$

Similarly

$$\mu'_m = \sum_{j=0}^{m} \binom{m}{j} \mu_j \mu^{m-j} \tag{13.8}$$

13.3.2 Standardized Moments

The m^{th} standardized moment of a PDF is defined as [Cramér, 1945]

$$\frac{\mu_m}{\sigma^m} \tag{13.9}$$

where μ_m is the m^{th} central moment and σ is the standard deviation of the PDF.

- The first standardized moment is zero, because the first moment about the mean is zero
- The second standardized moment is one, because the second moment about the mean is equal to the variance (the square of the standard deviation)
- The third standardized moment is the skewness
- The fourth standardized moment is the kurtosis

Note that for skewness and kurtosis alternative definitions exist, which are based on the third and fourth cumulants respectively.

13.3.3 Characteristic Function

In probability theory and statistics, the characteristic function of any random variable completely defines its probability distribution [Cramér, 1945]. Thus it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions. The characteristic function always exists when treated as a function of a real-valued argument, unlike the moment-generating function [Papoulis, 1991]. There are relations between the behavior of the characteristic function of a distribution and properties of the distribution, such as the existence of moments and the existence of a density function [Cramér, 1945]. The characteristic function of a PDF *p(k)* is the discrete Fourier transform of *p(k)*, defined as

$$\varphi\left(\xi\right) = \sum_{k=0}^{N-1} p\left(k\right) e^{jk\xi} \tag{13.10}$$

the series being absolutely and uniformly convergent for all ξ, since $\sum_k p\left(k\right) = 1$ [Cramér, 1945]. Each term of the series is a periodic function of ξ. Not every function $\varphi\left(\xi\right)$ may be the characteristic function of a distribution. Various necessary and sufficient conditions are known [Cramér, 1945], but they will not be discussed at length here.

One important property of the characteristic function is that, if the moment of order m of *p(k)* exists, equation (13.10) can be differentiated m times with respect to ξ, giving [Cramér, 1945]

$$\frac{d^m}{d\xi^m}\varphi\left(\xi\right) = \varphi^{(m)}\left(\xi\right) = j^m \sum_{k=0}^{N-1} k^m p\left(k\right) e^{jk\xi} \tag{13.11}$$

Furthermore, $\varphi^{(m)}(\xi)$ is continuous for all $\xi \in \mathbb{R}$ [Cramér, 1945], such that

$$\varphi^{(m)}(0) = j^m \sum_{k=0}^{N-1} k^m p(k) = j^m \mu_m \tag{13.12}$$

In the neighborhood of $\xi = 0$ there is thus a development in MacLaurin series [Cramér, 1945]

$$\varphi(\xi) = \sum_{m=0}^{\infty} \frac{\varphi^{(m)}(0)}{m!} (\xi)^m = \sum_{m=0}^{\infty} \frac{\mu'_m}{m!} (j\xi)^m = 1 + \sum_{m=1}^{M} \frac{\mu'_m}{m!} (j\xi)^m + \mathcal{O}(\xi^m) \tag{13.13}$$

where the error term $\mathcal{O}(\xi^m)$ divided by ξ^m tends to zero as $\xi \to 0$ [Cramér, 1945]. An alternative expression for the error term, given by

$$\mathcal{O}(\xi^m) = \int_0^{\xi} \varphi^{(m+1)}(0) \frac{(\xi - u)^m}{m!} du = \varphi^{(m+1)}(\Xi) \frac{\xi^{m+1}}{(m+1)!}; 0 < \Xi < \xi \tag{13.14}$$

If $\lim_{m \longrightarrow \infty} \mathcal{O}(\xi^m) = 0$ the series converges and $\varphi(\xi)$ is analytical. An alternative derivation can be obtained by expanding the complex exponential in equation (13.10) in MacLauren series, obtaining

$$e^{jk\xi} = \sum_{m=0}^{\infty} \frac{(jk\xi)^m}{m!} \tag{13.15}$$

and then substituting this expression into (13.10), getting

$$\varphi(\xi) = \sum_{k=0}^{N-1} p(k) \sum_{m=0}^{\infty} \frac{(jk\xi)^m}{m!} = \sum_{m=0}^{\infty} \frac{(j\xi)^m}{m!} \sum_{k=0}^{N-1} p(k) k^m = \sum_{m=0}^{\infty} \frac{\mu'_m}{m!} (j\xi)^m \tag{13.16}$$

It is important to know whether a distribution is uniquely defined by the sequence of its moments. This is known as the *moment problem*. It can be shown [Cramér, 1945] that the characteristic function is uniquely determined by the sequence of moments μ_m if the series converges absolutely near the origin. The moments of a distribution are not arbitrary numbers and must satisfy various conditions [Papoulis, 1991], such that only certain sequences represent distributions. The relationship between the characteristic function of a PDF and its moments allows us to write an analytical formula that establishes the connection between the cepstral coefficients and the spectral shape features of a given spectral envelope curve.

13.3.4 Analytical Formulation

Firstly, we observe that the inverse Fourier transform of equation (5.3) can be interpreted as the characteristic function of *p(k)* by inspecting equation (13.10) with $\xi = \omega = \frac{2\pi n}{N}; n = 0...N-1$, and thus it can be expressed as

$$\hat{x}(n) = \sum_{k=0}^{N-1} p(k) e^{jk\omega} = \sum_{m=0}^{\infty} \frac{\mu'_m}{m!} \left(j \frac{2\pi}{N} n \right)^m \tag{13.17}$$

where μ'_m represents the m^{th} order moment about the origin of *p(k)*. Upon closer examination, the real cepstrum can also be expressed in terms of a series of moments of the log magnitude spectrum

if we view the cepstrum as the characteristic function of $g(k)$. So we obtain equation (13.18) simply by substituting $p(k)$ in equation (13.10) with equation 5.3 in chapter 5, we get

$$c(n) = \sum_{k=0}^{N-1} g(k) e^{jk\omega} = \sum_{m=0}^{\infty} \frac{\mu_m^{'}}{m!} \left(j\frac{2\pi}{N} n \right)^m \tag{13.18}$$

where $\mu_m^{'}$ now represents the m^{th} moment about the origin of $g(k) = \log|X(k)|$, and *c(n)* is the cepstrum of the sequence *x(n)*. This equation establishes the analytical relation between the real cepstrum, a model of spectral envelope, and the statistical moments of the spectral envelope it defines, directly related to the spectral shape features we desire to control in transformations. The moments $\mu_m^{'}$ are directly linked to the spectral shape features by equations (13.8) and (13.9). Equation (13.18) can be viewed as the z-transform of the series of moments by substituting $z^{-1} = j\omega$. In words, the cepstral coefficients can be expressed as the z-transform of the series expansion of the statistical moments (around zero) of the log-magnitude spectrum evaluated on the imaginary axis. Schröeder [Schröeder, 1999] suggests a similar correspondence when deriving nonrecursive relations between the linear prediction coefficients and the cepstral coefficients of a sequence $x(n)$.

If we want to guarantee that we have enough constraints to narrow down the possible morphed spectral envelopes, we must add as many constraints as the number of dimensions we have in the original space the spectral envelopes reside. In other words, if we need M parameters (cepstral coefficients, lienear prediction coefficients, line spectral frequencies, etc) to specify the spectral envelope in the parameter space, we need to add at least M linearly independent constraints to uniquely specify the same spectral envelope in a different space. Supposing that the space we wish to project the spectral envelopes onto is the space of moments of probability distributions (which correspond to the spectral shape features), then we would need M moments to uniquely specify one spectral envelope.

The analytical formulation expressed in equation 13.18 theoretically allows us to obtain the cepstral coefficients corresponding to a given sequence of values of spectral shape features, given the condition that the series of moments is convergent. Nevertheless, spectral envelope curves are usually multimodal (they present more than one peak), while probability density functions (PDF) are usually unimodal (they only have one peak).

13.4 Manipulation of Spectral Envelope Representations

I derived the analytical relationship between the real cepstrum representation of a spectral envelope curve and its spectral shape features in the hopes of being able to perform spectral envelope transformations in the space of spectral shape features directly, and then retrieve the spectral envelope (curve or parameters) that corresponds to the specified values of spectral shape features. However, when this approach seemed to bear no fruit, I decided to adopt an error minimization approach instead.

Under the intermediateness and smoothness requirements of morphing algorithms, we want to select the spectral envelope morphing technique that gives minimal quadratic error according to the variation of the values of spectral shape features. The criterion adopted is that a linear variation of the morphing factor α should lead to a linear variation of the values of spectral shape features when we consider intermediateness and smoothness. The derivation of the quadratic error measure and the complete evaluation procedure will be presented in chapter 14, but right now we will see the motivation for the error minimization approach.

13.4.1 Interpolation of Spectral Envelope Representations

Figure 13.7 shows a comparison between many spectral envelope morphing techniques. On the left-hand side we see the morphed spectral envelope curves corresponding to $\alpha = [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$, and on the right-hand side we see the corresponding variation of the spectral shape features measured on the corresponding curves on the left. The solid lines correspond to the original spectral envelope curves for which $\alpha = 0$ or $\alpha = 1$.

Each row in figure 13.7 corresponds to a different spectral envelope morphing method, namely, interpolation of spectral envelope curves (ENV), interpolation of linear prediction coefficients (LPC), interpolation of cepstral coefficients (CC), interpolation of reflection coefficients (RC), interpolation of line spectral frequencies (LSF), and dynamic frequency warping (DFW). We should notice that there are spectral envelope morphing methods based on the spectral envelope curve (ENV and DFW), cepstral based representations (CC) and linear prediction based representations (LPC, RC, and LSF).

Among these methods, we find spectral peak shift (LPC, RC, and LSF) and spectral peak rise and wane (ENV, CC, and DFW) behavior. The spectral envelope curve was estimated using true envelope for all the spectral envelope morphing methods and then converted to LPC using equation 7.129 in section 7.7.2 of chapter 7. We want to investigate which of the spectral envelope morphing methods gives a variation of all spectral shape features as close as possible to linear when $\alpha = [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$. We should bear in mind that some of the spectral envelope morphing methods shown in figure 13.7 correspond to the main spectral morphing methods proposed in the literature. These are the envelope curve (ENV) [Tellman et al., 1995, Fitz et al., 2003, Osaka, 1995], linear prediction coefficients (LPC) [Moorer, 1978], cepstral coefficients (CC) [Slaney et al., 1996] and dynamic frequency warping (DFW) [Ezzat et al., 2005, Pfitzinger, 2004]. We will delve deeper into the evaluation procedure in the next chapter, devoted to answering this question.

The apparent difference between the original spectral envelope curves (solid lines) of linear prediction (LPC, RC, and LSF) and cepstral based (ENV, CC, and DFW) representations in figure 13.7 is due to the cepstral based estimation (true envelope) and posterior conversion to linear prediction coefficients. The conversion from cepstral to linear prediction based spectral envelope representation introduces distortions in the converted spectral envelope curve because the conversion method is not exact. The conversion from LPC to RC and LSF, on the other hand, is exact.

In the context of voice conversion, Villavicencio et al. [Villavicencio et al., 2006] showed that estimating the spectral envelope with "true envelope" and converting to LPC yields a better spectral envelope than simply estimating the spectral envelope directly with linear prediction. The rest of this chapter is dedicated to investigating the conversion between CC obtained with the "true envelope" estimation, and LPC representation. This conversion is referred to as TENV2LPC and it is important because the resynthesis step of the SF model uses the LPC representation, as explained in chapter 11.

13.4.2 Spectral Envelope Model Conversion

Spectral envelope model conversion can be considered as a spectral envelope manipulation technique whose main requirement is to preserve the spectral envelope curve as accurately as possible. Naturally there have been proposals to measure spectral distortion or spectral distance [Itakura and Saito, 1968]. We will analyze the result of conversion techniques using the Itakura-Saito distance, which is a well known measure of the perceptual difference between a given magnitude spectrum $|H(\omega)|$ and an approximation $\left|\widetilde{H}(\omega)\right|$. The Itakura-Saito distance will be used to evaluate the distortion introduced in the spectral envelope curve when the conversion between

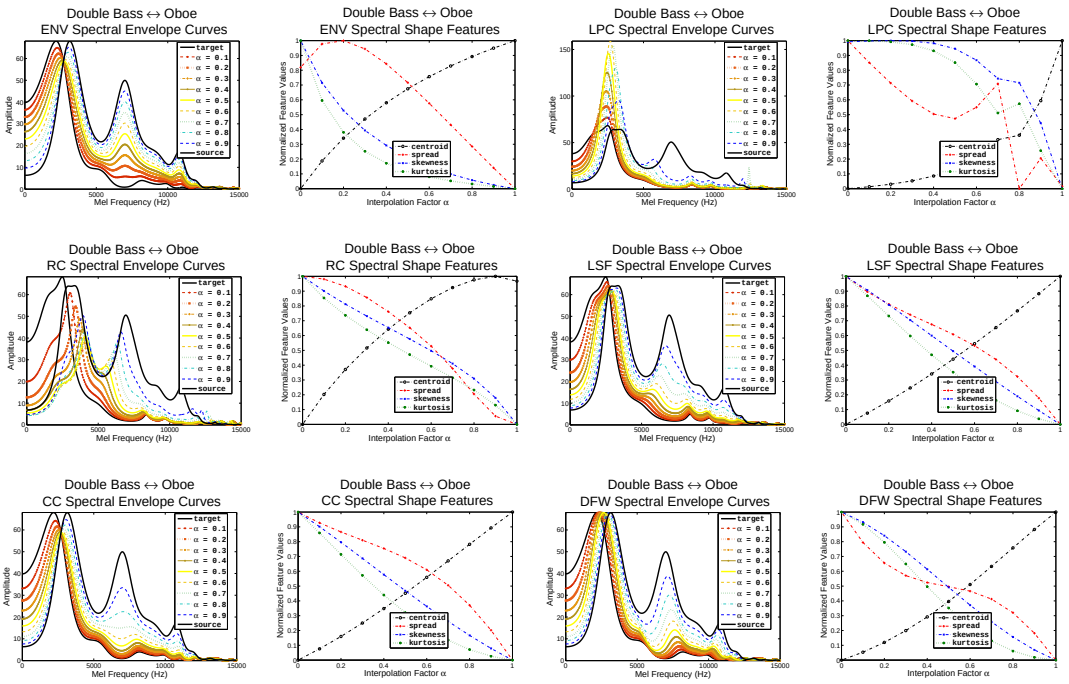

Figure 13.7: Spectral envelope morphing guided by spectral shape features. The figure shows the variation of the values of spectral shape features when morphing spectral envelopes using the main approaches proposed in the literature. The spectral envelope curves are shown on the left and the corresponding feature variation on the right. We want the spectral envelope morphing algorithm that leads to linear variation of spectral shape features.

spectral envelope representations is not exact, as is the case for the conversion between linear prediction coefficients (LPCs) and cepstral coefficients (CCs). The conversion operation allows us to use one spectral envelope estimation technique, while giving us the flexibility of manipulating the spectral envelopes using a different representation. So we can use a spectral envelope estimation technique that is optimal in some sense for the estimation problem, and then manipulate the spectral envelopes using another representation that is more appropriate than the one used in the estimation.

In the context of voice conversion, Villavicencio [Villavicencio et al., 2006] proposes to estimate the spectral envelope curves using true envelope [Röbel et al., 2007] and converting the result into LPCs to improve the quality of the voice conversion task. The idea is that true envelope gives an accurate estimation of the spectral envelope curve when we compare the distance between the peaks of the spectrum and the spectral envelope curve estimation at those points. Villavicencio compares the accuracy of true envelope estimation with other techniques (among them is LPC) and concludes that we obtain a more accurate spectral envelope curve when we estimate it using "true envelope" and convert the cepstral coefficients of the "true envelope" estimation to LPC, than when we estimate the spectral envelope curve using LPC directly. Let us perform a similar evaluation for the spectrum of musical instrument sounds. Figure 13.8 shows the spectrum of different musical instrument sounds and the spectral envelope curves superposed to it. The spectral envelope curves that we see were obtained by true envelope estimation directly (TENV), by linear prediction directly (LPC), and the result of the conversion of TENV into a linear prediction based

representation (TENV2LPC).

Supposing that true envelope gives the most accurate spectral envelope curve, we measure the Itakura-Saito distance TENV-TENV2LPC and TENV-LPC and compare them. The smallest IS distance value gives an accurate spectral envelope curve represented as LPC. We can see from the figure that the IS distance TENV-TENV2LPC is at least one order of magnitude smaller than TENV-LPC for all cases. Naturally, when we evaluate the conversion from CC to LPC, we are implicitly supposing that we will not use the CC representation obtained directly from the true envelope estimation. It is always the case in this work because the resyntheis of the SF model always uses the LPC representation of the filter. Therefore, we obtain more accurate results in the implementation of the SF model developed in this thesis when we estimate the spectral envelope with "true envelope" and convert it later to LPC. Chapter 14 will present the result of a listening test that verifies whether this statement is true perceptually.

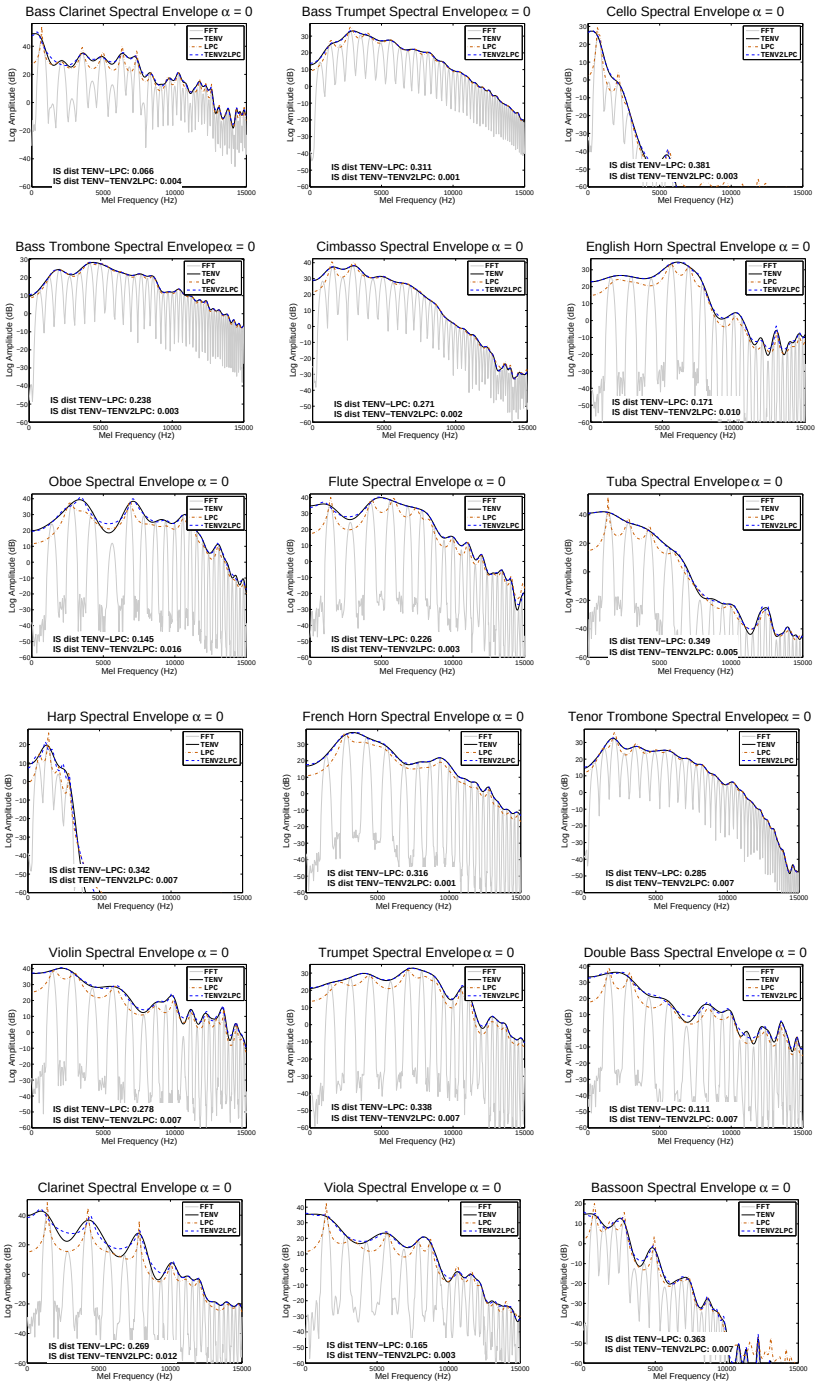

Figure 13.8: Comparison of the accuracy of spectral envelope representation in the linear prediction domain. The figure compares the spectral envelope curves estimated with "true envelope" (TENV) and (LPC), and the result of the conversion from TE to LPC (TENV2LPC). In the figure we see the Itakura-Saito (IS) distance between TENV - LPC and between TENV - TENV2LPC. The smallest IS distance indicates the most accurate representation in the linear prediction domain.

Chapter 14

Evaluation

In this chapter, we will see the results of the evaluation of the model and of the transformations we can obtain with it. The evaluation of the model consists in verifying that the sounds represented by the implementation of the source-filter (SF) model developed in this work actually sound the same as the original recordings. The morphing transformations are evaluated according to the criteria adopted, correspondence, intermediateness and smoothness, using objective and subjective (perceptual) criteria.

The main purpose of this chapter is to allow a comparison between the traditional sinusoidal model and the source-filter (SF) representation of isolated musical instrument sounds in terms of the transformations. We want to be able to control the transformation with the morphing factor α such that, when α varies linearly, the transformation is also linear. First and foremost, I will show the result of the evaluation of the SF representation of isolated acoustic musical instrument sounds I proposed. This first test, presented next in section 14.1, consists in evaluating whether sounds resynthesized with parameters derived from the SF representation are perceptually different from the original recordings. This task consisted in a simple listening test asking the subjects how perceptually similar the original recording and the same sound synthesized from the SF representation were.

Next, we will see the evaluation procedure concerning the transformation. This task concentrates on spectral aspects of the transformation for the sake of simplicity. Since evaluating the transformation is a much more complex task than the SF model evaluation, we need objective criteria and subjective perceptual tests for this task. The objective criteria adopted are related to the behavior of the spectral peaks and of the values of the features under the transformation. The behavior of the spectral peaks is just qualitatively analyzed in this step, permitting the separation of the spectral envelope representations into two distinct classes, rise and wane (RW) and peak shifting (PS). The values of the spectral shape features, on the other hand, are quantitatively evaluated. With the principles of intermediateness and smoothness in mind, we want the spectral shape features to vary linearly when the morphing factor α changes linearly. The error metric measures the deviation between the ideal interpolated values and the measured values of the spectral shape features for each spectral envelope representation studied. In this work, we want the method that gives minimum error.

Finally, the method with minimum average error was compared with the popular interpolation of parameters of sinusoidal models in a listening test. The aim of the listening test was to evaluate the perceptual linearity of the morphing transformation under both methods.

14.1 Evaluation of the Source-Filter Model

Since we know that the implementation of the source-filter (SF) model proposed in this work does not represent accurately the spectral information (mostly due to errors in the spectral envelope estimation and conversion steps), the first important step in the evaluation procedure is the perceptual validation of the source-filter representation. In other words, we need to investigate if the sounds represented with the SF model are perceptually different from the original recordings (and if they are, by how much). In general, the sinusoidal model provides a good representation of the sound on the perceptual sphere. However, the residual part is usually not represented in sinusoidal modeling.

We already know from section 7.6 that the spectral envelope estimation with "true envelope" presents small errors in the representation of the amplitudes of the spectral peaks corresponding to the partials. If we were to do a simple test and substitute the amplitudes of the partials in the traditional sinusoidal representation of a sound with those given by the "true envelope" estimation and resynthesize the sound with these small variations, would the resynthesized sound be perceptually identical to the original? If not, how perceptually similar are they? When we include the residual modeling, the question becomes even more intriguing. Is the white noise filtered with the LPC estimation of the spectral envelope of the residual from a sinusoidal analysis of a sound a good perceptual representation of the residual?

Naturally, it is the result of the combination sinusoidal plus residual components heard together that is of interest in this work. Thus we will see the result of a simple listening test that aimed to evaluate how perceptually similar the sounds resynthesized from the SF representation are to the original recordings. The quality of the morphing transformation is directly linked to the perceptual validation of the SF representation. If the the original sounds represented by the SF model are perceptually very different, then we could argue that we are not morphing between the original sounds on the perceptual plane when we use the SF model. Rather, we are morphing between the SF perceptual representation of the original sounds.

The listening test presented 20 pairs of sounds and asked the participant to rate the perceptual similarity between them. There were 5 possible choices: identical, slightly different, fairly different, significantly different, and very different. These verbal labels were adapted from the comparison category rating (CCR) test [ITU-T Recommendation P.800, 1996] (CMOS). Table 14.1 shows the numerical values associated with each verbal label to compare the similarity assessments quantitatively.

Participants were asked to listen to all the sounds once before starting the test to get used to the range of differences across all pairs. Most pairs contained the original recording on the left and its SF representation on the right. However, 6 identical pairs were used among the 20 pairs presented. The participants were not informed of the presence of identical pairs. The listening test is available online `http://recherche.ircam.fr/anasyn/caetano/survey/similarity.html`. Appendix F has the instructions used in the test. After taking the test, the participants were asked whether they used headphones, whether they listened once to all sounds, and whether they were experienced in music or audio evaluation. The results of participants who answered "no" to any of the questions were not used. This does not guarantee uniformity because we don't control the experimental setup, but at least we can make sure that the results of participants who didn't follow the instructions are not included. In total, the results of 80 participants aged between 22 and 67 were used.

The identical pairs are important for several reasons, namely, to verify the accuracy of the similarity judgments, to help calibrate the scale, and to validate the results. First of all, we learn a lot about how reliable and accurate certain participants' results are simply by confirming whether they rated an identical pair as perceptually identical.

Identical pairs also help calibrate the perceptual scale. When the identical pairs are present, any other pairs that were also judged identical by individuals have the same status. Naturally, the

identical	slightly different	fairly different	significantly different	very different
5	4	3	2	1

Table 14.1: Numerical scale for similarity test. The table shows the values associated with each label of the scale.

presence of identical pairs challenges the participants' ability to detect barely noticeable differences. A consequence is that some participants tend to rate pairs that present barely noticeable differences as "slightly different", which makes the results more reliable.

The other end of the scale is free, though. This means that the most dissimilar pair might become the reference for some listeners and be consistently rated very differently from the other sounds. Other listeners, however, will use another reference. The introduction of a systematically degraded version of the sounds (such as low bit encoded mp3) as reference may help make the results more uniform. However, it would also warp the scale for some users to accommodate the degraded version.

When the degraded version is definitely very different from the original recording, it might also have the side effect of artificially raising the mean evaluation of the most dissimilar pair. For this reason, a degraded version was not included. Finally, the verbal labels might also be used as reference. In this case, some people might associate them to previous exposures to sound recordings, especially when quality was important.

Figure 14.1 shows the results of the listening test used to evaluate the SF model representation. In figure 14.1 we see the ratings for individual sounds, together with a global average representing the SF model as a whole. The labels in figure 14.1 can be found in table 14.2. First of all we notice in figure 14.1 that the identical pairs were identified as such for most cases. There is an interesting difference in assessment between the two identical pairs presented right at first and the others. In average, the two first identical pairs were found to be a little more different than the others. This is probably due to the lack of previous context. The perceptual scale tends to be adjusted during the test as we are exposed to the stimuli.

In general, the implementation of SF model used in this work was rated between 'slightly different' and 'fairly different'. Except for the bass trumpet sound, which was very "brassy". The global average only includes the 14 pairs with the model (it does not include the assessment of the identical pairs). The global average was almost four, which corresponds to slightly different. Thus the results of the similarity test validate the SF model as perceptually similar to the original sounds.

14.2 Evaluation of the Transformation

The final goal of any sound morphing algorithm is to allow control of the transformation with the morphing factor α. We usually want a morphing algorithm to produce linear morphs when α varies linearly. Ideally, we would like to have control of perceptual features of the sounds being morphed by simple manipulation of the morphing factor, such that a morphing factor of $\alpha = 0.5$ would produce a morphed sound perceptually halfway between source and target. This calls for a measure of perceptual intermediateness. When we extrapolate this condition and require that the linear variation of the morphing factor α should lead to a perceptually linear transformation, we are imposing the requirement of perceptual smoothness.

Therefore, the evaluation of the transformation aims at intermediateness and smoothness of the morphing algorithm. Both criteria considered together correspond to linearity. Linearity will be evaluated both objectively and subjectively. The objective evaluation requires that the spectral shape features vary linearly. The subjective evaluation consisted in a listening test.

0	Identical
Ce	Clarinet
BT	Bass Trumpet
Ba	Bassoon
Ci	Cimbasso
CT	Contrabass Tuba
Fl	Flute
Vi	Viola
DB	Double Bass
Ob	Oboe
FH	French Horn
Tr	Trumpet
Vl	Violin
BC	Bass Clarinet
Tu	Tuba
G	Global

Table 14.2: Sounds used in the listening test.

Since it is extremely difficult to objectively evaluate the perceptual impact of a morphing procedure, I proposed to use perceptually related features instead. Now the morphing transformation can be quantitatively evaluated using the values of the features. If the features capture perceptually relevant information, their values should reflect the perceptual impact of the morphing procedure. One important question remains, what features are necessary to represent enough perceptually relevant information about a sound such that the inspection of the behavior of the feature values alone would point us toward the perceptual impact? Since the answer to this question is out of the scope of this work, I will present a cross-evaluation method that uses quantitative and qualitative objective criteria together with a listening test to evaluate the results.

The evaluation of the transformation takes into consideration the spectral and temporal envelope morphing procedures. The temporal alignment procedure is not included in the evaluation because it is considered preprocessing and is totally independent of the subsequent morphing steps. In this section we concentrate on the evaluation of the spectral envelope morphing. The evaluation of the temporal envelope morphing procedure will be seen later on in this chapter because it is analogous.

14.2.1 Objective Evaluation

This section focuses on the evaluation of the spectral envelope morphing procedure. The objective evaluation comprises a qualitative and a quantitative analysis. The qualitative analysis consists in verifying the behavior of the spectral envelope peaks when interpolating different spectral envelope representations. On the other hand, the quantitative analysis uses the variation of the values of the spectral shape features when morphing spectral envelopes.

In accordance with the theory of timbre perception, the position of the formant peaks is perceptually relevant. As such, it is important to consider the behavior of the formant peaks of the spectral envelope morphing techniques we are studying. As presented in chapter 13, the formant peaks can present two distinct behaviors when the parameters of the spectral envelope representation are interpolated, spectral peak shifting (PS) and spectral peak rise and wane (RW). In this work, we postulate that the spectral peaks of hybrid musical instruments should be in intermediate positions between those of the sounds being morphed. This hypothesis means that we favor the

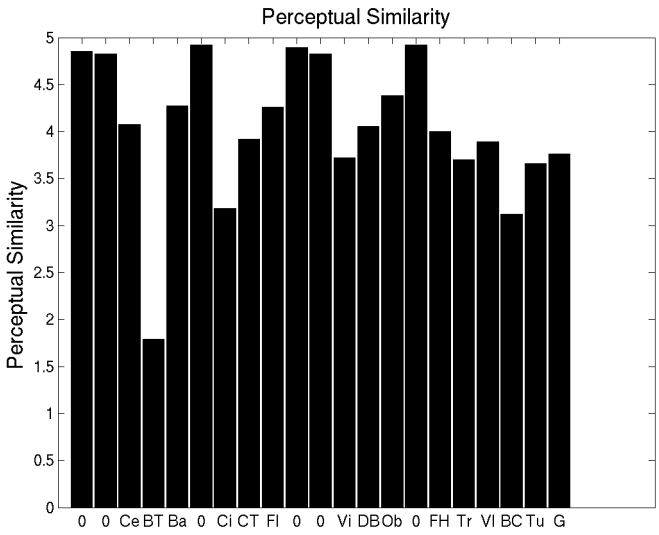

Figure 14.1: Perceptual similarity.

spectral peak shifting (PS) transition rather than the spectral peak rise and wane (RW). Therefore, the qualitative evaluation consists in classifying the spectral envelope morphing techniques studied into two groups, labeled PS and RW.

The variation of the values of the spectral shape features when the parameters of a given spectral envelope representation are interpolated is an important aspect of the morph. When the features used to guide the transformation capture perceptually relevant information, sounds whose features are intermediate should be perceived as intermediate regarding those features. The spectral shape features are a measure of the balance of the distribution of spectral energy. Chapter 5 explained the correlation between the distribution of spectral energy and the perception of timbre. In this work, the values of the spectral shape features are considered as an objective measure of the perceptual impact of the morphing transformation. Thus, under the assumption that the spectral shape features we chose to monitor (namely spectral centroid, spread, skewness and kurtosis) are correlated to salient dimensions of timbre perception [McAdams et al., 2005], we want the feature values to vary as close to linearly as possible when the interpolation factor is varied in equal steps (i.e, linearly) to guarantee that the morphing factor α controls perceptually relevant aspects of the morphing transformation.

The criterion of linearity stems from the intermediateness and smoothness requirements discussed earlier. The linearity criterion adopted allows us to introduce an error metric that objectively measures the deviation between the values of the transition considered ideal and the actual values measured for each spectral envelope representation. In this work, under the constraints of intermediateness and smoothness adopted, we postulate that the ideal transition corresponds to fitting a straight line between the values of the spectral shape features of the sounds used in the morph. Then, the error measure adopted is simply the distance between the ideal interpolated value and the value measured for each spectral envelope morphing method. A spectral envelope morphing method corresponds to interpolating between the parameters of a particular representation of a spectral envelope model. Therefore, we will study the spectral shape feature interpolation properties of the spectral envelope representations to investigate if there is one representation that consistently gives a small error when compared to the others.

14.2.1.1 Qualitative Analysis: Spectral Peaks

In this section, we will see some figures with examples of spectral envelope morphing techniques applied to different musical instrument sounds from the Vienna sound database. Each figure will present morphed spectral envelopes using several techniques for one specific pair of sounds on the left, and the corresponding variation of the values of the spectral shape features on the right. The morphing factor varies linearly $\alpha = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. The spectral envelopes of the sounds used in the morph are shown as solid lines and correspond to $\alpha = 0$ and $\alpha = 1$. The nine intermediate curves shown correspond to the intermediate values of α.

In this section, the aim of the figures is to qualitatively analyze the behavior of the spectral envelope peaks and the variation of the values of the spectral shape features for each representation. We can classify the representations into the peak-shift (PS) or rise and wane (RW) paradigms according to the behavior of the formant peaks. We can qualitatively evaluate if the spectral shape features present a tendency to vary linearly or not.

The behavior of the formant peaks under the spectral envelope morphing operation is intrinsically dependent on the way a particular spectral envelope representation encodes information. For example, cepstral coefficients represent the oscillations of the spectral envelope curve at different frequencies and it is the combination of the values of the coefficients representing the amplitude of oscillation for each frequency that results in the spectral envelope peaks and valleys. Changing the value of one coefficient will most likely have the effect of increasing or decreasing the amplitude of peaks or valleys, but it will definitely not shift them in frequency. Thus, when interpolating the values of cepstral coefficients, we would expect the peaks to behave under the rise and wane (RW) fashion rather than peak shift (PS).

LSFs, on the other hand, present a tendency to encode information directly about each spectral envelope peak. Each pair of LSFs usually represents the absolute position of the peaks (or valleys) of the spectral envelope in frequency and the distance between each line spectral pair (LSP) is proportional to the amplitude of the peaks (or valleys). Therefore, changing the value of one pair of LSFs equally usually shifts a specific spectral peak (or valley) in frequency and changing the distance between a pair of LSFs increases the amplitude of a peak (or decreases the depth of a valley). So, we can expect the interpolation of LSFs to lead to spectral peak shift (PS) rather than rise and wane (RW).

Figures 14.2 to 14.9 compare spectral envelope morphing techniques corresponding to several proposals found in the literature. The techniques are interpolation of the spectral envelope curve (ENV) [Osaka, 1995, Tellman et al., 1995, Hatch, 2004, Fitz and Haken, 1996, Fitz et al., 2003], interpolation of linear prediction coefficients (LPC) [Moorer, 1978], interpolation of reflection coefficients (RC) [Moorer, 1978], interpolation of line spectral frequencies (LSF) [Itakura, 1975, Itakura and Saito, 1970, Caetano and Rodet, 2011b], interpolation of cepstral coefficients (CC) [Slaney et al., 1996], and dynamic frequency warping (DFW) [Ezzat et al., 2005, Pfitzinger, 2004]. Notice that the linear prediction based techniques LPC, LSF, and RC are presented at the bottom row of figures 14.2 to 14.9, while the techniques that use the spectral envelope curve obtained directly from the "true envelope" estimation CC, DFW, and ENV are shown at the top. The original envelope curves for LPC, LSF, and RC are slightly different from their counterparts at the top because of the conversion from cepstral base to linear prediction based representation. Notice that the conversion between linear prediction based representations (LPC to LSF and LPC to RC) does not present distortions. Finally, interpolating the amplitudes of partials in sinusoidal models corresponds to interpolating the spectral envelope curve (ENV) in this case. Therefore, the conclusions for ENV will be extrapolated to sinusoidal models in general. Let us evaluate each figure in turn.

Figure 14.2 shows that CC and ENV produce morphed envelope curves that fit the RW paradigm, while DFW clearly shifts the spectral envelope peak (PS). LPC and RC do not give

Peak Shift (PS)	Rise and Wane (RW)
DFW, RC, LPC, and LSF	ENV and CC

Table 14.3: Behavior of spectral envelope peaks. The table shows which paradigm (PS or RW) fits best each spectral envelope morphing technique.

envelope curves that gradually change. In this case, LSF produces gradually changing curves that fit the PS paradigm. Except for kurtosis, the spectral shape features present a general tendency to vary rather linearly.

Figure 14.3 shows a challenging example. Once again, CC and ENV present a typical RW behavior, while DFW, LPC, LSF and RC present PS behavior. In this case the spectral shape features present a tendency to vary linearly for all representations.

Figure 14.4 shows a convincing example of the RW behavior of CC and ENV on the one hand, and PS behavior for DFW, LSF, and RC on the other hand. Again, LPC produced morphed envelope curves with unexpected shapes. Here again the spectral kurtosis is the only feature to present an odd behavior.

Figure 14.5 confirms once again the RW behavior of CC and ENV, and PS behavior for DFW, LSF and RC. It is interesting to examine figure 14.5 more carefully, though. We can see an interesting difference between the PS behavior of DFW and LSF in figure 14.5 when we examine which peaks are matched. DFW and LSF give very different morphed spectral envelopes for this particular case. Apart from LPC, most methods present a general tendency to give linear variation of spectral shape features.

Figure 14.6 shows a case where only DFW presents a behavior that varies radically from the other methods. In this particular case, the RW behavior of CC and ENV does not contrast very much from the result we get with LPC, LSF, or RC when compared to the previous examples. Interestingly though, this is also reflected in the variation of the spectral shape features. CC and ENV tend to behave rather nonlinearly in this case.

Figure 14.7 shows clearly the contrast between the RW behavior of CC and ENV and the PS behavior of DFW and LSF. Upon close inspection though, we see again a different peak matchin between DFW and LSF. The variation of the spectral shape features reflects the difference between the DFW and LSF morphed spectral envelope curves.

Figure 14.8 confirms the RW behavior of CC and ENV, and PS of DFW, LPC, LSF, and RC. Again the peak matching differs between DFW and LPC, LSF, and RC, but this does not seem to affect much the variation of the spectral shape descriptors. Finally, figure 14.9 presents an interesting example where all representations give similar results both in terms of behavior of spectral peaks and variation of spectral shape descriptors.

From the figures 14.2 to 14.9, we can broadly classify the representations according to the general behavior of the spectral peaks, as can be seen in table 14.3, which shows that DFW, RC, LPC, and LSF present PS behavior, and ENV and CC present RW behavior. It is not evident to decide which transition would sound "more natural" for morphed musical instrument sounds. The variation of the values of spectral shape features will be used to help decide. In this work, we are looking for the representation that gives linear variation of the spectral shape descriptors.

14.2.1.2 Quantitative Analysis: Spectral Shape Feature Values

The main objective in this section is to use the feature values to guide the transformation. The principles of intermediateness and smoothness dictate that the values of the features should vary linearly when the morphing factor α varies linearly. The requirement of linearity led to the adoption of a simple objective error measure that uses the quadratic deviation between the measured feature

values and the ideal interpolated ones, as illustrated in Figure 14.10. The requirement of linearity allows us to investigate which representation of the spectral envelope leads to the smallest error when interpolated, which is considered the closest to linear.

Figure 14.10 shows the error calculation applied to each spectral frame. For each spectral shape feature δ used as guide, we obtain a straight line that connects the values of that particular feature for the source and target sounds, represented as a capital "X" for $\alpha = 0$ and $\alpha = 1$ in figure 14.10.

Next we calculate the ideal interpolated values of the feature for each value of the morphing factor α considered using the straight line as linear regression. These are represented as small case "x" in figure 14.10. Then we calculate the values of the features for the spectral envelopes obtained as the interpolation of the parameters of a given spectral envelope representation, using the same values of the morphing factor α as before. These values are represented as small "o" in figure 14.10. Finally, we measure the deviation presented by the values of "x" and "o" for each spectral envelope representation used, as illustrated in figure 14.10.

The error calculation can be seen as a measure of the deviation between the calculated feature values (represented by "o" in figure 14.10) and the target interpolated values (represented by "x" in figure 14.10).

$$\varepsilon\left(\delta_i\right) = \sqrt{\sum_{m=1}^{M}\left(\hat{\delta}_i\left(m\right) - \delta_i\left(m\right)\right)^2} = \sqrt{\sum_{m=1}^{M}\varepsilon_m^2} \tag{14.1}$$

Where ε represents the error measure, δ is a particular feature (i.e, centroid, kurtosis, etc), M is the number of equal steps the morphing factor α has between 0 and 1, and $\delta_i\left(m\right)$ represents the m^{th} measured value of the i^{th} spectral shape feature (whose m values are represented by "o" in figure 14.10) while $\hat{\delta}_i\left(m\right)$ is the m^{th} interpolated value of the same i^{th} feature (whose m values are represented by "x" in figure 14.10) for the same value of α.

From equation 14.1, we obtain an estimate of the error measure $\varepsilon\left(\delta_i\right)$ for each i^{th} spectral shape descriptor δ_i. One important thing to notice is that each spectral shape feature has a different range of values. The spectral centroid, for example, is measured is Hertz, while the spectral spread in Hertz squared and both spectral skewness and kurtosis are nondimensional. This fact would make the error evaluation meaningless because the range of values of each individual error $\varepsilon\left(\delta_i\right)$ depends intrinsically on the spectral shape feature δ_i used in the calculation. However, we can normalize the range of values of each spectral feature between 0 and 1 using equation 14.2 below

$$\Delta_i\left(m\right) = \frac{\delta_i\left(m\right) - \min\delta_i\left(m\right)}{\max\left[\delta_i\left(m\right) - \min\delta_i\left(m\right)\right]} = \lambda\left[\delta_i\left(m\right) - \eta\right] \tag{14.2}$$

where $\lambda_i = \frac{1}{\max[\delta_i(m) - \min\delta_i(m)]}$ and $\eta = \min\delta_i\left(m\right)$ are two scalar constants. Equation 14.2 is thus simply a linear transformation of the values of the spectral shape features and a new normalized error measure can be defined with it as below

$$\tilde{\varepsilon}_i = \varepsilon\left(\Delta_i\right) = \sqrt{\sum_{m=1}^{M}\left(\hat{\Delta}_i\left(m\right) - \Delta_i\left(m\right)\right)^2} = \sqrt{\sum_{m=1}^{M}\epsilon_m^2} \tag{14.3}$$

where $\tilde{\varepsilon}_i$ is the normalized error of the i^{th} spectral shape feature, $\Delta_i\left(m\right)$ is the m^{th}normalized measured value of the i^{th} spectral shape feature and$\hat{\Delta}_i\left(m\right)$ is the m^{th} normalized interpolated value of the same i^{th} feature. It is interesting to notice that the normalized interpolated values of each feature Δ_i coincide with the values of the mophing factor $\alpha\left(m\right)$, such that equation 14.3 can be rewritten as

$$\tilde{\varepsilon}_i = \sqrt{\sum_{m=1}^{M} \left(\hat{\Delta}_i(m) - \alpha(m)\right)^2} \tag{14.4}$$

It can be shown that there is a simple linear relationship between the normalized error measure $\tilde{\varepsilon}_i$ calculated from equation 14.4 and $\varepsilon(\delta_i)$ calculated from equation 14.1, given below

$$\tilde{\varepsilon}_i = \lambda_i \varepsilon(\delta_i) \tag{14.5}$$

Equation (14.5) tells us that the individual normalized error $\tilde{\varepsilon}_i$ calculated from the normalized spectral feature values Δ_i is simply the error calculation $\varepsilon(\delta_i)$ for each feature value δ_i normalized by the factor λ_i. The individual normalized error $\tilde{\varepsilon}_i$ is calculated for each frame of the source-filter representation and then averaged over the frames as shown in equation 14.11

$$\tilde{\varepsilon}_N(i) = \frac{1}{N} \sum_{n=1}^{N} \tilde{\varepsilon}_i(n) \tag{14.6}$$

where $\tilde{\varepsilon}_N(i)$ is the averaged individual normalized error of the i^{th} spectral shape feature, and $\tilde{\varepsilon}_i(n)$ is the individual normalized error for the i^{th} spectral shape feature of the n^{th} spectral frame.

Figure 14.11 exemplifies the calculation. Now we can define the total error ε_T simply as a weighted average of the averaged individual normalized errors $\tilde{\varepsilon}_N(i)$ for each descriptor as expressed below.

$$\varepsilon_T = \sum_{i=1}^{K} \omega_i \tilde{\varepsilon}_N(i) \tag{14.7}$$

In equation 14.7, ϵ_T is the total error computed for the K spectral shape features from each individual error $\tilde{\varepsilon}_N(i)$ weighted by ω_i. The weights ω_i give us the possibility of adjusting the individual influence of each descriptor in the final error, such that if we decide that the influence of the centroid is more important than that of the other descriptors in the final result, we would adjust the weights ω_i to reflect this. In this thesis the weights are the same.

The averaged individual normalized errors $\tilde{\varepsilon}_N(i)$ is a measure of the linearity of the spectral envelope morph for a given spectral shape feature for each pair of sounds. The total error ε_T is a measure of the linearity of the spectral envelope morph for a given spectral envelope representation for each pair of sounds. For both, the smaller the error, the more linear the spectral envelope morphing transformation.

The objective evaluation then becomes a simple comparison of the averaged individual normalized error values $\tilde{\varepsilon}_N(i)$ together with the total error ε_T estimated for each spectral envelope representation for each pair of sounds. The objective criterion adopted is to determine which spectral envelope representation leads to the minimum error value for a large collection of pairs of sounds. The idea is to study the spectral shape feature interpolation properties of the spectral envelope representations to investigate if there is one representation that consistently gives a small error when compared to the others.

Figure 14.12 compares the values of the averaged individual normalized error values $\tilde{\varepsilon}_N(i)$ together with the total error ε_T estimated for each spectral envelope representation for different pairs of sounds. Notice that for each spectral shape feature δ we compare the mean value of $\tilde{\varepsilon}_N$ over N frames, and the confidence interval, calculated as $1.96\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation of the the averaged individual normalized error values $\tilde{\varepsilon}_N(i)$. On the righ-hand side of each figure we see the total error for the pair of sounds in question. The rightmost bottom plot is the average value of the total error ε_T for all pairs of sounds tested. This plot is of utmost

importance because its lowest bar indicates the most linear spectral envelope morphing method for all spectral shape features for all sounds. After analyzing figure 14.12, we conclude that the spectral envelope morphing technique with minimum error is interpolation of LSFs. Therefore interpolation of LSFs was the most linear spectral envelope morphing technique concerning the variation of spectral shape features. Interpolation of LSFs was adopted as the spectral envelope morphing technique in this thesis.

The minimal error only takes the spectral shape features into consideration, independently of the behavior of the spectral peaks. Therefore, an online listening test was performed to cross-validate the result of the qualitative and quantitative analyses. In this listening test, we compare the SF model with a sinusoidal model. For the SF model, LSFs were used as the spectral envelope morphing method. Notice that ENV represents the interpolation of sinusoidal models, very popular in the literature. The listening test investigates which sound morphing algorithm leads to a more perceptually linear transformation. The minimum feature interpolation error criterion is considered the objective evaluation procedure, while the listening test is regarded as a mean subjective perceptual evaluation.

14.2.2 Subjective Perceptual Evaluation

The main purpose of the subjective perceptual evaluation is to verify whether the cyclostationary morphs obtained with the SF model are perceptually linear when the morphing factor varies linearly. This would mean that the morphing factor α allows perceptual control of the morph and that the spectral shape features selected to guide the transformation do measure perceptually relevant information concerning the morphing transformations performed. Naturally, the result of the minimum error evaluation presented previously was used to select the spectral envelope morphing technique used in the morph.

Evaluating the linearity of the morph can be a very difficult task, depending on how we choose to do it. Hikichi and Osaka [Hikichi, 2001] compare the perceptual distance between the steps of the morph and use MDS spaces to verify if the result is a straight line. Their conclusion is that generally the morph gives a rather curved line in the MDS space. This implies that the results were not intermediate. However, it may also be due to perceptual phenomena. This experimental setup seems to raise more questions than it answers, so a simple comparison betwen the popular interpolation of sinusoidal models and the SF model was chosen instead.

The listening test compares the linearity of morphing transformations between musical instrument sounds obtained with sinusoidal analysis and the SF model. The SF model used LSFs to morph the spectral envelopes, while the sinusoidal morphing used the standard interpolation of partials frequency and amplitude values. The temporal alignment step is the same for both methods (and uses markers annotated by hand), only the spectral morphing procedure changes. The test is available online (http://recherche.ircam.fr/anasyn/caetano/survey/smoothness.html).

The listening test presented 11 pairs of cyclostationary morphs and asked the participants which was "smoother". Participants could either choose a method, or have no preference. The instructions presented an example of a cyclostationary morph with uneven perceptual intervals between steps, and another one that was considered "smoother" to explain what the participants should listen for. The example sounds were not used in the test, and the "uneven" morph used sounds from the "smoother" cyclostationary morph shuffled in order.

The instructions also explicitly said that one column did not correspond to an algorithm to avoid biasing the results. Appendix G has the instructions used in the test. After taking the test, the participants were asked whether they used headphones, and whether they were experienced in music or audio evaluation. The results of participants who answered "no" to any of the questions were not used. This does not guarantee uniformity because we don't control the experimental setup,

but at least we can make sure that the results of participants who didn't follow the instructions are not included. In total, the results of 58 participants aged between 22 and 53 were used.

Figure 14.13 shows the results of the listening test used to evaluate the linearity of the morphing algorithms. In figure 14.13 we see the mean of percentage ratings of each pair, together with a global average representing the performance of the models for all sounds. The labels in figure 14.13 can be found in table 14.2, and the legend stands for source-filter model (SF), no preference (=), and sinusoidal model (Sin). First of all we notice in figure 14.13 that the performance of the algorithms depends on the pair of sounds used. It is also important to notice that, for some pairs, many participants manifested no preference. In fact, figure 14.13 shows that there is no clearly predominant algorithm, especially when we see that the result of the global average shows that the SF model outperforms the sinusoidal model by a narrow margin. Also, no preference represents a significant percentage of the choices for most sounds, as well as for the global average. If the participants were forced to choose between one algorithm or the other, we would probably produce artificial results because there was no clear preference.

14.2.3 Temporal Envelope Morphing

In this section we will see the results of morphing the temporal envelope. The temporal envelope curves were estimated using RMS and then morphed. The temporal envelope morphing techniques compared in this section are morphing the temporal envelope curve directly or interpolating the cepstral coefficients that represent it. Since the techniques for estimation and representation of temporal envelope curves used in this thesis are analogous to spectral envelope, the objective evaluation presented will be similar.

Figure 14.14 shows the morphed temporal envelope curves and the variation of the temporal centroid for the interpolation of curves and CCs for the instruments marked. Figure 14.15 shows the the same information for other instruments. Finally, figure 14.16 compares the linearity error of the temporal centroid between both temporal envelope morphing techniques. The result of the comparison shown in figure 14.16 indicates that the interpolation of the cepstral coefficient representation of the temporal envelope is the most linear temporal envelope morphing technique when we consider linearity of the temporal centroid.

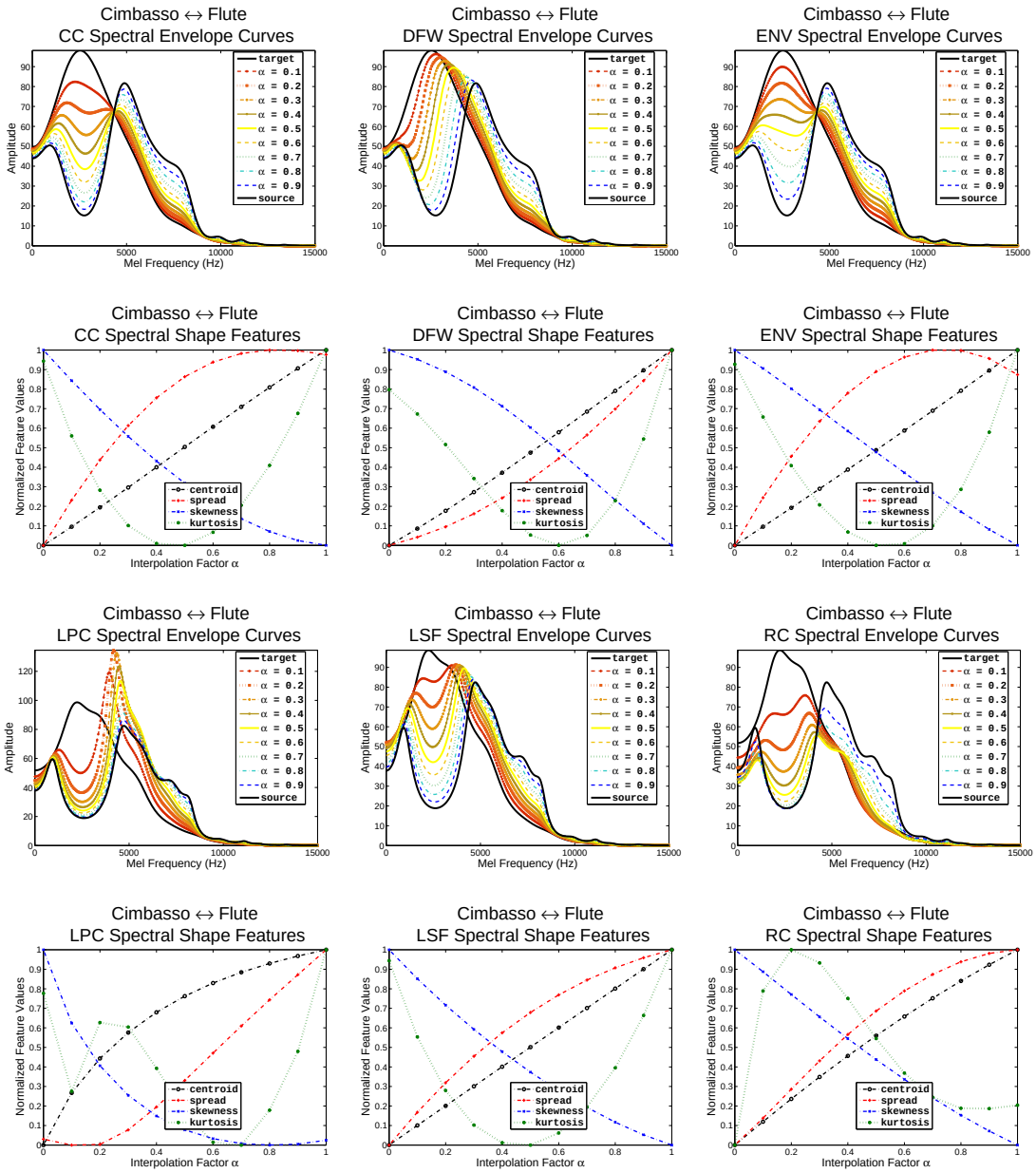

Figure 14.2: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

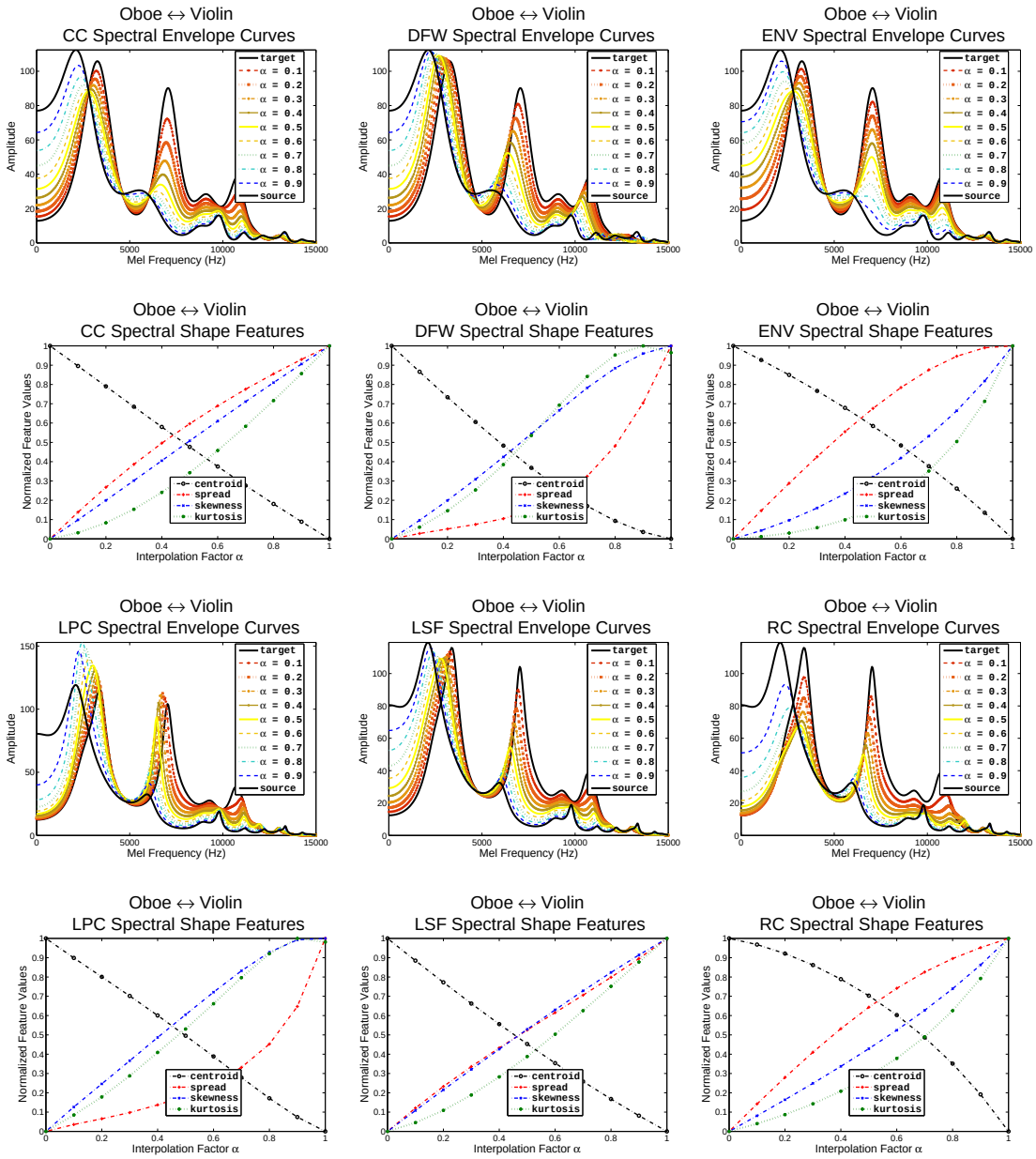

Figure 14.3: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

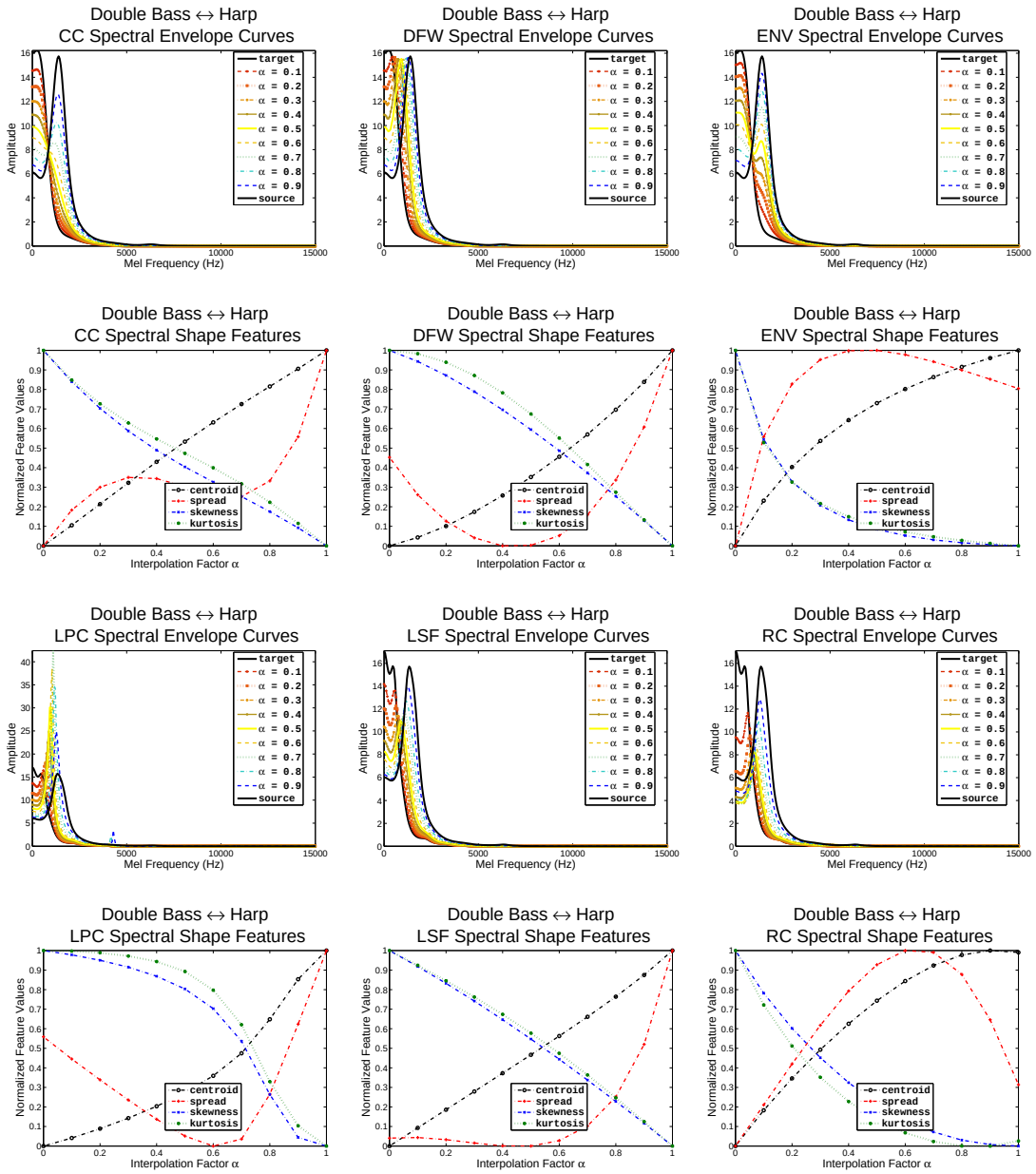

Figure 14.4: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

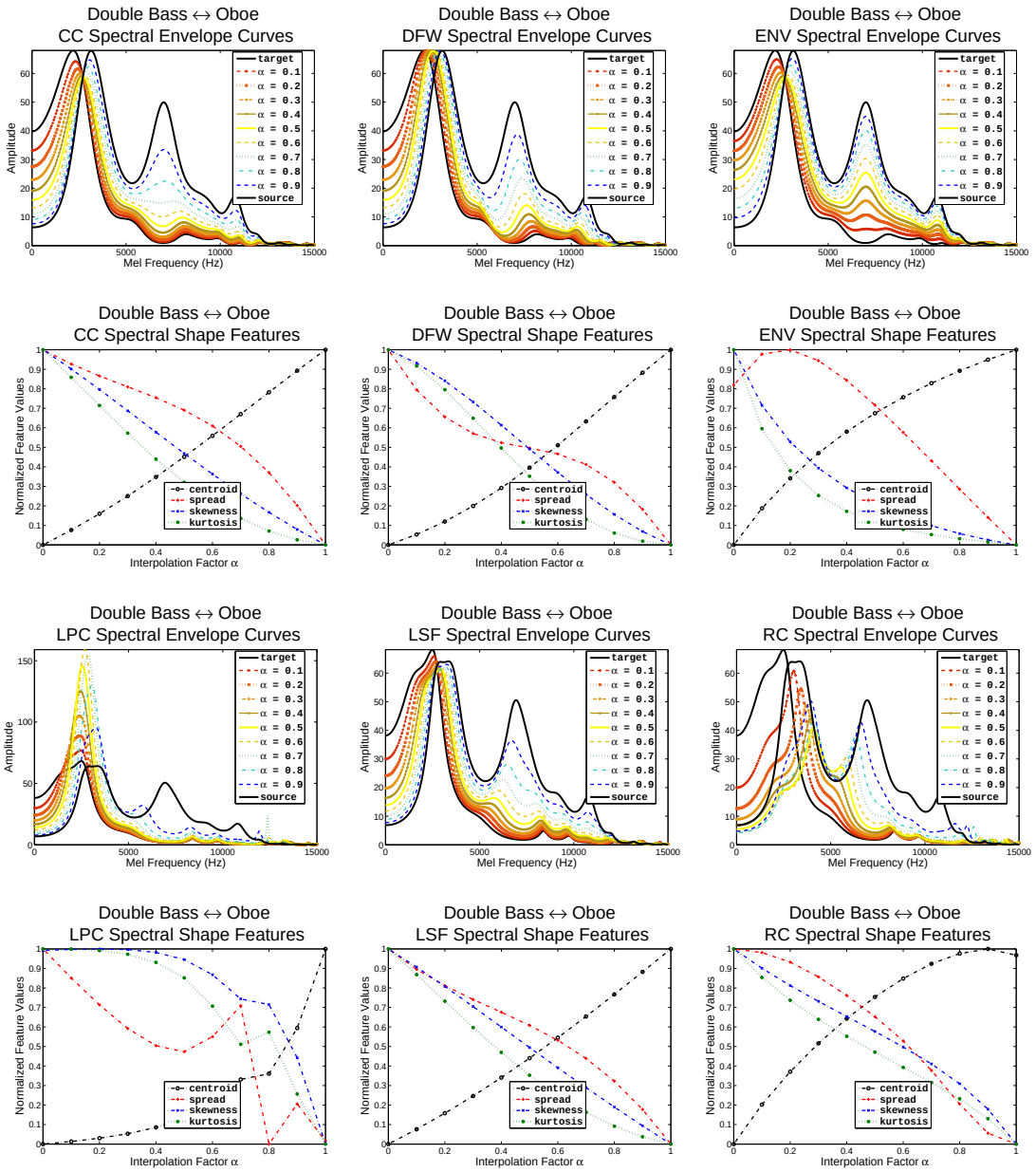

Figure 14.5: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

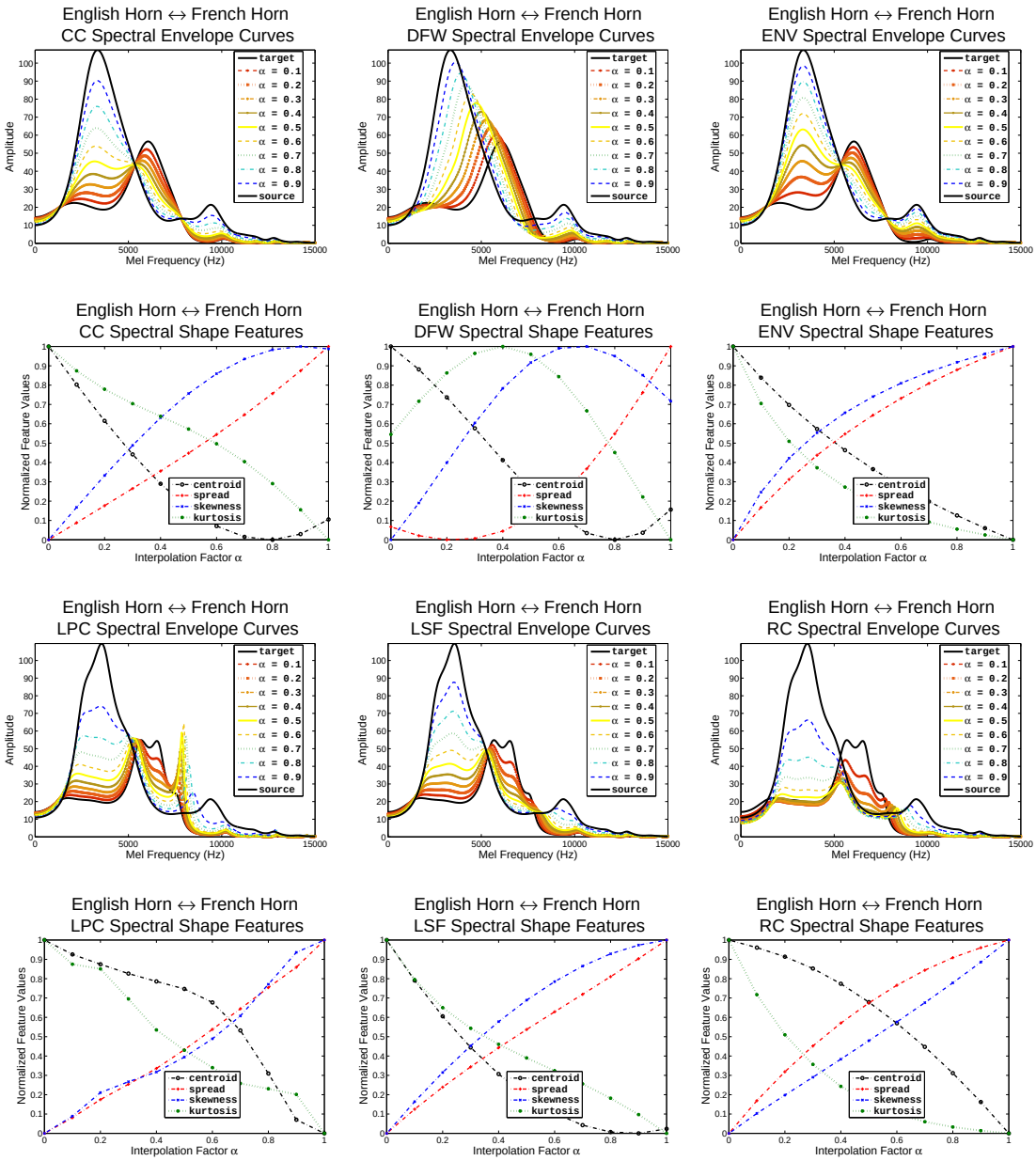

Figure 14.6: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

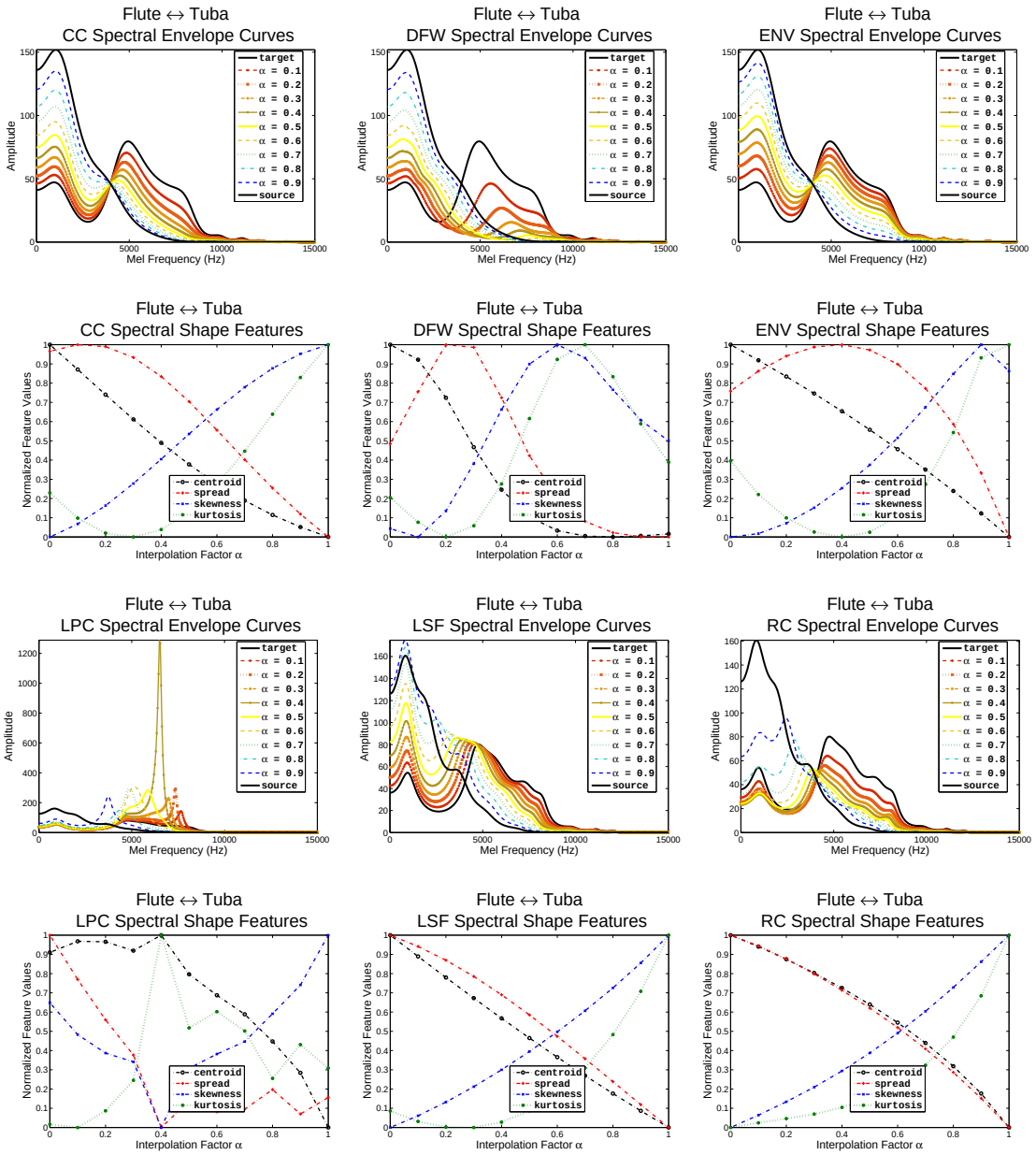

Figure 14.7: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods.

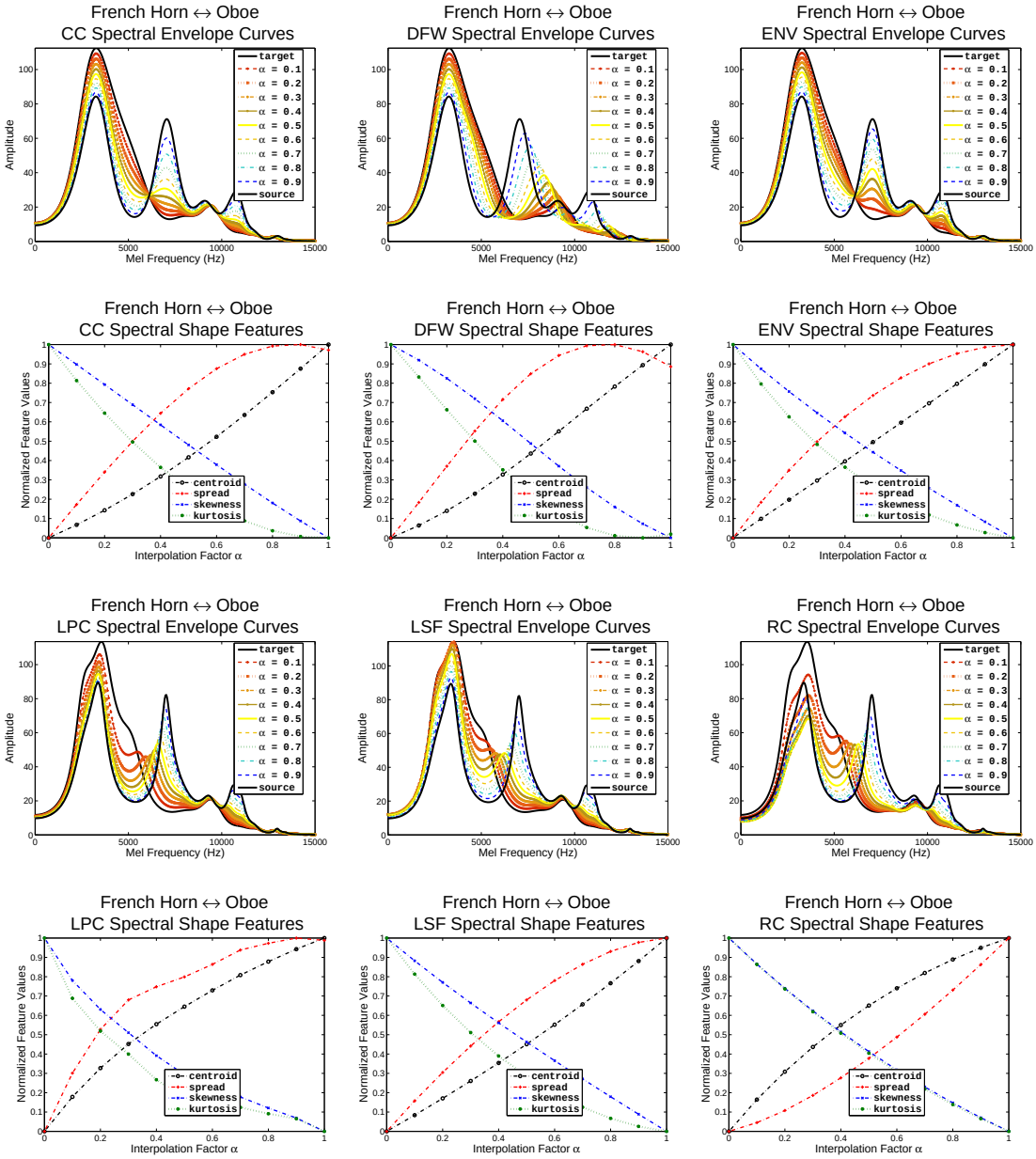

Figure 14.8: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods

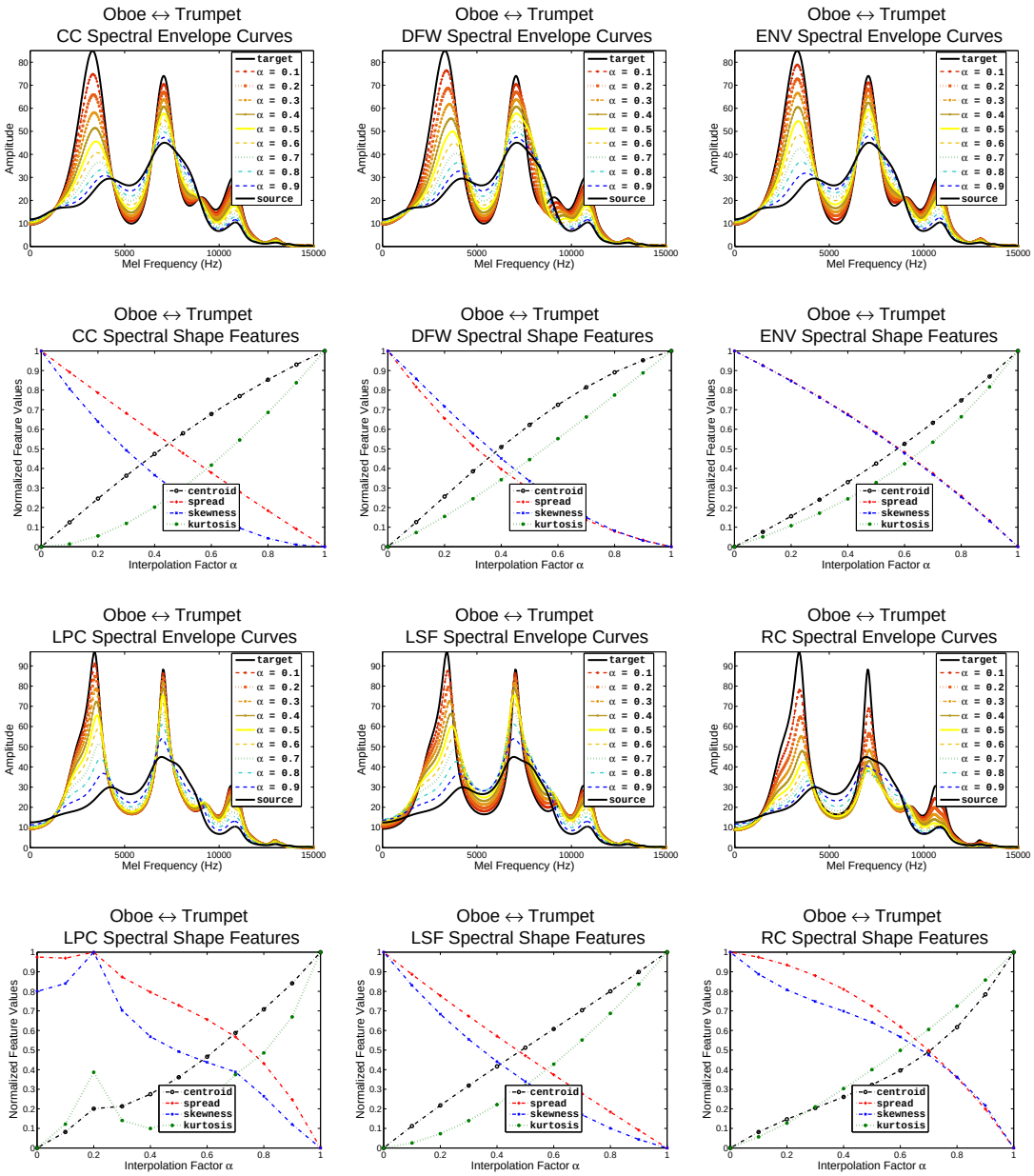

Figure 14.9: Behavior of spectral peaks. The figure compares the behavior of the spectral peaks for several spectral envelope morphing methods.

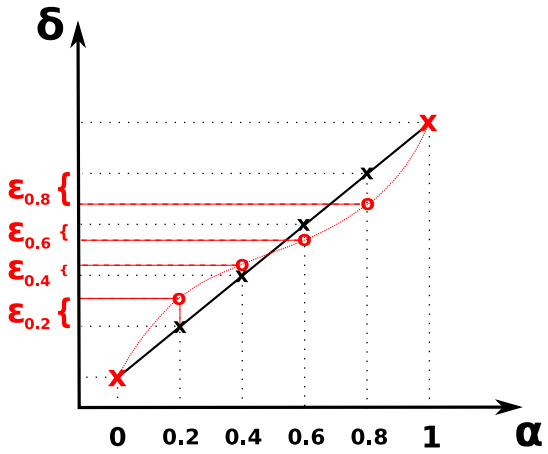

Figure 14.10: Error calculation. This figure depicts the calculation of the feature interpolation error. The ideal values obtained as a linear regression are represented as "x", while the calculated values are represented as "o".

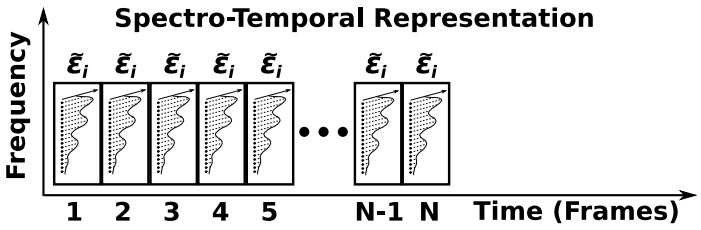

Figure 14.11: Error calculation for each spectral frame of the source-filter representation. The figure shows that the normalized error $\tilde{\varepsilon}_i$ is calculated for each spectral frame.

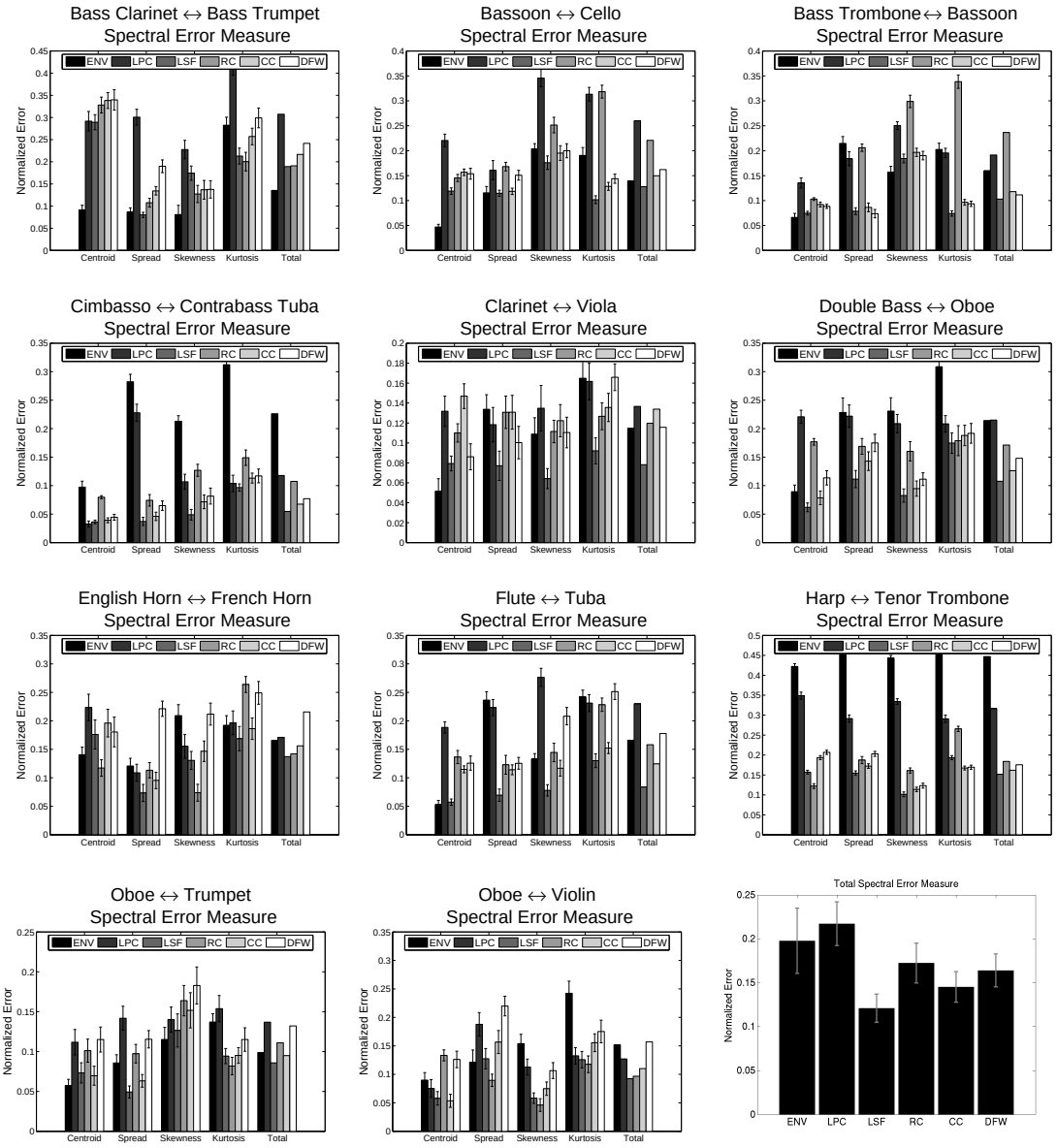

Figure 14.12: Error analysis. The figure shows the deviation between the ideal and the measured values of spectral shape features for each sound.

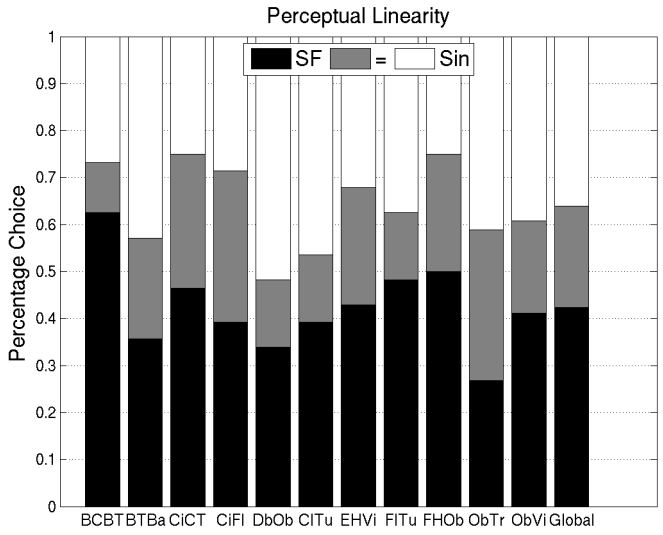

Figure 14.13: Perceptual linearity. The figure shows the percentage of participants who rated the SF algorithm smoother, the sinusoidal algorithm (Sin) smoother, or manifested no preference (=) for all pairs of sounds and a global average.

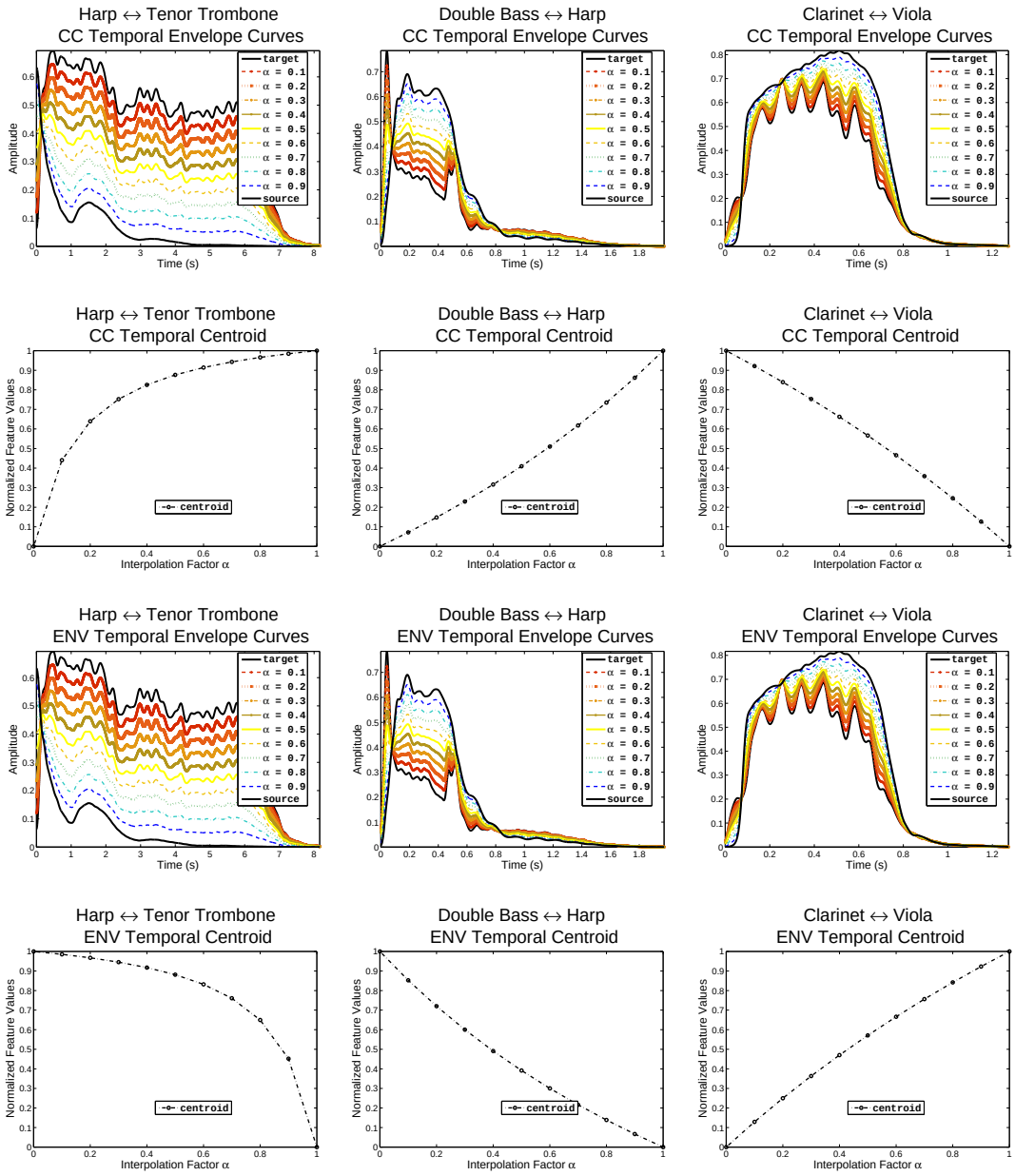

Figure 14.14: Morphing the temporal envelope curve directly. The figure shows the temporal envelope curves on the left-hand side and the corresponding variation of the temporal centroid on the right-hand side.

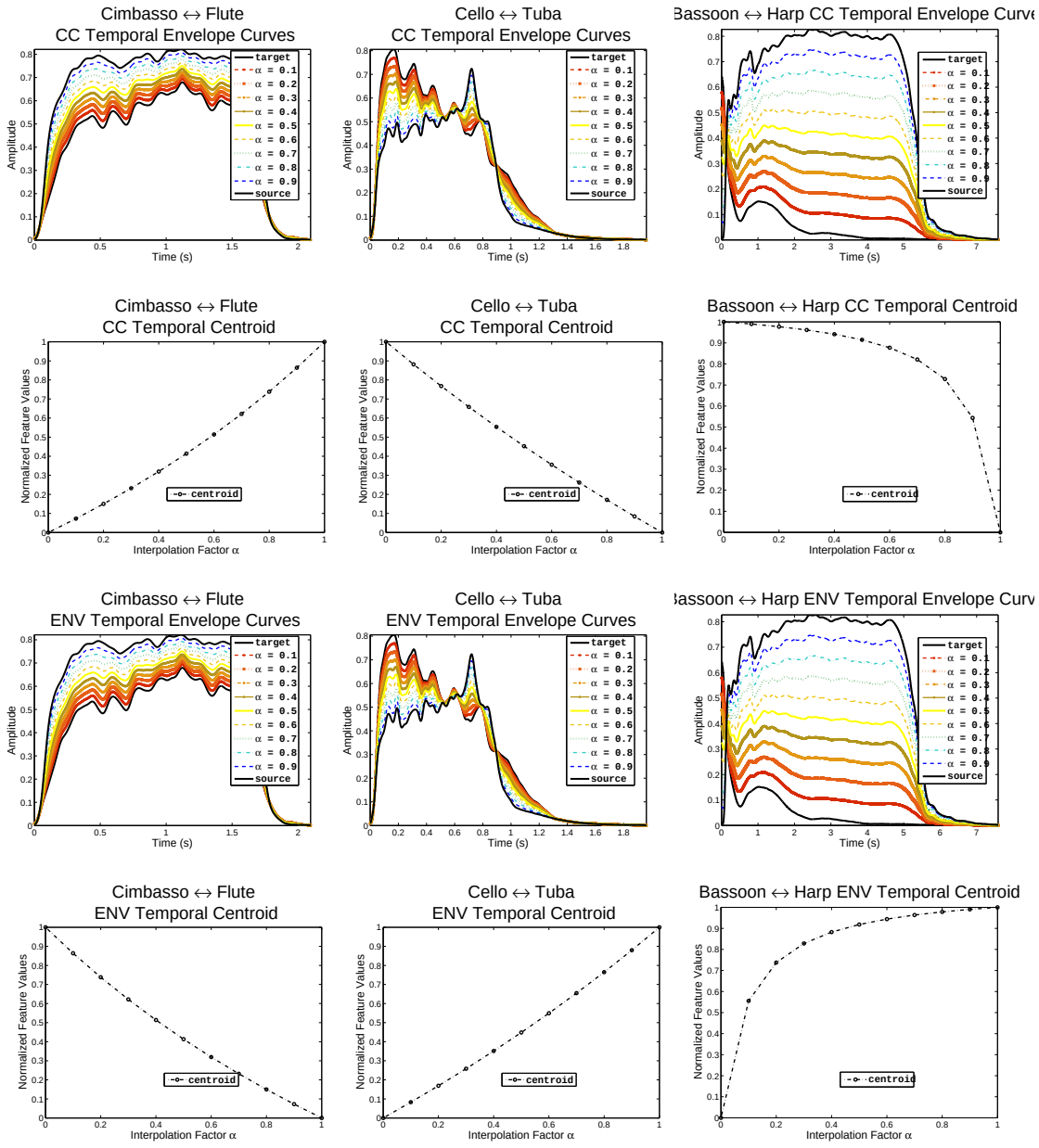

Figure 14.15: Morphing the temporal envelope cepstral coefficient representation. The figure shows the temporal envelope curves on the left-hand side and the corresponding variation of the temporal centroid on the right-hand side.

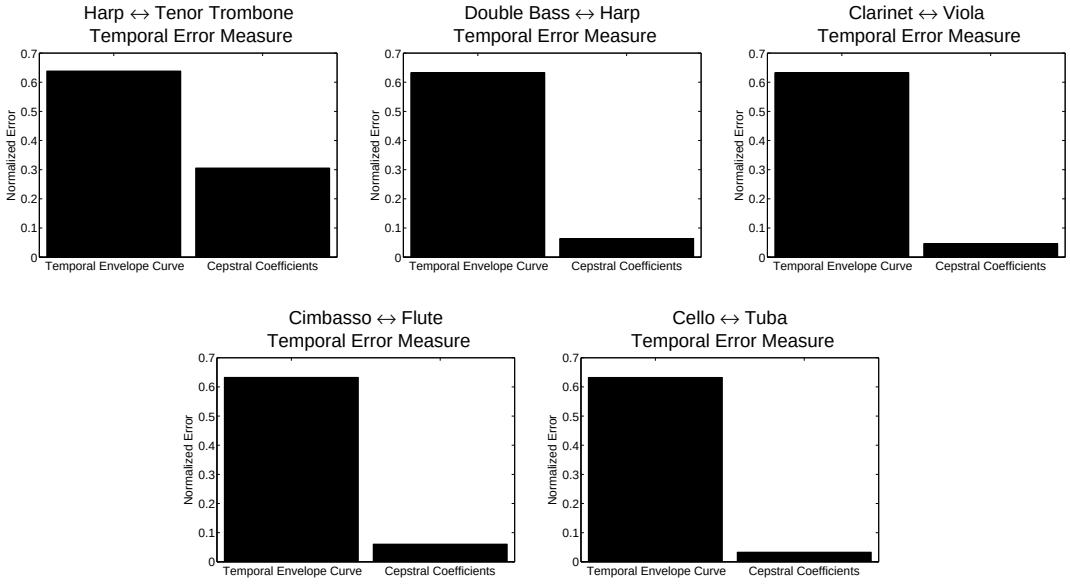

Figure 14.16: Error analysis for temporal envelope morphing. The figure shows the deviation between the ideal and calculated values of the temporal centroid for both temporal envelope morphing methods.

Chapter 15

Conclusions and Future Perspectives

This thesis was about morphing isolated quasi-harmonic acoustic musical instrument sounds guided by perceptually motivated features. The focus was on timbral features of musical instrument sounds and how to control them . The ultimate goal was to develop a method that gives the user perceptually intuitive control of the morph by means of the morphing factor alone. So, when the morphing factor varies linearly, we wanted the morph to be as perceptually linear as possible.

The usual approach to morphing sounds uses the interpolation principle, which consists of representing the sounds with a model, interpolating the parameters of the model representation, and resynthesizing the morphed sound with the interpolated parameter values. When the parameters of the model capture perceptually relevant features of sounds, the morphed sound might be perceptually intermediate. However, most models used in sound morphing tend to produce nonlinear morphs, so this work aimed to develop a method to obtain more perceptually linear morphs guided by perceptually motivated features.

There seems to be no consensus in the literature about what sound morphing is, or equivalently, what transformations can be considered morphing. This thesis approached this question from a theoretical and technical perspectives, discussing thoroughly the requirements of morphing and the difference between morphing and other hybridization processes. This thesis reviewed thoroughly the different transformations usually referred to as morphing in the literature, and proposed a classification system according to conceptual criteria. The cyclostationary morph figures prominently among the different morphing transformations considered very challenging because we need to accurately control temporal and spectral aspects of the morph to obtain a perceptually linear result. Transformations that happen during the course of a sound are more artistically appealing, but do not allow investigation of temporal aspects such as attack time.

The formalization of morphing proposed does not exclude more than one possible morphing transformation between sounds and this can render the evaluation of the results very difficult. A very challenging aspect of evaluating morphed sounds lies in the subjectivity usually applied in the evaluation. Usually, listeners have their own expectations about morphing, and the evaluation reflects merely whether the morph met those expectations or not. This work proposed to adopt evaluation criteria to evaluate morphing, namely, correspondence, intermediateness, and smoothness.

This thesis formalized the concept of morphing sounds, proposed a general algorithm, and a framework to objectively evaluate morphing using the criteria adopted. There were two important steps considered before actually performing the morphing itself, temporal alignment followed by spectral modeling. In this thesis, a temporal alignment procedure based on a perceptually motivated temporal segmentation model was proposed. Then, the spectral modeling was introduced, along with the motivation for the development of a model dedicated to morphing sounds, namely,

the source-filter (SF) model.

This thesis presented a SF model for musical instrument sounds that gives independent control of the spectral envelope and frequency of the partials to perform the transformations. The sounds to be morphed are decomposed into a sinusoidal and a residual parts, which are represented independently with the SF model. The sinusoidal component comprises a time-varying spectral envelope model (filter) and the frequencies of the partials (source), while the residual component is modeled as white noise (source) shaped by a time-varying spectral envelope model (filter). The SF representation was validated with a perceptual similarity test. Participants were presented the original and SF representation of sounds and asked to rate their perceptual similarity.

With a few exceptions, most works about morphing in the literature skip the evaluation of the results, usually considered too difficult and subjective. The evaluation is considered a crucial part of this work, responsible for the validation of the results. In this thesis, the evaluation consisted in verifying the linearity of the morph using objective measures and subjective tests. Three criteria were adopted to evaluate the results, namely, correspondence, intermediateness, and smoothness. This thesis proposed to use perceptually related features to objectively evaluate the linearity of the morph. There was a clear focus on morphing the spectral envelope, an important part of the SF model used to represent the sounds that is related to the perception of the timbral subset of attributes called sound color. A major part of the evaluation lay in the comparison of the linearity across several spectral envelope morphing techniques found in the literature together with others prosed in the scope of this thesis.

Lastly, perhaps the most essential aspect of sound morphing to be taken into consideration is the perceptual impact. Morphing depends essentially on perceptual phenomena, and the key to a perceptually relevant morph seems to lie in tricking the ear and the brain into forming a convincing auditory image. However, the creation of this apparent "auditory illusion" involves other questions about the perceptual or conceptual distance between the mental representation of sounds. The evaluation criteria adopted implicitly supposes that the mental representation of musical instruments sounds is metric, just like the MDS timbre spaces proposed to represent it. The evaluation criteria also implicitly supposes that the perception of musical instrument sounds is continuous rather than categorical when it assumes that it is always possible to obtain an intermediate representation of two sounds or a gradual transition between them.

The conclusions and future perspectives of this thesis should reflect how important perception is in sound morphing. The comments and remarks made by the participants of the listening tests seem to be a good starting point to evaluate the perceptual impact of the method proposed. Let us see why.

15.1 Comments and Remarks

Both tests met with a very positive response. Apart from the data, the participants also gave some extremely important feedback on the task and how they performed it. Some of it can be found below.

15.1.1 Perceptual Similarity

A natural consequence of adopting the source-filter (SF) model to perform the morphs is that the SF representation is morphed rather than the original musical instrument sounds. Thus I performed an online listening test to evaluate the perceptual similarity between the original recordings and their SF representations. The SF model is appropriate for sound morphing (among other sound transformations) if the original sounds and their SF representations are indeed perceptually similar. The result of this test validated the SF representation of the sounds in general. Some sounds were

found to be perceptually closer than others. It was generally agreed that the woodwinds are well represented. Some interesting comments included:

- "The differences in 'slightly different' meant that I had to listen a couple of times in a different order to sense and confirm minor differences in spectrum."
- "The horns and winds were generally better than the strings, but that's to be expected."
- "The string instruments were almost always different on the whole. The wind instruments seem to be reproduced rather nicely."
- "I tried really hard, but often I couldn't hear any differences at all. With some of the wind instruments, the model doesn't get the onset quite right."
- "For some instruments, the attack envelope is not perfectly identical, and for some the vibrato seems to be stronger in the model than in the original recording. Other than that, the timbre is modeled so well that I could not find a difference for most samples."
- "Often, 'slightly different' comes because the attack is less "abrupt" or "shar" "
- "I believe string instruments are reproduced with an impressive similarity. In my ears, the brass instruments tend to differ more. I believe this is due to the fact that the start of the blow is not reproduced accurately and therefore sounds more artificial."
- "I would say the sounds differ the most in their attack segments and this gives the modeled one a different touch."
- "The subjective similarity classes are a bit hard to handle, so I put 'significantly different' only to the one I thought wasn't even resembling the same thing (although possibly simulating the same instrument)"
- "I chose 'fairly different' for some sounds to use the full range of the five choices given. But, fundamentally, the pairs are very close. I would say that the 'very different' are rather 'fairly different"'
- "In my understanding of English, 'significantly different' is not as strong a difference as 'very different'. I rated some sounds as slightly different based on my perception of slight differences in attack."
- "I would say (as a native English speaker) that 'significantly different' is LESS strong than 'very different'. I've ticked the 'very different boxes' in the sense of them being different on a 4 out of 5 scale, even though orally I would have described them as 'significantly' rather than 'very' different."

It is important to bear in mind that the instructions did not warn the participants of the presence of identical pairs. A number of participants rated identical pairs slightly different. Interestingly, even though the names of the instruments were not explicitly written on the web page, most participants still refer to the sounds by instrument family.

Some participants reported difficulty in fitting the proposed similarity scale into their cognitive representations of differences. The labels used in the subjective scale conflicted with the interpretation of the scale, which can be confusing. Also, the number of divisions was mentioned as a possible factor that impaired judgment.

15.1.2 Perceptual Linearity

The aim of this thesis is to develop a method that gives the user perceptually intuitive control of the morph by means of the morphing factor alone. That is, the morph should be perceptually linear when the morphing factor varies linearly (in equal steps). Additionally to objectively evaluating the objective linearity of several spectral envelope morphing techniques, it is important to evaluate the perceptual linearity of the results. Therefore, an online listening test was proposed in which the SF model is compared to a sinusoidal model. The perceptual linearity test was overall judged to be too difficult. Some people found that other perceptual cues, such as loudness differences, impaired the assessment of the timbral differences.

- "For some of the sequences, the perceived loudness of the tones seemed to differ between renderings, making it sound bumpy although the timbre might have been good... I think my judgment may have been influenced by this, don't know if this is in your interest or not."
- "I found myself identifying instruments and then listening for when I noticed their sound disappear or appear in the sequence."
- "The method relies a lot on memory, as well as discrimination. It took a little bit to familiarize myself with the task. I think it would be better if it didn't automatically repeat. It would be helpful to be able to play the endpoints so that you would know where the morph is supposed to be going. In many cases, it didn't seem like the morph was changing the sound very much. The sounds were very synthetic. Have you tried morphing high quality instrument sounds?"
- "The task was pretty difficult. Needed a few listenings for some of the pairs."
- "Some of the transitions are not evenly spread over the scale (0.1-1), to my ear both algorithms have a slight preference to make "bumpy" transitions later (near 1) - maybe it is a pure perceptual issue."
- "Sounds like you are morphing string, brass, and woodwinds by attack and overtones."
- "I feel like the results are less accurate towards the last tasks because it's hard to keep concentrated."
- "Quite a difficult test! Some sound examples had clicks in the beginning, which may cause the transitions to sound unsmooth (I assume that was not intended)."
- "The difference is too subtle to tell."
- "The artifacts were sometimes annoying and may have had an influence on some occasions. I listened to the 'curve' of the entry (or suppression) of the upper partials for most of my judgments. I preferred the upper partials to start to change with the first transformation rather than being delayed."
- "I have to say that the task wasn't very easy for me to do as I didn't have a very obvious preference in most cases. One thing I would have liked was to have source and target sound equalized in loudness. In most cases I thought that they were differing in terms of loudness and I was mostly judging smoothness in loudness change rather than in the overall timbre."
- "I found it very hard to judge, and I believe it's due to three factors. 1: The timbres themselves are pretty dull. 2: The morphing is done in sequence, rather than seamlessly over a period of time. 3: The frequency remains constant, i.e. no melodic elements."
- "I found it hard to judge smoothness."

- "For most examples that I don't specify a preference, both series sound definitely smooth to me, and I can't choose the smoother"

- "This was tough. I tried to go on gut instinct when I wasn't sure why I thought one was better than the other. Often they sounded the same."

Again, many participants still refer to the sounds by instrument family even though the names of the instruments were not explicitly written on the web page. The task was generally considered very difficult or too difficult. There are many likely reasons. Firstly, the task itself was difficult to understand. For cyclostationary morphs, the evaluation of perceptual linearity implies judging and comparing the intervals between the steps of the morph. To avoid the cumbersome task of evaluating each step individually, the proposed task was to compare two different morphing algorithms. Evaluation of perceptual linearity involves judging several characteristics of sounds at the same time and remembering them for comparison across steps. The evaluation of the intervals between steps of the transformation might use one single feature (e.g., attack time) or mutiple criteria (attack time, spectrum, etc).

The big cognitive load of the task compromised the evaluations in some cases. The task of judging the intervals between steps is already difficult, and the comparison across all steps can be complex. Moreover, the participants were asked to perform this task twice for each pair of sounds (there were 11 pairs) and compare them, which relies heavily on memory. Another key factor was that the number of steps used was considered too big. The more steps, the more the task relies on memory to perform the comparison.

The sound material used in the test also influenced the evaluation for some participants. In general, the source and target sounds were considered very similar, which made the task even more difficult. For some participants, even the terminology was considered confusing. The term perceptual linearity was avoided in the instructions and replaced by "smoothness", considered more appropriate for the task. However, most participants seemed to have grasped well the concept of "smoothness", probably thanks to the example.

There are different aspects to be discussed. First of all, the type of sound transformation investigated, cyclostationary morphing, is particularly challenging in many respects because it involves temporal and spectral manipulations. This choice affects not only the techniques needed to achieve the transformations, but also the impact of the result and consequently its evaluation.

Secondly, the sounds are decomposed into a sinusoidal and residual parts that are modeled separately. This decomposition permits to treat both sinusoidal and residual parts independently, but also needed the development of separate models. One important part of the work concentrated on the implementation of different instances of the source-filter model for the sinusoidal and residual components. The perceptual quality of the SF representation plays a fundamental role on the quality of the morphed sounds.

The automatic segmentation model developed in the scope of this work together with the temporal alignment technique also have a significant impact on the quality of the results. The temporal alignment technique focuses on four perceptually inspired regions of musical instrument sounds, and a simple time stretch/compress procedure.

A large effort in this work, however, concerned the investigation of spectral morphing techniques. The linearity of the transformation obtained with several spectral envelope morphing techniques was investigated. Linearity was measured using acoustic correlates of timbre dimensions obtained from psychoacoustic studies. The choice of sound material and how it influenced the evaluation of the results will be discussed.

Finally, the choice of evaluation criteria deserves some comments, given that most previous work on sound morphing hardly evaluate the results.

15.2 Type of Transformation

Cyclostationary morphs are very challenging sound transformations because they involve temporal and spectral manipulations. Dynamic transformations are more artistically appealing, but do not allow the investigation of manipulations of global temporal aspects such as attack time. When the transformation occurs during the course of a sound (such that the sound changes while we hear it), it usually changes only the steady state portion. This thesis developed techniques to independently manipulate the global temporal and spectral characteristics of the musical instrument sounds being morphed.

The attack is responsible for one of the most perceptually salient dimensions of timbre perception. Consequently, manipulations of the attack usually have a major impact on the perception of musical instrument sounds. In a cyclostationary morph, different steps of the morph are expected to present intermediate perceptual features that gradually change from source to target sound. The constraint that the transformation must be perceptually linear requires that the perceptual difference across steps of the morph be the same. Thus the attack time needs to be manipulated accordingly to be perceived as linearly varying from source to target.

However, manipulation of the attack is not enough to obtain a perceptually linear cyclostationary morph. The interpolation of spectral information also has a significant impact on the perception of linearity. The spectral centroid, a measure of the baricenter of the distribution of spectral energy, was found to be correlated with a perceptually salient dimension of timbre perception in studies of musical instrument sound (dis)similarity. This is an indication that the distribution of spectral energy has to be controlled appropriately.

15.3 Sinusoidal plus Residual Decomposition

The attack is the only event that is present in all sounds, be they environmental, acoustic, synthetic, etc. For musical instrument sounds, the attack contains transients that take place between the onset and the moment when more sustained vibrations occur. These transients are not very well modeled by sinusoids, such that the sinusoidal component of a musical instrument sound notoriously lacks the noisy sudden attack characteristics.

The inclusion of a SF model of the residual was therefore considered vital to achieve a perceptually similar representation of the musical instrument sounds used in the morphs. The residual component was represented as white noise shaped by a time-varying filter that models the residual from the sinusoidal component. The residual component was morphed independently from the sinusoidal component using the same spectral envelope morphing technique.

In the similarity test, many participants reported using the onset or attack cues to differentiate between the sounds. Specially important seemed to be the noisy cues such as blowing or bowing, which most participants used to assess the dissimilarity between the sounds. Although it confirms the well established fact that the attack is essential in musical instrument sound perception, it also implies that the residual component does not capture perfectly the perceptual noisiness. It might be interesting to perform listening tests to evaluate the SF representation of the sinusoidal and residual components separately.

15.4 Source-Filter Model

Most sound morphing techniques found in the literature apply the interpolation principle directly on the parameters of a sinusoidal model. Sinusoidal models are very popular in part due to the quality of the representation of a broad class of sounds. The perceptual similarity between the sinusoidal representation and the original sound is well known. However, a drawback of using

sinusoidal models in sound transformations is that the number of parameters is proportional to the number of partials. Each partial is modeled independently, and the number of parameters to control can grow fast.

The SF representation, in turn, models the amplitude of the partials with a spectral envelope curve, which has a limited number of parameters that depend solely on the fundamental frequency of the spectrum, not on the number of partials. The partials, on the other hand, are modeled as sinusoids that sample the spectral envelope curve at certain frequency values. When the spectral envelope is represented as a linear shift-invariant (LSI) system upon resynthesis, the amplitudes of the partials are readily retrieved because sinusoids are the eigenfuctions of LSI systems.

Perceptually, the SF representation must be as close as possible to the original recordings. The are several studies on the perception of sounds with slight alterations [Horner et al., 2009, Grey and Gordon, 1978, McAdams et al., 1999]. Usually, the authors of these studies are interested in investigating whether alterations to the sounds can be perceived or not, using discrimination tests [McAdams et al., 1999]. Each of these studies proposes specific alterations, and the analysis of the results is application and domain dependent. The SF representation of sounds proposed in this thesis corresponds mainly to modifications to the amplitudes of the partials on each frame. Instead of looking at each frame individually, we can focus on the temporal evolution of the amplitude of each partial. In this case, the SF representation corresponds to alterations to the amplitude envelope of each partial. Grey [Grey and Gordon, 1978] proposed a similar investigation where he approximated the amplitude envelope of the partials by straight line segments. He concluded that these alterations generally are not perceptually relevant. The perceptual test presented in this thesis also showed that the SF representation of the musical instrument sounds tested was generally considered perceptually accurate enough.

The SF model is very appropriate for sound transformations because of its independent and compact spectral representation. One advantage of the SF model over traditional sinusoidal models is the independent representation of the amplitudes and frequencies of the partials. The amplitudes of the partials are represented as a spectral envelope curve, whose parameters normally give smooth continuous transformed spectral envelope curves when manipulated. The manipulation of the amplitudes of the partials via the spectral envelope model can also be beneficial for transformations guided by features. Different spectral envelope representations will generally lead to different behaviors when manipulated, and we are free to choose whatever representation is more appropriate to a certain application or to obtain a desired effect.

15.5 Temporal Processing

The temporal processing steps are responsible for many perceptually important parts of the mothod developed in this work. The temporal processing happens in two distinct steps, before spectral morphing in a pre-processing step called temporal alignment, and during spectral morphing the temporal envelope procedure modulates the morphed frames. Each one of them will be considered separately.

15.5.1 Automatic Segmentation

The segmentation task consisted in automatically detecting the boundaries between perceptually motivated segments of musical instrument sounds, such as attack and steady state, according to an underlying model. The automatic segmentation task uses a model to define the regions and their boundaries, and identifies the boundaries using detection functions. The behavior of the detection function must stand out during the events to be detected. It is notoriously difficult to find a robust

model of segmentation of musical instrument sounds and detection functions that present good performance for a broad range of sounds.

The model used to define the regions and their boundaries plays an essential role in the results of the automatic segmentation task. For example, the attack/decay/sustain/release (ADSR) model supposes that there is always a decay region after the attack, which might impair the segmentation results when the sound being segmented does not fit the model. In this thesis, the amplitude/centroid trajectory (ACT) model of segmentation of musical instrument sounds [Hajda, 1996] was adopted.

The ACT model, which was originally proposed for sustained sounds, uses the temporal envelope and temporal variation of spectral centroid as detection functions. In general, for sustained musical instrument sounds, both the temporal envelope and the spectral centroid behave as the model predicts. In this work, it was empirically verified that percussive sounds do not fit the model well and render generally poor segmentation results. The temporal envelope and centroid tend to behave differently than predicted by the model for these cases. The same applies to most plucked string sounds, with a few exceptions, such as guitar sounds. For these sounds, the spectral centroid behaves like predicted by the model, but the temporal envelope does not.

Even for sustained musical instrument sounds, it was verified that it is very difficult to achieve robust automatic detection of the boundaries for a broad class of musical instruments. The results depend on instrument family and even pitch. The behavior of the spectral centroid, in particular, fit the model well for some cases, while others revealed to be more challenging. The results of the automatic segmentation algorithm proposed in this work were compared with the baseline AR model [Peeters, 2004]. In general, the algorithm proposed outperformed the baseline method. However, only empirical tests were used to compare the segmentation results. Formal validation of the segmentation results would require a more careful comparison.

15.5.2 Temporal Alignment

First of all, it is important to note that he temporal alignment procedure developed in this work is totally independent of the temporal segmentation task. All the temporal alignment technique needs as input is the time markers corresponding to the boundaries of the regions of both sounds, and the output will be two sounds whose corresponding regions will be properly aligned in time. Therefore, we can annotate the sounds by hand, for example, if we choose to use sounds that are not very well handled by the automatic segmentation method presented.

Naturally, the temporal alignment procedure needs a reliable time stretch/compress algorithm to work properly, specially when we consider the accuracy of the results. That is, to guarantee that the boundaries of the regions will be accurately aligned, we need a time stretch/compress algorithm that is capable of delivering the required accuracy. For instance, we cannot expect the time-aligned sounds to have the same number of samples if we are using a time stretch/compress algorithm that only guarantees accuracy down to the frame level.

Also, the temporal alignment procedure supposes that the time stretch/compress algorithm will give similar results when stretching and compressing the different portions of the sounds. This was not the case in this work, especially during the attack and transient parts. When longer attacks were compressed, the results were quite satisfactory. Time-stretching shorter attacks, however, produced more artificially sounding results. This does not constitute a problem when morphing because the temporally aligned sounds are never heard. They are just intermediaries used to combine the corresponding segments during the spectral morphing procedure. Given the perceptual salience of the attack in musical instrument sound perception, it might be beneficial to apply other time stretch/compress manipulation techniques that handle the attack transients differently.

15.5.3 Temporal Envelope Estimation

The temporal envelope is estimated twice, before and after temporal alignment of the sounds being morphed. Before temporal alignment, the estimated temporal envelope is used in the automatic segmentation step. After temporal alignment, the temporal envelope is morphed and the result is used to modulated the morphed frames. In both cases the RMS energy envelope was used in the estimation to account for temporal evolution of spectral energy.

The true amplitude envelope (TAE) introduced in this work gives very responsive and robust estimation of the amplitude envelope. However, estimations of amplitude take phase into consideration, and phase information has little perceptual impact. TAE can easily be adapted to calculate instantaneous energy envelope using the instantaneous energy signal $x^2(t)$ instead of the signal $x(t)$. But the instantaneous energy does not take perceptual effects such as the ear integration time into consideration. RMS was found to be the best choice of temporal envelope estimation algorithm.

15.5.4 Temporal Envelope Morphing

The representation of the temporal envelope is important in the temporal envelope morphing step. Analogously to the spectral envelope techniques investigated, this work proposed to represent the temporal envelope (obtained with any available estimation method) with different models. So the temporal envelope estimation and manipulation procedures are independent. This thesis investigated two temporal envelope morphing techniques, interpolation of the temporal envelope curves directly, and interpolation of the cepstral representation of the temporal envelope curve. The temporal centroid was used to evaluate the linearity of the temporal envelope morphing techniques.

It was found that the interpolation of the cepstral coefficients representing the temporal envelope generated intermediate temporal envelopes whose temporal centroid varies more linearly than simply interpolating the curves. Thus interpolation of the cepstral representation of the temporal envelope was adopted as the temporal envelope morphing procedure. It would be interesting to investigate the temporal envelope morphing procedure for sounds that present tremolo, for example.

15.6 Spectral Morphing

A perceptually important aspect of any sound morphing algorithm lies in how it represents spectral information and how it handles the spectral morphing task. The accuracy of representation is an important factor when we want to achieve high-quality results. On the other hand, the spectral morphing algorithm itself should produce morphed spectral envelopes that are perceived as a gradual transition between those of the source and target sounds.

15.6.1 Spectral Envelope Estimation

In the SF model, the spectral envelope models the filter. An accurate representation of the amplitudes of the partials for the sinusoidal component is very important to guarantee quality results. The residual component, however, requires a spectral envelope curve that follows the distribution of spectral energy without matching the spectral peaks. This thesis proposed to use "true envelope" to estimate the parameters of the spectral envelope of the sinusoidal component, and linear prediction for the residual component.

For the sinusoidal component, the filter represents the amplitudes of the partials independently from their frequencies. The robustness and accuracy of the "true envelope" estimation guaranteed a spectral envelope curve that matches well the amplitudes of the partials.

"True envelope" proved to be an appropriate choice also on the perceptual level. The residual, on the other hand, was found to be more dissimilar. The online similarity test asked participants to assess the similarity between the original recordings of musical instrument sounds and their SF representation. The aim of the test was to investigate whether the SF representation was perceptually similar to the original recordings, thus validating the model. The result confirmed that the representations are indeed similar, validating perceptually the SF representation of musical instrument sounds.

15.6.2 Spectral Envelope Morphing

This thesis presented a strong focus on the spectral morphing procedure and devoted a great deal of research effort to the investigation of different spectral envelope morphing algorithms. The main goal was to determine which spectral morphing algorithm leads to the most perceptually linear spectral transitions. The main spectral morphing algorithms proposed in the literature were compared. Among them, dynamic frequency warping, interpolation of the amplitudes of the partials, and interpolation of parameters of different spectral envelope representations. The representations compared were cepstral coefficients (CC), reflection coefficients (RC), line spectral frequencies (LSF), and linear prediction coefficients (LPC).

Linearity (or "spectral smoothness", as it was referred to in the listening test) was measured objectively and perceptually. The objective measure used the values of the spectral shape features, which measure the distribution of spectral energy and were found to be correlated with timbre perception in psychoacoustic studies. The perceptual evaluation, on the other hand, asked participants to compare pairs of transformations that only differed in the spectral morphing algorithm used.

The result of the objective spectral linearity evaluation indicated that interpolation of line spectral frequencies (LSF) are the spectral envelope morphing procedure that leads to the variation of spectral shape features closest to linear. On the other hand, the result of the "spectral smoothness" test was inconclusive. The participants reported that the task was too difficult and it was verified that, in fact, it relied too heavily on memory.

15.7 Sound Material

All the samples used in this work were taken from the Vienna symphonic library sound database, which is generally considered very high quality. The samples were recorded in controlled conditions to be used in (sample based) synthesizers. So it is possible to select a set of sounds that corresponds to rigid criteria. In this work, the focus on timbre demands several musical instruments to be compared. The sounds should be equalized in pitch, loudness (dynamics), and duration to allow comparison of timbral features due to temporal and spectral differences only.

I selected the dullest versions whenever possible so the differences would be due to spectral envelope (color as defined by Slawson) and attack times mainly. This means that the sounds had no vibrato (except the strings) or other ornaments. There were different attacks for most instruments (such as long, normal, and staccato) and different durations. Normal attack and short duration were selected. All the pairs had the same pitch (C3 or C4) and dynamics (forte or fortissimo). Especially for the "spectral smoothness" (linearity) listening test, I wanted the differences to be due mainly to spectral envelope changes to allow comparison between the two algorithms. Such controlled experimental setup had a negative impact for some participants. In musical contexts, expressivity plays an essential role, so the performance of model should be evaluated with expressive sounds.

15.8 Formalization

One of the main contributions of this thesis lies in the formalization of morphing to standardize musical instrument sound morphing, address the lack of consensus in the literature about many aspects of morphing, and help promote communication in the morphing community. This work proposed a standardization of nomenclature, algorithm, morphing transformations, and more importantly, evaluation.

A standardized terminology will definitely promote communication by establishing common grounds for future research. We need to make sure we are talking about the same things when we are using the same terms to further the scientific knowledge in any field, otherwise most of the effort will gravitate around trying to understand each other. This applies to the more technical issues as well. We need a clear terminology to address specific steps of the morphing process and the results.

The most important contribution of this thesis concerns the formalization of the evaluation. After a thorough review of the literature, it became evident that most works about morphing proposed models and techniques, bu rarely addressed the formal evaluation of the results. Therefore, this work proposed a formal framework to evaluate morphing results according to three criteria, correspondence, intermediateness, and smoothness.

15.8.1 Correspondence

Correspondence is key to a quality morph. When there is no correspondence, the unmatched feature tends to stand out during the transformation. In general, it is very difficult to guarantee correspondence in every step of the morphing algorithm. There are many levels to be considered, and correspondence is necessary in all of them. For example, correspondence between sound objects (notes of a melody, sound events in a soundscape, etc), between perceptually salient events during the course of the sound, between spectral peaks, etc. In this thesis, each of those levels was addressed.

First of all, this work investigates morphing between isolated sounds, which guarantees correspondence between sonic events. Temporal correspondence was approached as correspondence between perceptually salient events during the course of the sound, such as attack, steady state, etc. This specific problem was the subject of the temporal segmentation procedure developed, explained in chapters 8 and 12. Naturally, the temporal segmentation model adopted, called ACT, applies to a restricted class of musical instruments, namely sustained instruments.

Spectral correspondence was quite successfully achieved with the SF model. Sinusoidal models call for correspondence of number of partials, due to the intrinsic representation of each partial individually. For quasi-harmonic musical instrument sounds, correspondence between partials can be established with partial number. But that naturally restricts the result to contain only partials that have a match. The SF model solves elegantly the correspondence between partials by adopting a spectral envelope curve to represent the amplitudes of the partials. In the SF model, spectral correspondence is guaranteed as long as both spectral envelopes are represented with the same order (number of coefficients).

Moreover, the spectral envelope representation solves intrinsically the question about correspondence of spectral envelope peaks during the morphing transformation. The spectral envelope morphing technique used leads to a particular behavior of the spectral peaks, without the need to explicitly establishing correspondence between the formant peaks of the spectral envelope curve. Many different representations of the spectral envelope were investigated (envelope curve, LPC, CC, LSF, RC, CC). Two paradigms were proposed to explain the behavior of the spectral peaks during morphing, spectral peak shift (PS), and spectral peak rise and wane (RW). Cepstral representations were found to present RW behavior in general, while linear prediction representations

tend to present PS behavior.

Finally, the spectral envelope morphing technique proposed in this work only addresses the magnitude of the spectral representation. The frequency values of the partials in the morphed spectrum can depend on the correspondence between the number of partials. In this work, the sounds used in the morph were restricted to be quasi-harmonic, which means that we can use a simple approximation of the frequency of any partial using the fundamental frequency and the partial number for unmatched partials. Use of the interval in cents between the frequency of the matched partials makes it easy to handle the interpolation between quasi-harmonic and slightly inharmonic spectra, such as the piano. In this case, the frequency of the morphed partial will be a fraction of the interval in cents between the partials being combined.

15.8.2 Intermediateness

Intermediateness can become a complex issue conceptually. What needs to be measured or included when evaluating morphed sounds to guarantee intermediatess? This question was only partially addressed in this work because intermediateness was only considered from a spectral point of view. To guarantee temporal intermediateness, events such as attack time must be taken into consideration. Intermediateness for the attack times was implicitly assumed to result from the temporal alignment procedure. By using the log attack times in the temporal alignment procedure, each step would be perceptually linear because attack time was found to be perceived logarithmically [Caclin et al., 2005, Krimphoff et al., 1994, Luce and Clark, 1965, McAdams et al., 2005, Krumhansl, 1989, Grey and Gordon, 1977].

Spectral intermediateness was measured via the values of the spectral shape features. The variation of the spectral shape features was compared across spectral envelope morphing techniques with the aim of determining which technique leads to the most linear variation when the morphing factor varies linearly. A simple quadratic error measure was used after a fruitless attempt to use an analytic formulation of the relationship between the parameters of a spectral envelope model and the associated spectral envelope curve. The minimum quadratic error across several traditional instruments revealed that interpolation of line spectral frequencies (LSFs) outperforms all the others in average, and individually for most cases tested. The conclusion was that interpolation of LSFs is the best spectral envelope morphing strategy tested.

Partial frequency intermediateness was obtained by interpolating the interval in cents between each pair of partials. In this case, even when interpolating between musical instruments that present slightly inharmonic spectra, such as the piano, intermediate steps would be gradually more or less inharmonic. The same goes for effects that appear in the temporal variation of the frequencies of the partials, such as vibrato. These would gradually appear or disappear. Naturally, the inclusion of a specific model of vibrato would make the transitions between vibrato and no vibrato more gradual.

15.8.3 Smoothness

Smoothness was the term adopted to refer to the gradual change required when the morphing factor varies gradually. When combined with intermediateness, the much stronger constraint of linearity is achieved. Smoothness was investigated using the values of the spectral shape features and perceptually. One of the listening tests asked participants to compare the "smoothness" between two morphing algorithms. The sounds were annotated by hand to guarantee that the results of the automatic segmentation procedure would not interfere. The sounds were temporally aligned before the application of the algorithms, so the differences between the morphs were mainly due to the spectral morphing step. The algorithms compared used the SF and sinusoidal model for the spectral morphing procedure.

The results of the comparison was rather inconclusive. The preference for each pair of sounds varied significantly across participants, and was close to half in average. Most participants found the task too difficult, which might partially explain why the test failed to provide a definitive answer.

15.9 Memory Effect

One surprising conclusion was that the cyclostationary morph is perceived differently depending on the direction of the transformation. That is, we hear the sequences from A to B and from B to A differently even though they have the same sounds. One of the participants reported that both algorithms have the tendency to be nonlinear, increasing the rate of change toward the end of the transformation. Interestingly, it was empirically verified that this phenomenon is somewhat linked to a memory effect, whereby the transition changes more radically near the end independent of the order of presentation of the sounds. This memory effect should definitely be explored to study the cognitive mechanisms associated with musical instrument sound perception and representation.

15.10 Categorical perception

Sound morphing could help answer one intriguing question about musical instrument sound perception: "Is musical instrument sound perception categorical or continuous?" We might never be able to achieve perceptually linear transformations if the answer is categorical simply because perceptually linear transformations require continuous perception. An important part of the evaluation of the linearity of the morphs was done using the acoustic correlates of timbres spaces proposed by McAdams and others (spectral centroid, log attack time, etc) [Grey and Gordon, 1977, Krimphoff et al., 1994, Krumhansl, 1989, McAdams et al., 2005]. A very important consequence of guiding the morph using the acoustic correlates of timbre spaces is the possibility to investigate whether the morphed sounds validate the correlates of timbre dimensions. In this case the question to be answered is "Would (morphed) sounds with intermediate values of correlates be placed in intermediate positions in the underlying timbre space?" We could probably repeat the MDS space using the original and the morphed sounds and check the resultant space.

15.11 Conceptual Distance

The idea of conceptual distance in morphing, briefly introduced in chapter 2, may be explored further. Chapter 2 states that there is an inversely proportional relationship between the quality of the morph and the conceptual distance between the objects being morphed. However true for most cases, this is hardly a precise description of the whole morphing scenario. There are two key elements to consider, the initial conceptual distance between the objects, and the conceptual distances between the morph and the original objects.

When the objects being morphed are initially far apart conceptually (or perceptually), we need more steps to fill in the space between them equally. That is, to have the same distance across steps as when they are closer. Naturally, if we want each division to be one centimeter long, one meter will have more divisions than, say, ten centimeters. However, when the objects being morphed are initially close together, the more steps it takes to transition between source and target, the less they seem to change across steps. This was clearly the case for some of the transitions in the listening test. One participant remarked that "in many cases, it didn't seem like the morph was changing the sound very much." This is probably due to the choice of sound material (no expressivity, etc)

and focus on timbral features. That is, in many cases, there wasn't a big difference between the source and target sounds to begin with.

15.12 Timbre Spaces and Morphing

Another interesting perceptual phenomenon to be explored concerns the morphing space, where the original and the morphed objects reside. However, there is a much more difficult question about the different possible hybrids concerning their conceptual distance. It is known that perceptual and conceptual spaces are usually nonmetric, which means that the sum of the distances between a perceptually intermediate objects and the objects combined to produce it are not equal. Let us explore further this question with an example. In figure 15.1 we see that the distance between the man and the horse (d1) is much larger than the sum of the distances between the man and the centaur (d2) and the horse and the centaur (d3). In other words, the triangle inequality does not hold in this case, and the perceptual/conceptual space is nonmetric.

15.12.1 Is Perception of Morphing Metric?

What is the result of the morph between a man and a horse? Is it a centaur? In chapter 2 we saw that there are different possible hybridization processes. A centaur uses the natural hybridization process that takes parts from the the man and from the horse to compose the result. In morphing, on the other hand, each part should be the result of the combination of the corresponding parts from the man and from the horse.

The question about whether the perceptual/conceptual space in which these combinations exist is metric or not is extremely relevant in the context of morphing. For example, an intriguing question is: Is there any other hybrid for which the space is metric? Or, alternatively, if we morph between the man and the horse, is the resultant space metric? In this thesis we assumed that the morph follows a straight line. Therefore, implicitly we assume that the morph space is metric, just like the underlying MDS timbre spaces (shown in figure 15.2) used to guide the morphing transformation. If we use the acoustic correlates of timbre dimensions from these spaces (log attack time, spectral centroid, etc) to guide the morph, will the morphed sounds follow a straight line between them in the underlying timbre space?

The question of similarity of objects in (non)metric spaces is crucial in multimedia databases and information retrieval. For sounds, music information retrieval is a possible candidate. Interestingly, the cross evaluation of the feature values under the morphing transformation could also be used to validate the features themselves in MDS studies, for example. But this is unfortunately also out of the scope of this work.

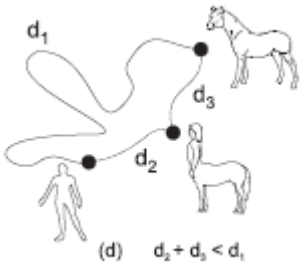

Figure 15.1: Nonmetric representation of conceptual objects. The figure shows a man, a horse, and a centaur, a mental representation of a hybrid between the man and the horse. The figure illustrates the case when the conceptual space in which the mental representations of the three is nonmetric.

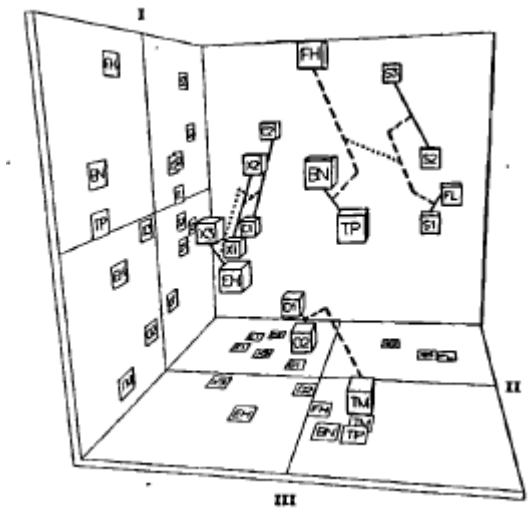

Figure 15.2: Example of multidimensional timbre spaces. After Grey [Grey and Gordon, 1977]

Bibliography

[Ahmad et al., 2009] Ahmad, M., Hacihabiboglu, H., and Kondoz, A. (2009). Morphing of transient sounds based on shift-invariant discrete wavelet transform and singular value decomposition. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*.

[Alexa and Müller, 1999] Alexa, M. and Müller, W. (1999). The morphing space. In *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, pages 329–336, Plzen, Czech Republic.

[Amatriain et al., 2003] Amatriain, X., Bonada, J., Álex Loscos, Arcos, J. L., and Verfaille, V. (2003). Content-based transformations. *Journal of New Music Research*, 32(1):95–114.

[Amatriain et al., 2002] Amatriain, X., Bonada, J., Loscos, A., and Serra, X. (2002). Spectral processing. In Zolzer, U., editor, *DAFX - Digital Audio Effects*, chapter 10, pages 373–438. John Wiley and Sons.

[ANSI, 1960] ANSI (1960). Usa standard acoustical terminology (including mechanical shock and vibration) sl.1-1960 (r1976).

[Athineos and Ellis, 2003] Athineos, M. and Ellis, D. P. W. (2003). Frequency-domain linear prediction for temporal features. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*.

[Aucouturier et al., 2005] Aucouturier, J.-J., Pachet, F., and Sandler, M. (2005). The way it sounds: Timbre models for analysis and retrieval of polyphonic music signals. *IEEE Transactions of Multimedia*, 7(6):1028–1035.

[Backström and Magi, 2006] Backström, T. and Magi, C. (2006). Properties of line spectrum pair polynomials - a review. *Signal Processing*, 86:3286–3298.

[Bello et al., 2005] Bello, J., Daudet, L., Abdullah, S., Duxbury, C., Davies, M., and Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5).

[Boccardi and Drioli, 2001] Boccardi, F. and Drioli, C. (2001). Sound morphing with gaussian mixture models. In *Proceedings of the International Conference on Digital Audio Effects*, pages 44–48.

[Bogert et al., 1963] Bogert, B., Healy, M., and Tukey, J. (1963). The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In Rosenblatt, M., editor, *Time Series Analysis*, chapter 15, pages 209–243. New York: Wiley.

[Bregman, 1990] Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts.

[Brown, 1998] Brown, J. C. (1998). Musical instrument identification using autocorrelation coefficients. In *Proceedings of the International Symposium on Musical Acoustics*.

[Brown, 1999] Brown, J. C. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *Journal of the Acoustical Society of America*, 105(3):1933–41.

[Burred et al., 2010] Burred, J., Röbel, A., and Sikora, T. (2010). Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3).

[Caclin et al., 2005] Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1):471–482.

[Caetano and Rodet, 2009] Caetano, M. and Rodet, X. (2009). Evolutionary spectral envelope morphing by spectral shape descriptors. In *Proceedings of the International Computer Music Conference*, Montreal, Canada.

[Caetano and Rodet, 2010a] Caetano, M. and Rodet, X. (2010a). Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In *Proceedings of the International Conference on Digital Audio Effects*, Graz, Austria.

[Caetano and Rodet, 2010b] Caetano, M. and Rodet, X. (2010b). Automatic timbral morphing of musical instrument sounds by high-level descriptors. In *Proceedings of the International Computer Music Conference*, New York/Stony Brook, USA.

[Caetano and Rodet, 2010c] Caetano, M. and Rodet, X. (2010c). Independent manipulation of high-level spectral envelope shape features for sound morphing by means of evolutionary computation. In *Proceedings of the International Conference on Digital Audio Effects*, Graz, Austria.

[Caetano and Rodet, 2011a] Caetano, M. and Rodet, X. (2011a). Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing. In *Proceedings of the International Conference on Audio, Speech, and Signal Processing*, Prague, Czech Republic.

[Caetano and Rodet, 2011b] Caetano, M. and Rodet, X. (2011b). Sound morphing by feature interpolation. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic.

[Cappé et al., 1995] Cappé, O., Laroche, J., and Moulines, E. (1995). Regularized estimation of cepstrum envelope from discrete frequency points. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Cappé and Moulines, 1996] Cappé, O. and Moulines, E. (1996). Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4):100–102.

[Childers et al., 1977] Childers, D. G., Skinner, D. P., and Kemerait, R. C. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443.

[Cohen, 1970] Cohen, T. (1970). Sourcedepth determinations using spectral, pseudoautocorrelation and cepstral analysis. *Geophysics Journal of the Royal Astronomical Society*, 20:223–231.

[Cooley and Tukey, 1965] Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301.

[Cramér, 1945] Cramér, H. (1945). *Mathematical Methods of Statistics*. Princeton University Press.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

[D'haes and Rodet, 2003] D'haes, W. and Rodet, X. (2003). Discrete cepstrum coefficients as perceptual features. In *Proceedings of the International Computer Music Conference*.

[Dolson, 1986] Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27.

[El-Jaroudi and Makhoul, 1969] El-Jaroudi, A. and Makhoul, J. (1969). Discrete all-pole modeling. *IEEE Transactions on Communication Technolology*, COM-17:481–488.

[Ezzat et al., 2005] Ezzat, T., Meyers, E., Glass, J., and Poggio, T. (2005). Morphing spectral envelopes using audio flow. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*.

[Fant, 1960] Fant, G. (1960). *Acoustic theory of speech production*. Mouton, The Hague, Netherlands, 2nd edition.

[Fitz and Haken, 1996] Fitz, K. and Haken, L. (1996). Sinusoidal modeling and manipulation using lemur. *Computer Music Journal*, 20(4):44–59.

[Fitz et al., 2003] Fitz, K., Haken, L., Lefvert, S., Champion, C., and O'Donnel, M. (2003). Cell-utes and flutter-tongued cats: Sound morphing using loris and the reassigned bandwidth-enhanced model. *Computer Music Journal*, 27(3):44–65.

[Flanagan and Golden, 1966] Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. *Bell System Technical Journal*, pages 1493–1509.

[Fletcher, 1934] Fletcher, H. (1934). Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *Journal of the Acoustical Society of America*, 6:59–69.

[Fletcher and Rossing, 1998] Fletcher, N. H. and Rossing, T. D. (1998). *The Physics of Musical Instruments*. Springer, New York, 23nd ed edition.

[Foote, 1997] Foote, J. T. (1997). Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II, Proceedings of SPIE*, pages 138–147.

[Galas and Rodet, 1990] Galas, T. and Rodet, X. (1990). An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals. In *Proceedings of the International Computer Music Conference*.

[Grey, 1975] Grey, J. M. (1975). *An Exploration of Musical Timbre*. PhD thesis, Stanford University. Dept. of Psychology, Dept. of Music Report STAN-M-2.

[Grey and Gordon, 1977] Grey, J. M. and Gordon, J. W. (1977). Multidimensional perceptual scaling of musical timbre. *Journal of the Acoustical Society of America*, 61(5):1270–1277.

[Grey and Gordon, 1978] Grey, J. M. and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63(5):1493–1500.

[Hahn et al., 2010] Hahn, H., Röbel, A., Burred, J. J., and Weinzierl, S. (2010). A source-filter model for quasi-harmonic instruments. In *Proceedings of the International Conference on Digital Audio Effects*.

[Hajda, 1996] Hajda, J. (1996). A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited. In *Audio Engineering Society Convention 101*.

[Haken et al., 2006] Haken, L., Fitz, K., and Christensen, P. (2006). Beyond traditional sampling synthesis: Real-time timbre morphing using additive synthesis. In Beauchamp, J. W., editor, *Sound of Music: Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin.

[Handel, 1995] Handel, S. (1995). Timbre perception and auditory object identification. In Moore, B., editor, *Hearing*, pages 425–461. Academic Press, New York.

[Hartmann, 1978] Hartmann, W. (1978). The effect of amplitude envelope on the pitch of sine wave tones. *Journal of the Acoustical Society of America*, 63:1105–1113.

[Hartmann, 2007] Hartmann, W. (2007). Acoustic signal processing. In Rossing, T. D., editor, *Springer Handbook of Acoustics*, pages 503–530. Springer.

[Hartmann, 1998] Hartmann, W. M. (1998). *Signals, Sound, and Sensation*. Springer-Verlag, New York, NY.

[Harvey, 1981] Harvey, J. (1981). "Mortuos Plango, Vivos Voco": A realization at ircam. *Computer Music Journal*, 5(4):22–24.

[Hassab and Boucher, 1976] Hassab, J. and Boucher, R. (1976). A probabilistic analysis of time delay extraction by the cepstrum in stationary gaussian noise. *IEEE Transactions on Information Theory*, 22(4):444–454.

[Hatch, 2004] Hatch, W. (2004). High-level audio morphing strategies. Master's thesis, Music Technology Dept., McGill University.

[Helmholtz, 1885] Helmholtz, H. v. (1885). *On the Sensations of Tone*. Longman, London.

[Herrera et al., 1999] Herrera, P., Serra, X., and Peeters, G. (1999). Audio descriptors and descriptors schemes in the context of mpeg-7. In *Proceedings of the International Computer Music Conference*.

[Hikichi, 2001] Hikichi, T. (2001). Sound timbre interpolation based on physical modelling. *Acoustical Science and Technology*, 22(2):101–111.

[Hope and Furlong, 1997] Hope, C. and Furlong, D. (1997). Time-frequency distributions for timbre morphing: The wigner distribution versus the stft. In *Proceedings of the Brazilian Symposium of Computer Music*.

[Hope and Furlong, 1998] Hope, C. J. and Furlong, D. J. (1998). Endemic problems in timbre morphing processes: Causes and cures. In *Proceedings of the Irish Signals and Systems Conference*.

[Horner et al., 2009] Horner, A. B., Beauchamp, J. W., and So, R. H. Y. (2009). Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones. *Journal of the Acoustical Society of America*, 125(1):882–897.

[Itakura, 1975] Itakura, F. (1975). Line spectrum representation of linear prediction coefficients of speech signals. *Journal of the Acoustical Society of America*, 57:S35.

[Itakura and Saito, 1968] Itakura, F. and Saito, S. (1968). Analysis synthesis telephony based on the maximum likelihood. In *Proceedings of the International Congress on Acoustics*, pages C17–C20.

[Itakura and Saito, 1970] Itakura, F. and Saito, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communication in Japan*, 53-A:36–43.

[ITU-T Recommendation P.800, 1996] ITU-T Recommendation P.800 (1996). Methods for subjective determination of transmission quality. ITU-T SG12.

[Jensen, 1999] Jensen, K. (1999). Envelope model of isolated musical sounds. In *Proceedings of the International Conference on Digital Audio Effects*.

[Jensen, 2001] Jensen, K. (2001). The timbre model. In *Workshop on Current Research Directions in Computer Music*.

[Kaplan and Ulrych, 2007] Kaplan, S. T. and Ulrych, T. J. (2007). Phase unwrapping: a review of methods and a novel technique. In *Proceedings of Canadian Society of Exploration Geophysics Convention*.

[Kemerait, 1971] Kemerait, R. (1971). *Signal Detection and Extraction by Cepstrum Techniques*. PhD thesis, University of Florida.

[Kemerait, 1972] Kemerait, R. (1972). Signal detection and extraction by cepstrum techniques. *IEEE Transactions on Information Theory*, IT-18:745–759.

[Klapuri et al., 2010] Klapuri, A., Virtanen, T., and Heittola, T. (2010). Sound source separation in monaural music signals using excitation-filter model and em algorithm. In *Proceedings of the International Conference on Audio, Speech, and Signal Processing*, pages 5510–5513.

[Kolassa, 2006] Kolassa, J. E. (2006). *Series Approximation Methods in Statistics*. New York: Springer Lecture Notes in Statistics 88, 3rd ed. edition.

[Konvalina and Matache, 2004] Konvalina, J. and Matache, V. (2004). Palindrome-polynomials with roots on the unit circle. *Comptes Rendus Mathematiques*, 26:39–44.

[Krimphoff et al., 1994] Krimphoff, J., McAdams, S., and Winsberg, S. (1994). Charactérisation du timbre des sons complexes. ii: Analyses acoustiques et quantification psychophysique. *Journal de Physique IV*, 4(1):C5.625–C5.628.

[Krumhansl, 1989] Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In Nielzén, S. and Olsson, O., editors, *Structure and perception of electroacoustic sound and music*, pages 43–54. Excerpta Medica, New York.

[Landy, 2011] Landy, L. (accessed in 02/2011). Sound transformations in electroacoustic music. `http://people.bath.ac.uk/masjpf/CDP/landyeam.htm`.

[Laroche and Meillier, 1998] Laroche, J. and Meillier, J. L. (1998). A simplified source/filter model for percussive sounds. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Laura and Rodet, 1990] Laura, C. and Rodet, X. (1990). Appariement des pics spectraux et règles pour la synthèse de la parole par concaténation de diphones. In *Journal de Physique Colloques*, page 50.

[LeBlanc, 1969] LeBlanc, L. R. (1969). Narrow-band sampled-data techniques for detection via the underwater acoustic communication channel. *IEEE Transactions on Communication Technology*, COM-17:481–488.

[Letowski, 1992] Letowski, T. (1992). Timbre, tone color, and sound quality: Concepts and definitions. *Archives of Acoustics*, 17(1):17–30.

[Luce and Clark, 1965] Luce, D. and Clark, Jr., M. (1965). Durations of attack transients of nonpercussive orchestral instruments. *Journal of the Audio Engineering Society*, 13(3):194–199.

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63:561–580.

[Markel and Gray, 1976] Markel, J. D. and Gray, A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag, New York.

[Martin and Kim, 1998] Martin, K. D. and Kim, Y. E. (1998). Musical instrument identification: A pattern-recognition approach. In *Proceedings of the 136th Meeting of the Acoustical Society of America*.

[McAdams et al., 1999] McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, 105(2):492–502.

[McAdams et al., 2006] McAdams, S., Bruno, G., Susini, P., Peeters, G., and Rioux, V. (2006). A meta-analysis of acoustic correlates of timbre dimensions (a). *The Journal of the Acoustical Society of America*, 120(5):3275–3275.

[McAdams et al., 2005] McAdams, S., Winsberg, S., Donnadieu, S., de Soete, G., and Krimphoff, J. (2005). Perceptual scaling of synthesized musical timbres: Common dimensions, specifities and latent subject classes. *Psychological Research*, 58:177–192.

[McAulay and Quatieri, 1986] McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34:744–754.

[McAuley, 1984] McAuley, R. (1984). Maximum likelihood spectral estimation and its application to narrow-band speech coding. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, pages 243–251.

[McLoughlin, 2008] McLoughlin, I. V. (2008). Review: Line spectral pairs. *Signal Processing*, 88(3):448–467.

[McNabb, 1981] McNabb, M. (1981). "Dreamsong": The composition. *Computer Music Journal*, 5(4):36–53.

[Misdariis et al., 1998] Misdariis, N., Smith, B., Pressnitzer, D., Susini, P., and McAdams, S. (1998). Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. *The Journal of the Acoustical Society of America*, 103(5):3005–3006.

[Moorer, 1977] Moorer, J. A. (1977). Signal processing aspects of computer music: A survey. *Proceedings of the IEEE*, 65:1108–1137.

[Moorer, 1978] Moorer, J. A. (1978). The use of linear prediction of speech in computer music applications. *Journal of the Audio Engineering Society*, 26(1).

[Moorer, 1979] Moorer, J. A. (1979). The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society*, 27(3):134–140.

[Moorer and Grey, 1977a] Moorer, J. A. and Grey, J. M. (1977a). Lexicon of analyzed tones (part 1: A violin tone). *Computer Music Journal*, 1(2):39–45.

[Moorer and Grey, 1977b] Moorer, J. A. and Grey, J. M. (1977b). Lexicon of analyzed tones (part 2: Clarinet and oboe tones). *Computer Music Journal*, 1(3):12–29.

[Moorer and Grey, 1977c] Moorer, J. A. and Grey, J. M. (1977c). Lexicon of analyzed tones (part 3: The trumpet). *Computer Music Journal*, 2(2):23–31.

[Morris and Clements, 2002] Morris, R. and Clements, M. (2002). Modification of formants in the line spectrum domain. *IEEE Signal Processing Letters*, 9(1).

[Noll, 1973] Noll, A. (1973). The cepstrum and some close relatives. In Griffiths, J. W. R., Stocklin, P. L., and Schooneveld, C. V., editors, *Signal Processing*, pages 11–22. London: Academic Press.

[Noll, 1964] Noll, A. M. (1964). Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection. *Journal of the Acoustical Society of America*, 36(2):296–302.

[Noll, 1967] Noll, A. M. (1967). Cepstrum pitch determination. *Journal of the Acoustical Society of America*, 41(2):293–309.

[Oppenheim, 1969] Oppenheim, A. (1969). Speech analysis-synthesis system based on homomorphic filtering. *Journal of the Acoustical Society of America*, 45(2):458–465.

[Oppenheim and Schafer, 1968] Oppenheim, A. and Schafer, R. (1968). Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, AU-16:221–226.

[Oppenheim et al., 1968] Oppenheim, A., Schafer, R., and T.G. Stockham, J. (1968). Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56:1264–1291.

[Oppenheim, 1964] Oppenheim, A. V. (1964). *Superposition in a Class of Nonlinear Systems*. PhD thesis, MIT.

[Oppenheim and Schafer, 2004] Oppenheim, A. V. and Schafer, R. W. (2004). From frequency to quefrency: a history of the cepstrum. *Signal Processing Magazine IEEE*, 21(5):95–106.

[Osaka, 1995] Osaka, N. (1995). Timbre interpolation of sounds using a sinusoidal model. In *Proceedings of the International Computer Music Conference*.

[Osaka, 1998] Osaka, N. (1998). Timbre morphing and interpolation based on a sinusoidal model. In *Proceedings of the 16th International Congress on Acoustics and 135th meeting Acoustical Society of America*, volume 1, pages 83–84.

[Osaka, 2005] Osaka, N. (2005). Concatenation and stretch/squeeze of musical instrumental sound using morphing. In *Proceedings of the International Computer Music Conference*.

[Paliwal, 1992] Paliwal, K. (1992). On the use of line spectral frequency parameters for speech recognition. *Digital Signal Processing*, 2:80–87.

[Papoulis, 1991] Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes*. New York:McGraw-Hill, 3rd ed. edition.

[Peeters, 2003] Peeters, G. (2003). Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Audio Engineering Society Convention 115*.

[Peeters, 2004] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM.

[Peeters et al., 2000] Peeters, G., McAdams, S., and Herrera, P. (2000). Instrument sound description in the context of mpeg-7. In *Proceedings of the International Computer Music Conference*.

[Peeters and Rodet, 2002] Peeters, G. and Rodet, X. (2002). Automatically selecting signal descriptors for sound classification. In *Proceedings of the International Computer Music Conference*.

[Peeters and Rodet, 2003] Peeters, G. and Rodet, X. (2003). Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. In *Proceedings of the International Conference on Digital Audio Effects*.

[Pfitzinger, 2004] Pfitzinger, H. (2004). Dfw-based spectral smoothing for concatenative speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, volume 2, pages 1397–1400.

[Pinch and Trocco, 2002] Pinch, T. and Trocco, F. (2002). *Analog Days: The Invention and Impact of the Moog Synthesizer*. Harvard University Press, Cambridge, MA.

[Plomp, 1966] Plomp, R. (1966). *Experiments On Tones Perception*. Van Gorcum Comp. N.V, Netherlands.

[Plomp and J.M.Steeneken, 1969] Plomp, R. and J.M.Steeneken (1969). Effect of phase on the timbre of complex tones. *Journal of the Acoustical Society of America*, 46:409–421.

[Potamianos and Maragos, 1994] Potamianos, A. and Maragos, P. (1994). A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 17(1):95–120.

[Pratt and Doak, 1976] Pratt, R. and Doak, P. (1976). A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45:317–328.

[Quatieri and McAuley, 2002] Quatieri, T. and McAuley, R. (2002). Audio signal processing based on sinusoidal analysis/synthesis. In Kahrs, M. and Brandenburg, K., editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 9, pages 343–416. Kluwer Academic Publishers.

[Rabiner, 1993] Rabiner, L. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice Hall.

[Risset and Mathews, 1969] Risset, J. and Mathews, M. (1969). Analysis of instrument tones. *Physics Today*, 22(2):23–30.

[Risset, 1966] Risset, J.-C. (1966). Computer study of trumpet tones (with sound examples on tape). Technical report, Bell Laboratories, Murray Hill, N.J.

[Risset and Wessel, 1982] Risset, J.-C. and Wessel, D. (1982). Exploration of timbre by analysis and synthesis. In Deustch, D., editor, *The Psychology of Music*, pages 26–58. Academic Press, Orlando, FL.

[Röbel, 1998] Röbel, A. (1998). Morphing dynamical sound models. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*.

[Röbel, 2003] Röbel, A. (2003). A new approach to transient processing in the phase vocoder. In *Proceedings of the International Conference on Digital Audio Effects*, pages 344–349.

[Röbel and Rodet, 2005] Röbel, A. and Rodet, X. (2005). Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proceedings of the International Conference on Digital Audio Effects*, pages 30–35.

[Röbel et al., 2007] Röbel, A., Villavicencio, F., and Rodet, X. (2007). On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11):1343–1350.

[Rodet and Depalle, 1992] Rodet, X. and Depalle, P. (1992). A new additive synthesis method using inverse fourier transform and spectral envelopes. In *Proceedings of the International Computer Music Conference*, San Jose, California.

[Russel, 2009] Russel, R. (2009). A sex difference in facial contrast and its exaggeration by cosmetics. *Perception*, 38(8):1211–1219.

[Schaeffer, 1966] Schaeffer, P. (1966). *Traité des objets Musicaux*. Paris: Seuil.

[Schafer, 1968] Schafer, R. (1968). *Echo Removal by Discrete Generalized Linear Filtering*. PhD thesis, MIT.

[Schafer and Rabiner, 1970] Schafer, R. and Rabiner, L. (1970). System for automatic formant analysis of voiced speech. *Journal of the Acoustical Society of America*, 47:634–648.

[Scheirer and Slaney, 1997] Scheirer, E. and Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, pages 1331–1334.

[Schloss, 1985] Schloss, A. (1985). *On the Automatic Transcription of Percussive Music-From Acoustic Signal to High-Level Analysis*. PhD thesis, Stanford University.

[Schroeder, 1980] Schroeder, M. R. (1980). Direct (nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(2):297–301.

[Schröeder, 1999] Schröeder, M. R. (1999). *Computer speech: recognition, compression, synthesis*. Springer-Verlag New York, Inc., New York, NY, USA.

[Schubert and Wolfe, 2006] Schubert, E. and Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica United with Acustica*, 92:820–825.

[Schwarz and Rodet, 1999] Schwarz, D. and Rodet, X. (1999). Spectral envelope estimation and representation for sound analysis/synthesis. In *Proceedings of the International Computer Music Conference*.

[Seashore, 1938] Seashore, C. E. (1938). *Psychology of Music*. Dover Publications, New York.

[Serra and Bonada, 1998] Serra, X. and Bonada, J. (1998). Sound transformations based on the sms high level attributes. In *Proceedings of the Digital Audio Effects Workshop*.

[Serra and Smith, 1990] Serra, X. and Smith, J. O. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):49–56.

[Skowronek and McKinney, 2006] Skowronek, J. and McKinney, M. (2006). Features for audio classification: Percussiveness of sounds. In Verhaegh, W., Aarts, E., and Korst, J., editors, *The language of electroacoustic music*, pages 103–118. Springer Netherlands.

[Slaney et al., 1996] Slaney, M., Covell, M., and Lassiter, B. (1996). Automatic audio morphing. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*.

[Slawson, 1985] Slawson, W. (1985). *Sound Color*. University of California Press, Berkerley.

[Smalley, 1986] Smalley, D. (1986). Spectromorphology and structuring processes. In Emmerson, S., editor, *Intelligent Algorithms in Ambient and Biomedical Computing*, pages 61–93. London: Macmillan.

[Smalley, 1997] Smalley, D. (1997). Spectromorphology: Explaining sound shapes. *Organized Sound*, 2(2):107–126.

[Smith, 2011] Smith, J. O. (accessed 04/2011). *Spectral Audio Signal Processing, May 2010 Draft*. http://ccrma.stanford.edu/~jos/sasp/ online book.

[Smith and Serra, 1985] Smith, J. O. and Serra, X. (1985). Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation.

[Soong and Juang, 1984] Soong, F. and Juang, B. (1984). Line spectrum pair (lsp) and speech data compression. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*.

[Steiglitz and Dickinson, 1977] Steiglitz, K. and Dickinson, B. (1977). Computation of the complex cepstrum by factorization of the z-transform. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, pages 723–726.

[Stevens and Volkman, 1940] Stevens, S. S. and Volkman, J. (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology*, 53:329–353.

[Stevens et al., 1937] Stevens, S. S., Volkman, J., and Newman, E. (1937). A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, 8(3):185–190.

[Stylianou, 2008] Stylianou, Y. (2008). Voice transformation. In Benesty, J., Sondhi, M. M., and Huang, Y. A., editors, *Springer Handbook of Speech Processing*, pages 489–504. Springer.

[Tellman et al., 1995] Tellman, E., Haken, L., and Holloway, B. (1995). Morphing between timbres with different numbers of features. *Journal of the Audio Engineering Society*, 43(9):678–689.

[Terasawa et al., 2005] Terasawa, H., Slaney, M., and Berger, J. (2005). The thirteen colors of timbre. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Tribolet, 1977] Tribolet, J. (1977). A new phase unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25:170–177.

[Tzanetakis and Cook, 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5).

[Verfaille et al., 2006] Verfaille, V., Zölzer, U., and Arfib, D. (2006). Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1817–1831.

[Villavicencio et al., 2006] Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving lpc spectral envelope extraction of voiced speech by true envelope estimation. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*.

[Villavicencio et al., 2007] Villavicencio, F., Rodet, X., and Röbel, A. (2007). On cepstral and all-pole based spectral envelope modeling with unknown order. *Pattern Recognition Letters*, 28:1343–1350.

[Wen and Sandler, 2010] Wen, X. and Sandler, M. (2010). Source-filter modeling in the sinusoidal domain. *Journal of the Audio Engineering Society*, 58(10).

[Williams and Brookes, 2007] Williams, D. and Brookes, T. (2007). Perceptually motivated audio morphing: brightness. In *Audio Engineering Society, 122nd convention*.

[Williams and Brookes, 2009] Williams, D. and Brookes, T. (2009). Perceptually motivated audio morphing: softness. In *Audio Engineering Society, 126nd convention*.

[Wishart, 1996] Wishart, T. (1996). *On Sonic Art*. On Sonic Art. Imagineering Press, simon emmerson edition.

[Wishart, 1997] Wishart, T. (1997). Soundhack. *Computer Music Journal*, 1(21):10–11.

[Wishart, 2011] Wishart, T. (accessed in 02/2011). Computer sound transformation. `http://www.trevorwishart.co.uk/transformation.html`.

[Wolberg, 1998] Wolberg, G. (1998). Image morphing: A survey. *The Visual Computer*, 14:360–372.

[Wold et al., 1996] Wold, E., Blum, T., Keislar, D., and Wheaton, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3:27–36.

Part IV

Appendices

Appendix A

Linear Prediction

Here we will see the stability of the all-pole filter, the spectral conversion characteristics, and the error analysis.

A.0.2 Filter Stability

After the predictor parameters are computed, the question of the stability of the resulting filter $H(z)$ arises. Filter stability is important for many applications. A causal all-pole filter is stable if all its poles lie inside the unit circle (in which case it is also a filter with minimum phase). The poles of $H(z)$ are simply the roots of the denominator polynomial $A(z)$, where

$$A(z) = 1 + \sum_{k=1}^{p} a(k) z^{-k} \tag{A.1}$$

and

$$H(z) = \frac{G}{A(z)} \tag{A.2}$$

$A(z)$ is also known as the *inverse filter*.

If the coefficients $R(i)$ in equation 7.16 are positive definite (which is assured if $R(i)$ is computed from a nonzero signal using equation 7.21 or from a positive definite spectrum, i.e., a spectrum that can be zero at most at a finite set of frequencies), the solution of the autocorrelation equation 7.16 gives predictor parameters which guarantee that all the roots of $A(z)$ lie inside the unit circle. In other words, it gives a stable $H(z)$ [Makhoul, 1975]. This result can also be obtained from orthogonal polynomial theory. In fact, if one denotes the inverse filter at step i in iteration 7.36 by $A_i(z)$, then it can be shown that the polynomials $z^i A_i(z)$ for $i = 0, 1, 2, \cdots$, form an orthogonal set over the unit circle, as expressed below

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) A_n\left(e^{j\omega}\right) A_m\left(e^{-j\omega}\right) e^{j(n-m)\omega} d\omega = E_n \delta_{nm}, \quad n, m = 0, 1, 2, \cdots \tag{A.3}$$

where E_n is the minimum error for an n^{th} order predictor, and $P(\omega)$ is any positive definite spectrum whose Fourier transform results in the autocorrelation coefficients $R(i)$ that are used in equation 7.16. The recurrence relation for these polynomials is as follows

$$A_i(z) = A_{i-1}(z) + k_i z^{-i} A_{i-1} z^{-1} \tag{A.4}$$

which is the same as the recursion in equation 7.38.

The positive definiteness of $R(i)$ can often be lost if one uses a small word length to represent $R(i)$ in a computer. Also, roundoff errors can cause the autocorrelation matrix to become illconditioned. Therefore, it is often necessary to check for the stability of $H(z)$. Checking if the roots of $A(z)$ are inside the unit circle is a costly procedure that is best avoided One method is to check if all the successive errors are positive. In fact, the condition $E_i > 0,\ 1 \leq i \leq p$, is a necessary and sufficient condition for the stability of $H(z)$. From equations 7.39 and 7.42 it is clear that an equivalent condition for the stability of $H(z)$ is that

$$|k_i| < 1,\ 1 \leq i \leq p \tag{A.5}$$

Therefore, the recursive procedure in equations 7.35 through 7.39 also facilitates the check for the stability of the filter $H(z)$.

The predictor parameters resulting from a solution to the covariance matrix equation 7.22 cannot in general be guaranteed to form a stable filter. The computed filter tends to be more stable as the number of signal samples N is increased, i.e., as the covariance matrix approaches an autocorrelation matrix. Given the computed predictor parameters, it is useful to be able to test for the stability of the filter $H(z)$. One method is to compute the reflection coefficients k_i from the predictor parameters by a backward recursion, and then check for stability using equation A.5. The recursion is as follows

$$\begin{gathered} k_i = a_i^{(i)} \\ a_j^{(i-1)} = \frac{a_j^{(i)} - a_i^{(i)} a i - j_i^{(i)}}{1 - k_i^2},\ 1 \leq j \leq i-1 \end{gathered} \tag{A.6}$$

where the index i takes values $p, p-1, \cdots, 1$ in that order. Initially, $a_j^{(p)} = a_j,\ 1 \leq j \leq i-1$. It is interesting to note that this method (or checking the stability of $H(z)$) is essentially the same as the Lehmer-Schur method [Makhoul, 1975] for testing whether or not the zeros of a polynomial lie inside the unit circle. An unstable filter can be made stable by reflecting the poles outside the unit circle inside, such that the magnitude of the system frequency response remains the same. Filter instability can often be avoided by adding a very small number to the diagonal elements in the covariance matrix.

A question always arises as to whether to use the autocorrelation method or covariance method in estimating the predictor parameters. The covariance method is quite general and can be used with no restrictions. The only problem is that of the stability of the resulting filter, which is not a severe problem generally. In the autocorrelation method, on the other hand, the filter is guaranteed to be stable, but problems of parameter accuracy can arise because of the necessity of windowing (truncating) the time signal. This is usually a problem if the signal is a portion of an impulse response. For example, if the impulse response of an all-pole filter is analyzed by the covariance method, the filter parameters can be computed accurately from only a finite number of samples of the signal. Using the autocorrelation method, one cannot obtain the exact parameter values unless the whole infinite impulse response is used in the analysis. However, in practice, very good approximations can be obtained by truncating the impulse response at a point where most of the decay of the response has already occurred.

A.1 Frequency Domain Formulations

In Section A, the stationary and nonstationary methods of linear prediction were derived from a time domain formulation. In this section we show that the same normal equations can be derived from a frequency domain formulation. It will become clear that linear prediction is basically a

correlation type of analysis which can be approached either from the time or frequency domain. The insights gained from the frequency domain analysis will lead to new applications for linear predictive analysis.

A.1.1 Stationary Case

The error e_n between the actual signal and the predicted signal is given by equation 7.11. Applying the z-transform to equation 7.11, we obtain

$$E(z) = \left[1 + \sum_{k=1}^{p} a(k) z^{-k}\right] S(z) = A(z) S(z) \tag{A.7}$$

where $A(z)$ is the inverse filter defined in equation A.1, and $E(z)$ and $S(z)$ are the z-transforms of $e(n)$ and $s(n)$, respectively. Therefore, $e(n)$ can be viewed as the result of passing $s(n)$ through the inverse filter $A(z)$. Assuming a deterministic signal $s(n)$, and applying Parseval's theorem, the total error to be minimized is given by

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|E\left(e^{j\omega}\right)\right|^2 d\omega \tag{A.8}$$

where $E\left(e^{j\omega}\right)$ is obtained by evaluating $E(z)$ on the unit circle $z = e^{j\omega}$. Denoting the power spectrum of the signal $s(n)$ by $P(\omega)$,where

$$P(\omega) = \left|S\left(e^{j\omega}\right)\right|^2 \tag{A.9}$$

we have from equations A.7 through A.9

$$E = \int_{-\pi}^{\pi} P(\omega) A\left(e^{j\omega}\right) A\left(e^{-j\omega}\right) d\omega \tag{A.10}$$

Following the same procedure as in Section 7.2.1.1, E is minimized by applying equation 7.13 to A.10. The result can be shown to be identical to the autocorrelation normal equations 7.16, but with the autocorrelation $R(i)$ obtained from the signal spectrum $P(\omega)$ by an inverse Fourier transform

$$R(i) = \int_{-\pi}^{\pi} P(\omega) \cos(i\omega) d\omega \tag{A.11}$$

Note that in equation A.11 the cosine transform is adequate since $P(\omega)$ is real and even. The minimum squared error E_p can be obtained by substituting equations 7.16 and A.10 in A.11, which results in the same equation as in 7.17.

A.1.2 Nonstationary Case

Here the signal s_n and the error e_n are assumed to be nonstationary. If $R(t, t')$ is the nonstationary autocorrelation of s_n, then we define the nonstationary two-dimensional (2D) spectrum $Q(\omega, \omega')$ of s_n by [Makhoul, 1975]

$$Q(\omega, \omega') = \sum_{t'=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} R(t, t') \exp\left[-j(\omega t - \omega' t')\right] \tag{A.12}$$

$R(t,t')$ can be recovered from $Q(\omega,\omega')$ by an inverse 2D Fourier transform

$$R(t,t') = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} Q(\omega,\omega') \exp\left[j\left(\omega t - \omega' t'\right)\right] d\omega d\omega' \tag{A.13}$$

As in the time domain formulation, we are interested in minimizing the error variance for time $n = 0$, which is now given by

$$E = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} Q(\omega,\omega') A\left(e^{j\omega}\right) A\left(e^{-j\omega}\right) d\omega d\omega' \tag{A.14}$$

Applying equation 7.13 to A.14 results in equations identical to the nonstationary normal equations 7.31, where $R(t,t')$ is now defined by equation A.13. The minimum error is then obtained by substituting 7.31 and A.13 in A.14. The answer is identical to 7.32.

A.1.3 Linear Predictive Spectral Matching

In this section we shall examine in what manner the signal spectrum $P(\omega)$ is approximated by the all-pole model spectrum, which we shall denote by $\hat{P}(\omega)$. From equation 7.9 and A.2

$$\hat{P}(\omega) = \left|H\left(e^{j\omega}\right)\right|^2 = \frac{G^2}{\left|A\left(e^{j\omega}\right)\right|^2} = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a(k)\, e^{-jk\omega}\right|^2} \tag{A.15}$$

From equations A.7 and A.9 we have

$$P(\omega) = \frac{\left|E\left(e^{j\omega}\right)\right|^2}{\left|A\left(e^{j\omega}\right)\right|^2} \tag{A.16}$$

By comparing equations A.15 and A.16 we see that if $P(\omega)$ is being modeled by $\hat{P}(\omega)$, then the error power spectrum $\left|E\left(e^{j\omega}\right)\right|^2$ is being modeled by a flat spectrum equal to G^2. This means that the actual error signal $e(n)$ is being approximated by another signal that has a flat spectrum, such as a unit impulse, white noise, or any other signal with a flat spectrum. The filter $A(\omega)$ is sometimes known as a "whitening filter" since it attempts to produce an output signal $e(n)$ that is white, i.e., has a flat spectrum.

From equations A.8, A.15, and A.16, the total error can be written as

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega \tag{A.17}$$

Therefore, minimizing the total error E is equivalent to the minimization of the integrated ratio of the signal spectrum $P(\omega)$ to its approximation $\hat{P}(\omega)$. An equivalent formulation using maximum likelihood estimation has been given by Itakura [Itakura and Saito, 1970]. Now, we can back up and restate the problem of linear prediction as follows. Given some spectrum $P(\omega)$, we wish to model it by another spectrum $\hat{P}(\omega)$ such that the integrated ratio between the two spectra as in equation A.17 is minimized. The parameters of the model spectrum are computed from the normal equations 7.16, where the needed autocorrelation coefficients $R(i)$ are easily computed from $P(\omega)$ by a simple Fourier transform. The gain factor G is obtained by equating the total energy in the two spectra, i.e., $\hat{R}(0) = R(0)$, where

$$\hat{R}(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{P}(\omega) \cos(i\omega)\, d\omega \tag{A.18}$$

Note that $\hat{R}(i)$ is the autocorrelation of the impulse response of $H(z)$.

The manner in which the model spectrum $\hat{P}(\omega)$ approximates $P(\omega)$ is largely reflected in the relation between the corresponding autocorrelation functions. From A.11, we have $\hat{R}(i) = R(i),\ 0 \leq i \leq p$. Since $\hat{P}(\omega)$ and $P(\omega)$ are Fourier transforms of $\hat{R}(i)$ and $R(i)$, respectively, it follows that increasing the value of the order of the model p increases the range over which $R(i)$ and $\hat{R}(i)$ are equal, resulting in a better fit of $\hat{P}(\omega)$ to $P(\omega)$. In the limit, as $p \to \infty$, $\hat{R}(i)$ becomes identical to $R(i)$ for all i, and hence the two spectra become identical. This statement says that we can approximate any spectrum arbitrarily closely by an all-pole model. Arbitrary frequency resolution in computing $\hat{P}(\omega)$ can be obtained by simply appending an appropriate number of zeros to this sequence before taking the FFT. An alternate method of computing $\hat{P}(\omega)$ is obtained by rewriting equation A.15 as

$$\hat{P}(\omega) = \frac{G^2}{\rho(0) + 2\sum_{i=1}^{p-i} \rho(i) \cos(i\omega)} \tag{A.19}$$

where

$$\rho(i) = \sum_{k=0}^{p-i} a(k)\, a(k+i),\ a(0) = 1,\ 0 \leq i \leq p \tag{A.20}$$

is the autocorrelation of- the impulse response of filter $A(z)$. From equation A.19, $\hat{P}(\omega)$ can be computed by dividing G^2 by the real part of the FFT of the sequence $[\rho(0), 2\rho(1), 2\rho(2), \cdots, 2\rho(p)]$. Note that the slope of $\hat{P}(\omega)$ is always zero a t $\omega = 0$ and $\omega = \pi$.

Another property of $\hat{P}(\omega)$ is obtained by noting that the minimum error $E_p = G^2$, and, therefore, from equation A.17 we have

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega = 1 \tag{A.21}$$

This relation is a special case of the more general result A.3 relating the fact that the polynomials $[A_0(z), A_1(z), \cdots, A_p(z), \cdots]$, form a complete set of orthogonal polynomials with weight $P(\omega)$. Equation A.21 is true for all values of p. In particular, it is true as $p \to \infty$, in which case equation A.21 becomes an identity. Another important case where equation A.21 becomes an identity is when $P(\omega)$ is an all-pole spectrum with p_0 poles, then $\hat{P}(\omega)$ will be identical to $P(\omega)$ for all $p \geq p_0$. Relation A.21 will be useful in discussing the properties of the error measure in Section IV.

The transfer functions $S(z)$ and $H(z)$ corresponding to $P(\omega)$ and $\hat{P}(\omega)$ are also related. It can be shown that as $p \to \infty$, $H(z)$ is given by

$$H_\infty(z) = \frac{G}{1 + \sum_{k=1}^{p} a(k) z^{-k}} = \sum_{n=0}^{N-1} h_\infty(n) z^{-n},\ p \to \infty \tag{A.22}$$

where $h_\infty(n),\ 0 \leq n \leq N-1$ is the minimum phase sequence corresponding to $s(n),\ 0 \leq n \leq N-1$. Note that the minimum phase sequence is of the same length as the original signal.

Another important conclusion is that since linear predictive analysis can be viewed as a process of spectrum or autocorrelation matching, one must be careful how to estimate the spectrum $P(\omega)$

or the corresponding autocorrelation that is to be modeled. Since the signal is often weighted or windowed' before either the autocorrelation or the spectrum is computed, it can be quite important to properly choose the type and width of the data window to be used. The choice of window depends very much on the type of signal to be analyzed. If the signal can be considered to be stationary for a long period of time (relative to the effective length of the system impulse response), then a rectangular window suffices. However, for signals that result from systems that are varying relatively quickly, the time of analysis must necessarily be limited. For example, in many transient speech sounds, the signal can be considered stationary for a duration of only one or two pitch periods. In that case a window such as Hamming or Hanning is more appropriate.

A.1.4 Modeling Discrete Spectra

Thus far we have assumed that the spectrum $P(\omega)$ is a continuous function of frequency. More often, however, the spectrum is known at only a finite number of frequencies. For example, FFT-derived spectra and those obtained from many commercially available spectrum analyzers have values at equally spaced frequency points. On the other hand, filter bank spectra, and, for example, third-octave band spectrum analyzers have values at frequencies that are not necessarily equally spaced. In order to be able to model these discrete spectra, only one change in our analysis need be made. The error measure E in equation A.17 is defined as a summation instead of an integral. The rest of the analysis remains he same except that the autocorrelation coefficients $R(i)$ are now computed from

$$R(i) = \frac{1}{M} \sum_{m=0}^{M-1} P(\omega_m) \cos(i\omega_m) \tag{A.23}$$

where M is the total number of spectral points on the unit circle. The frequencies ω, are those for which a spectral value exists, and they need not be equally spaced. Below we demonstrate the application of LP modeling for filter bank and harmonic spectra.

A.2 Error Analysis

An important aspect of any fitting or matching procedure is the properties of the error measure that is employed, and whether those properties are commensurate with certain objectives. In this section we shall examine the properties of the error measure used in LP analysis and we shall discuss its strengths and weaknesses in order to be able to fully utilize its capabilities. The analysis will be restricted to the stationary (autocorrelation) case, although the conclusions can be extrapolated to the nonstationary (covariance) case. The error measure used in Section 7.2.1.2 to determine the predictor parameters is the least squares error measure due to Gauss, who first reported on it in the early 1800's. This error measure has been used extensively since then, and is quite well understood. Its major asset is its mathematical tractability. Its main characteristic is that it puts great emphasis on large errors and little emphasis on small errors. Purely from the time domain, it is often difficult to say whether such an error measure is a desirable one or not for the problem at hand. Many would probably agree that it does not really matter which error measure one uses as long as it is a reasonable function of the magnitude of the error at each point. For the linear prediction problem, we are fortunate that the error measure can also be written in the frequency domain and can be interpreted as a goodness of fit between a given signal spectrum and a model spectrum that approximates it. The insights gained in the frequency domain should enhance our understanding of the least squares error criterion.

A.2.1 The Minimum Error

For each value of p, minimization of the error measure E in equation A.17 leads to the minimum error E_p in equation 7.17, which is given in terms of the predictor and autocorrelation coefficients. Here we derive an expression for E_p in the frequency domain, which will help us determine some of its properties. Other properties will be discussed when we discuss the normalized minimum error.

Let

$$\hat{c}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \hat{P}(\omega)\, d\omega \tag{A.24}$$

be the zeroth coefficient (quefrency) of the cepstrum (inverse Fourier transform of log spectrum) [Childers et al., 1977] corresponding to $\hat{P}(\omega)$. From equation A.15, equation A.24 reduces to

$$\hat{c}(0) = \log G^2 - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left| A\left(e^{j\omega}\right) \right|^2 d\omega \tag{A.25}$$

$A(z)$ has all its zeros inside the unit circle. Therefore, the integral in equation A.25 is equal to zero. Since $G^2 = E_p$, we conclude from equation A.25 that

$$E_p = e^{\hat{c}(0)} \tag{A.26}$$

From equations A.26 and A.24, E_p can be interpreted as the geometric mean of the model spectrum $\hat{P}(\omega)$. From equation 7.42 we know that E_p decreases as p increases. The minimum occurs as $p \to \infty$ and is equal to

$$E_{\min} = E_\infty = e^{c_0} \tag{A.27}$$

where c_0 is obtained by substituting $P(\omega)$ for $\hat{P}(\omega)$ in equation A.24. Therefore, the absolute minimum error is a function of $P(\omega)$ only, and is equal to its geometric mean, which is always positive for positive definite spectra. This is a curious result, because it says that the minimum error can be nonzero even when the matching spectrum $\hat{P}(\omega)$ is identical to the matched spectrum $P(\omega)$. Therefore, although E_p is a measure of fit of the model spectrum to the signal spectrum, it is not an absolute one. The measure is always relative to $E_{\min}$. The nonzero aspect of $E_{\min}$ can be understood by realizing that, for any p, E_p is equal to that portion of the signal energy that is not predictable by a p^{th} order predictor. For example, the impulse response of an all-pole filter is perfectly predictable except for the initial nonzero value. It is the energy in this initial value that shows up in E_p. (Note that in the covariance method one can choose the region of analysis to exclude the initial value, in which case the prediction error would be zero for this example.)

A.2.2 Spectral Matching Properties

The LP error measure E in equation A.17 has two major properties, a global property and a local property.

1. Global Property: Because the contributions to the total error are determined by the ratio of the two spectra, the matching process should perform uniformly over the whole frequency range, irrespective of the general shaping of the spectrum. This is an important property for spectral estimation because it makes sure that the spectral match at frequencies with little energy is just as good, on the average, as the match at frequencies with high energy. If the

error measure had been of the form $\int \left| P(\omega) - \hat{P}(\omega) \right| d\omega$, the spectral matches would have been best at high energy frequency points.

2. Local Property: This property deals with how the match is done in each small region of the spectrum.

Let the ratio of $P(\omega)$ to $\hat{P}(\omega)$ be given by

$$E(\omega) = \frac{P(\omega)}{\hat{P}(\omega)} \tag{A.28}$$

Then from equation A.21we have

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} E(\omega)\, d\omega = 1, \ \forall p \tag{A.29}$$

$E(\omega)$ can be interpreted as the "instantaneous error" between $P(\omega)$ and $\hat{P}(\omega)$ at frequency ω. Equation A.29 says that the arithmetic mean of $E(\omega)$ is equal to 1, which means that there are values of $E(\omega)$ greater and less than 1 such that the average is equal to l. In terms of the two spectra, this means that $P(\omega)$ will be greater than $\hat{P}(\omega)$ in some regions and less in others such that equation A.29 applies. However, the contribution to the total error is more significant when $P(\omega)$ is greater than $\hat{P}(\omega)$ than when $P(\omega)$ is smaller, e.g., a ratio of $E(\omega) = 2$ contributes more to the total error than a ratio of $1/2$. We conclude that after the minimization of error, we expect a better fit of $\hat{P}(\omega)$ to $P(\omega)$ where $P(\omega)$ is greater than $\hat{P}(\omega)$, than where $P(\omega)$ is smaller (on the average). For example, if $P(\omega)$ is the power spectrum of a quasi-periodic signal, then most of the energy in $P(\omega)$ will exist in the harmonics, and very little energy will reside between harmonics. The error measure in equation A.21 insures that the approximation of $\hat{P}(\omega)$ to $P(\omega)$ is far superior at the harmonics than between the harmonics. If the signal had been generated by exciting a filter with a periodic sequence of impulses, then the system response of the filter must pass through all the harmonic peaks. Therefore, with a proper choice of the model order p, minimization of the LP error measure results in a model spectrum that is a good approximation to that system response. This leads to one characteristic of the local property, minimization of the error measure E results in a model spectrum $\hat{P}(\omega)$ that is a good estimate of the spectral envelope of the signal spectrum $P(\omega)$. It should be clear from the above that the importance of the local property is not as crucial when the variations of the signal spectrum from the spectral envelope are much less pronounced.

In the modeling of harmonic spectra, we can encounter examples where, although the all-pole spectrum resulting from LP modeling is a reasonably good estimate of the harmonic spectral envelope, it does not yield the unique all-pole transfer function that coincides with the line spectrum at the harmonics. This is a significant disadvantage of LP modeling, and is an indirect reflection of another characteristic of the local property, the cancellation of errors. This is evident from equation A.29 where the instantaneous errors $E(\omega)$ are greater and less than 1 such that the average is 1. To help elucidate this point, let us define a new error measure E' that is the logarithm of E in equation A.17

$$E' = \log \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega \right] \tag{A.30}$$

where the gain factor has been omitted since it is not relevant to this discussion. It is simple to show that the minimization of E' is equivalent to the minimization of E. For cases where $P(\omega)$ is smooth relative to $\hat{P}(\omega)$ and the values of $P(\omega)$ are not expected to deviate very much from

$\hat{P}(\omega)$, the logarithm of the average of spectral ratios can be approximated by the average of the logarithms, i.e.,

$$E' \simeq \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \frac{P(\omega)}{\hat{P}(\omega)} d\omega \tag{A.31}$$

From equation A.31 it is clear that the contributions to the error when $P(\omega) > \hat{P}(\omega)$ cancel those when $P(\omega) < \hat{P}(\omega)$. The above discussion suggests the use of an error measure that takes the magnitude of the integrand in equation A.31. One such error measure is

$$E'' = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\log \frac{P(\omega)}{\hat{P}(\omega)} \right]^2 d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\log P(\omega) - \log \hat{P}(\omega) \right]^2 d\omega \tag{A.32}$$

E'' is just the mean squared error between the two log spectra. It has the important property that the minimum error of zero occurs if and only if $\hat{P}(\omega)$ is identical to $P(\omega)$. However, while the error measure E solves one problem, it introduces an other. Note that the contributions to the total error in equation A.32 are equally important whether $P(\omega) > \hat{P}(\omega)$ or vice versa. This means that if the variations of $P(\omega)$ are large relative to $\hat{P}(\omega)$, the resulting model spectrum will nor be a good estimate of the spectral envelope. In addition, the minimization of E'' in equation A.32 results in a set of nonlinear equations that must be solved iteratively, thus increasing the computational load tremendously.

Our conclusion is that the LP error measure in equation A.21 is to be preferred in general, except for certain special cases where an error measure such as E'' in equation A.32 can be used, provided one is willing to carry the extra computational burden.

The global and local properties described here are properties of the error measure in equation A.21 and do not depend on the details of $P(\omega)$ and $\hat{P}(\omega)$. These properties apply on the average over the whole frequency range. Depending on the detailed shapes of $P(\omega)$ and $\hat{P}(\omega)$, the resulting match can be better in one spectral region than in another. For example, if $\hat{P}(\omega)$ is an all-pole model spectrum and if the signal spectrum $P(\omega)$ contains zeros as well as poles, then one would not expect as good a match at the zeros as at the poles. This is especially true if the zeros have bandwidths of the same order as the poles or less. (Wide bandwidth zeros are usually well approximated by poles.) On the other hand, if $\hat{P}(\omega)$ is an all-zero spectrum then the preceding statement would have to be reversed.

A.2.3 The Normalized Error

The normalized error has been a very useful parameter for the determination of the optimal number of parameters to be used in the model spectrum. This subject will be discussed in the following section. Here we shall present some of the properties of the normalized error, especially as they relate to the signal and model spectra.

1. Relation to the Spectral Dynamic Range

The normalized error was defined in Section 7.2.1.8 as the ratio of the minimum error E_p to the energy in the signal $R(0)$. Keeping in mind that $R(0) = \hat{R}(0)$, and substituting for E_p from equation A.26, we obtain

$$V_p = \frac{E_p}{R(0)} = \frac{e^{\hat{c}_0}}{\hat{R}(0)} \tag{A.33}$$

Also, from equation A.27, we have in the limit as $p \to \infty$

$$V_{\min} = V_{\infty} = \frac{e^{c_0}}{R(0)} \tag{A.34}$$

Therefore, the normalized error is always equal to the normalized zero quefrency of the model spectrum. From equations 7.42 and A.33 it is clear that V_p is a monotonically decreasing function of p, with $V_0 = 1$ and $V_{\infty} = V_{\min}$ in equation A.34.

It is instructive to write V_p as a function of $\hat{P}(\omega)$. From equations A.18 and A.24, equation A.33 can be rewritten as

$$V_p = \frac{\exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \hat{P}(\omega)\, d\omega\right]}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{P}(\omega)\, d\omega} \tag{A.35}$$

It is clear from equation A.35 that V_p depends completely on the shape of the model spectrum, and from equation A.35, $V_{\min}$ is determined solely by the shape of the signal spectrum. An interesting way to view equation A.35 is that V_p is equal to the ratio of the geometric mean of the model spectrum to its arithmetic mean. This ratio has been used in the past as a measure of the spread of the data. When the spread of the data is small, the ratio is close to 1. Indeed, from equation A.35 it is easy to see that if $\hat{P}(\omega)$ is flat, $V_p = 1$. On the other hand, if the data spread is large, then V_p becomes close to zero. Again, from equation A.35 we see that if $\hat{P}(\omega)$ is zero for a portion of the spectrum (hence a large spread), then $V_p = 1$. (Another way of looking at V_p is in terms of the flatness of the spectrum [Makhoul, 1975].) Another measure of data spread is the dynamic range. We define the spectral dynamic range d as the ratio of the highest to the lowest amplitude points on the spectrum

$$d = {}^{H}/_{L} \tag{A.36}$$

where

$$H = \max_{\omega} \hat{P}(\omega) \qquad L = \min_{\omega} \hat{P}(\omega) \tag{A.37}$$

A.2.4 A Measure of Ill-Conditioning

In solving the autocorrelation normal equations 7.16, the condition of the autocorrelation matrix is an important consideration in deciding the accuracy of the computation needed. An ill-conditioned matrix can cause numerical problems in the solution. An accepted measure of ill-conditioning in a matrix is given by the ratio

$$d' = {}^{\lambda_{\max}}/_{\lambda_{min}} \tag{A.38}$$

where $\lambda_{\max}$ and λ_{min} are the maximum and minimum eigenvalues of the matrix. It is possible to show [Makhoul, 1975] that all the eigenvalues of an autocorrelation matrix lie in the range $\lambda_i \in [H, L]$, $1 \leq i \leq p$, where H and L are defined in equation A.37. In addition, as the order of the matrix p increases, the eigenvalues become approximately equal to $\hat{P}(\omega)$ evaluated at equally spaced points with separation ${}^{2\pi}/_{(p+1)}$. Therefore, the ratio d' given in equation A.38 can be well approximated by the dynamic range of $\hat{P}(\omega)$

$$d' \simeq d \tag{A.39}$$

Therefore, the spectral dynamic range is a good measure of the ill-conditioning of the autocorrelation matrix. The larger the dynamic range, the greater is the chance that the matrix is illconditioned.

But in the previous section we noted that an increase in d usually results in a decrease in the normalized error V_p. Therefore, V_p can also be used as a measure of ill-conditioning: the ill-conditioning is greater with decreased V_p. The problem becomes more and more serious as $V_p \to 0$, i.e., as the signal becomes highly predictable.

If ill-conditioning occurs sporadically, then one way of patching the problem is to increase the values along the principal diagonal of the matrix by a small fraction of a percent. However, if the problem is a regular one, then it is a good idea if one can reduce the dynamic range of the signal spectrum. For example, if the spectrum has a general slope, then a single zero filter of the form $1 + az^{-1}$ applied to the signal can be very effective. The new signal is given by

$$s'(n) = s(n) + as(n-1) \tag{A.40}$$

An optimal value for a is obtained by solving for the filter $A(z)$ that "whitens" (flattens) $s'(n)$. This is, of course, given by the first order predictor, where

$$a = -\frac{R(1)}{R(0)} \tag{A.41}$$

$R(1)$ and $R(0)$ are autocorrelation coefficients of the signal $s(n)$. The filtered signal $s'(n)$ is then guaranteed to have a smaller spectral dynamic range. The above process is usually referred to as *preemphasis*.

One conclusion from the above concerns the design of the low-pass filter that one uses before sampling the signal to reduce aliasing. In order to ensure against aliasing, it is usually recommended that the cutoff frequency of the filter be lower than half the sampling frequency. However, if the cutoff frequency is appreciably lower than half the sampling frequency, then the spectral dynamic range of the signal spectrum increases, especially if the filter has a sharp cutoff and the stop band is very low relative to the pass band. This increases problems of ill-conditioning. Therefore, if one uses a lowpass filter with a sharp cutoff, the cutoff frequency should be set as close to half the sampling frequency as possible.

Appendix B

Discrete All-Pole Model

In this appendix we will discuss the properties of the error measure for the discrete all-pole model.

B.1 Properties of The Error Measure

The Itakura-Saito (IS) error measure was defined originally for continuous spectra [Itakura and Saito, 1968], [Itakura and Saito, 1970]. However, it can be adapted to the discrete case as follows

$$E_{IS} = \frac{1}{N} \sum_{m=1}^{N} \frac{P(\omega_m)}{\hat{P}(\omega_m)} - \ln \frac{P(\omega_m)}{\hat{P}(\omega_m)} - 1 \tag{B.1}$$

where, just like before, $P(\omega_m)$ is the given discrete spectrum defined at N frequencies ω_m and $\hat{P}(\omega_m)$ is the all-pole model spectrum defined in equation A.15 and evaluated at the same frequencies. This error measure is always nonnegative and is equal to zero only when $\hat{P}(\omega_m) = P(\omega_m)$, $\forall \omega_m \in \Omega$. Notice that $\hat{P}(\omega_m) = P(\omega_m)$ gives a minimum for E_{IS} but not necessarily for E_{LP}, as demonstrated in the previous section.

The continuous form of this error measure was originally presented as part of a maximum likelihood approach to linear prediction and was shown to produce the same result as LP for continuous spectra. Later, the discrete version shown in equation B.1 was derived by McAuley [McAuley, 1984] for the maximum likelihood spectral modeling of periodic speech signals with Gaussian statistics.

A spectral flatness interpretation of this discrete error measure makes it a very reasonable choice for the problem of fitting an envelope to a set of discrete spectral values. It can be shown that minimizing the error in equation B.1 is equivalent to maximizing the spectral flatness of the error spectrum $\hat{P}(\omega_m)/P(\omega_m)$, where the spectral flatness is defined as the geometric mean of the spectral samples divided by their arithmetic mean. The proof parallels the one for the continuous case given by Markel and Gray [Markel and Gray, 1976]. The major consequence of this property is that our optimal model is the one which makes the residual (error) spectrum as flat as possible.

It has been shown that, for small values of E_{IS}, the IS error approximates the mean-squared distance between log spectra [Itakura and Saito, 1968]. Based on this property, El-Jaroudi defines

$$E_{dB} = 6.142\sqrt{E_{IS}} \approx \sqrt{\frac{1}{N} \sum_{m=1}^{N} \left[10 \log_{10} P(\omega_m) - 10 \log_{10} \hat{P}(\omega_m)\right]^2} \tag{B.2}$$

for small E_{IS}, such that we can use E_{dB} when comparing error values since it provides an approximate estimate of the spectral error in decibels. It is important to note that, for the continuous case of the IS measure, the optimal all-pole model is the same as the one produced by LP. Therefore, by using this error measure, we do not sacrifice any of the advantages or performance of LP in unvoiced segments of speech.

B.1.1 Error Minimization

First, we will minimize the error measure in equation B.1 with $\hat{P}(\omega)$ expressed as

$$\hat{P}(\omega) = \frac{1}{D(\omega)} = \frac{1}{\sum_{k=0}^{p} d_k \cos \omega k} \tag{B.3}$$

where $\{d_k\}$ can be shown to be equal to

$$d_0 = \sum_{k=0}^{p} a_k^2 \tag{B.4}$$

$$d_i = 2 \sum_{k=0}^{p-i} a_k a_{k+1},\ 1 \leq i \leq p \tag{B.5}$$

Note that d_i is twice the autocorrelation of $\{a_k\}$ for $1 \leq i \leq p$, and d_0 is equal to the zero-lag autocorrelation. We then set $\partial E_{IS}/\partial d_i = 0$ for $i = 0, \cdots, p$. The result can be shown to yield a set of correlation matching conditions, given by

$$\hat{R}(i) = R(i),\ 0 \leq i \leq p \tag{B.6}$$

where $R(i)$ is the autocorrelation corresponding to the given discrete spectrum defined in equation A.23 and $\hat{R}(i)$ is the autocorrelation corresponding to the all-pole model sampled at the same discrete frequencies as the given spectrum

$$\hat{R}(i) = \frac{1}{N} \sum_{m=1}^{N} \hat{P}(\omega_m) \cos \omega_m i \tag{B.7}$$

Equation B.6 looks deceptively similar to the matching condition expressed in equation 7.52. The major difference, however, is that in LP, $\hat{R}_{LP}(i)$ is the autocorrelation of the continuous all-pole spectrum $\hat{P}(\omega)$, while here, $\hat{R}(i)$ in equation B.7 is the autocorrelation of a discrete sampling of the all-pole spectrum. From equation B.6, we see that DAP requires matching the given aliased autocorrelation to the autocorrelation of the all-pole model aliased in the same manner. According to El-Jaroudi and Makhoul [El-Jaroudi and Makhoul, 1969], it is this improved correlation matching condition, which incorporates the autocorrelation aliasing, that makes DAP better suited than LP for analyzing discrete spectral.

While the matching condition B.6 provides us with insight into the solution of the modeling problem, it does not give us a way of computing the parameters of the optimal all-pole model. The all-pole model is obtained by using the definition of $\hat{P}(\omega)$ in equation A.15 and setting $\partial E_{IS}/\partial a_i = 0,\ i = 0, \cdots, p$. This yields the following set of equations relating the predictor coefficients $\{a_k\}$ to the autocorrelations of the given discrete spectrum and the sampled all-pole model

$$2 \sum_{k=0}^{p} a_k \Big[R(i-k) - \hat{R}(i-k) \Big] = 0,\ 0 \leq i \leq p \tag{B.8}$$

The conditions in equations B.8 can be expressed in matrix notation as

$$2\left(\mathbf{R}-\hat{\mathbf{R}}\mathbf{a}\right)=\mathbf{0} \tag{B.9}$$

where $\mathbf{a}$ is the column vector of predictor coefficients, and $\mathbf{R}$ and $\hat{\mathbf{R}}$ are symmetric Toeplitz matrices with elements $R(i-j)$ and $\hat{R}(i-j)$, $0 \leq i,\, j \leq p$, respectively. Since $\hat{\mathbf{R}}$ is a function of $\mathbf{a}$, equation B.9 constitutes a set of $p+1$ unknowns. In the next section we will derive the minimum error, followed by the solution to equation B.9.

B.1.2 Minimum Error

The expression for minimum error is obtained by substituting the condition for minimization in equation B.9 into equation B.1. We begin by simplifying the first term in the error measure B.1

$$\frac{1}{N}\sum_{m=1}^{N}\frac{P(\omega_m)}{\hat{P}(\omega_m)}=\sum_{k=0}^{p}\sum_{j=0}^{p}a_k a_j R(k-j)=\mathbf{a}^T\mathbf{R}\mathbf{a}=\mathbf{a}^T\hat{\mathbf{R}}\mathbf{a} \tag{B.10}$$

and

$$\mathbf{a}^T\hat{\mathbf{R}}\mathbf{a}=\sum_{k=0}^{p}\sum_{j=0}^{p}a_k a_j \hat{R}(k-j)=\frac{1}{N}\sum_{m=1}^{N}\frac{P(\omega_m)}{\hat{P}(\omega_m)}=1 \tag{B.11}$$

Consequently, from equation B.1 we have

$$E_{IS\min}=\frac{1}{N}\sum_{m=1}^{N}-\ln\frac{P(\omega_m)}{\hat{P}(\omega_m)}=\ln\frac{\left[\prod_{m=1}^{N}\hat{P}(\omega_m)\right]^{1/N}}{\left[\prod_{m=1}^{N}P(\omega_m)\right]^{1/N}} \tag{B.12}$$

We conclude from equations B.11 and B.12 that, at the minimum, the energy in the residual spectrum $P(\omega_m)/\hat{P}(\omega_m)$ is automatically normalized to 1 and the minimum error is equal to the logarithm of the ratio of the geometric mean of the model spectrum and the geometric mean of the given spectrum. Both these properties have their equivalent in continuous spectrum LP [Makhoul, 1975]. Based on the similarities between the two methods (DAP and LP) and the fact that DAP reduces to LP for the continuous spectrum case while LP does not reduce to DAP for discrete spectra, El-Jaroudi concludes that LP is just a special case of DAP modeling where the number of spectral points goes to infinity.

B.1.3 The Solution and its Uniqueness

Now we focus on the solution of the minimization conditions B.9. These equations allow one of two possible solutions:

1. A matching solution in which $\hat{\mathbf{R}}=\mathbf{R}$ and the model satisfies the conditions in equation B.6;
2. A singular solution in which $\hat{\mathbf{R}}\neq\mathbf{R}$, and therefore the predictor vector $\mathbf{a}$ will be an eigenvector of the difference matrix $\left(\mathbf{R}-\hat{\mathbf{R}}\mathbf{a}\right)$ corresponding to an eigenvalue equal to 0. Also, note that the trivial solution $\mathbf{a}=\mathbf{0}$ is not possible since it produces unbounded values for $\hat{\mathbf{R}}$.

Consequently, the optimal all-pole model will belong to one of these two classes of solutions. It will either have an aliased autocorrelation equal to that of the given discrete signal (corresponding to the matching solution), or it will not (corresponding to the singular solution). El-Jaroudi and

Makhoul [El-Jaroudi and Makhoul, 1969] examine the properties of the error function and of the optimal all-pole model. He concluded that the error function is convex and that the optimal all-pole model is unique depending on the number N of spectral points. Following, he presents an iterative algorithm to find the solution to equations B.8 and evaluates its convergence properties.

Appendix C

Cepstral Smoothing

This appendix examines the effect of several common procedures applied to the calculation of Fourier spectra on the cepstrum, such as phase unwrapping, windowing, zero-padding, and spectrum notching. We will also see the results of the above, such as aliasing and oversampling.

C.1 The Phase Cepstrum

The inverse transform of the phase of the complex logarithm yields peaks at multiples of the echo arrival time in much the same way that the inverse transform of the log magnitude does. This can be shown as follows for the single additive echo case:

$$\begin{aligned} \hat{X}\left(e^{j\omega}\right) = \log\left(X_1\left(e^{j\omega}\right)\right) + \log\left(\left[1 + ae^{-j\omega n_0}\right]\right) = \\ \log\left|X_1\left(e^{j\omega}\right)\right| + j\text{phase}\left(X_1\left(e^{j\omega}\right)\right) + \frac{1}{2}\log\left(1 + a^2 + 2a\cos\left(\omega n_0 T\right)\right) + \\ j\arctan\left(-\frac{a\sin\left(\omega n_0 T\right)}{1 + a\cos\left(\omega n_0 T\right)}\right) \end{aligned} \tag{C.1}$$

The fourth term on the right produces ripples in the phase, just as the third term produces ripples in the log magnitude. Since $\hat{X}\left(e^{j\omega}\right)$ is obtained from the transform of a real sequence, its real part (magnitude of the transform of the real sequence) is an even function of ω, and its imaginary part (phase of the transform of the real sequence) is an odd function of ω. Thus the inverse transform of $\Re\left\{\hat{X}\left(e^{j\omega}\right)\right\}$ will yield the even portion of the complex cepstrum and the inverse transform of $j\Im\left\{\hat{X}\left(e^{j\omega}\right)\right\}$ ill produce the odd portion of the complex cepstrum. Since the inverse transform of the term $\log\left(1 + ae^{-j\omega n_0 T}\right)$ produces peaks on one side of the origin only, the peaks produced by its real and imaginary parts must be equal in magnitude and opposite in sign on one side of the origin but of the same sign on the other side of the origin (depending upon whether the echo amplitude a is less or greater than unity).

From these observations we formally define the phase cepstrum of a data sequence as the square of the inverse z-transform of twice the phase (the imaginary part of the logarithm) of the z-transform of the data sequence. This may be written as

$$x\angle(n) = \left\{Z^{-1}\left[2\log X(z) - 2\log|X(z)|\right]\right\}^2 \tag{C.2}$$

where the factor of 2 has been introduced to eliminate any normalization factors in the relation between the phase and complex cepstra and $x\angle(0) = 0$. From equations 7.67, 7.70, and 7.73 the phase cepstrum can be easily shown to be

$$x\angle(n) = \left[\hat{x}(-n) - \hat{x}(-n)\right]^2 \tag{C.3}$$

Thus the phase cepstrum is to the phase as the power cepstrum is to the log magnitude. Once again the final squaring operation could be changed to magnitude squared or eliminated. Empirically, it has been determined that the phase cepstrum is less useful than the power cepstrum in the determination of echo arrival times [Childers et al., 1977]. This is apparently due to the phase unwrapping errors produced by additive noise and linear phase terms. The phase cepstrum is as difficult to compute as the complex cepstrum, since both require phase unwrapping. However, the phase cepstrum has proven valuable in evaluating the effects of noise on the signal phase. Significant differences in the appearance of the phase and power cepstra can be indicative of phase unwrapping problems which might otherwise go unnoticed [Childers et al., 1977]. Many problems arise in the computation of the phase sequence for the complex cepstrum. Here we address several of these problems along with their alleviation.

C.2 Linear Phase Components

The presence of a linear component in the phase sequence introduces rapidly decaying oscillations in the complex cepstrum, e.g., let the spectrum of such a signal be represented as $X\left(e^{j\omega}\right) = e^{-jr\omega}X'\left(e^{j\omega}\right)$ or $X(z) = z^{-r}X'(z)$. Then the cepstrum of the linear phase term alone is

$$\hat{x}\angle(n) = \begin{cases} 0, & n = 0 \\ \frac{-r}{nT}\cos n\pi = \frac{-r}{nT}(-1)^n, & n \neq 0 \end{cases} \tag{C.4}$$

This term is added to the cepstrum of the remaining portion of the data being analyzed. Note that it changes sign at each sample and although it does decay, it may be quite large depending upon r. Such a term may mask echo peaks in the complex cepstrum, and should be removed by subtraction from the composite signal phase. Basically, this is just trend removal, which is standard practice for improving spectral estimates. The removed linear phase term can be recorded and then reinserted during the inversion process if necessary.

The presence of a linear phase term may influence the choice of liftering to be applied to the complex cepstrum. If the echo is to be removed and the basic wavelet is to be recovered, then the echo peaks should not be notch liftered (removed) by simply replacing them with the average of their adjacent points, since these adjacent points have contributions from the linear phase component (if it has not been completely removed) which are opposite in sign to the contribution of the echo point to be removed. Instead, if the echo is located at n_0 in the complex cepstrum then this point should be replaced with the average of the $n_0 + 2$ and $n_0 - 2$ points. This form of liftering results in a smaller mean-square error (MSE) in the recovered wavelet than when the average of the points adjacent to the echo peak is used. This has been found to be the case even when the linear phase component has been completely removed [Childers et al., 1977]. This liftering procedure is not claimed to be optimum. In fact the liftering procedure is undoubtedly signal and noise dependent and would in general involve averaging more than just two points in the complex cepstrum.

A serious problem in phase unwrapping is encountered when discontinuities in the phase occur in calculating the phase modulo 2π via the arctan routine. The phase unwrapping algorithm previously described removes these discontinuities provided the phase changes by less than π between samples. Recently, it has been pointed out that a linear phase component with a large

slope will cause errors in this unwrapping procedure [Childers et al., 1977]. If the phase changes between samples are greater than π due to the presence of a linear phase term, then this unwrapping problem can be alleviated by increasing the record length with the addition of zeros [Childers et al., 1977]. This is equivalent to sampling the z-transform more frequently. If one is unsure whether the phase change between samples is less than π, then one can check such an hypothesis with the above procedure by comparing the unwrapped phase before and after the record length has been appended with zeros.

One example of where the linear phase component gives problems is when $x(n) = x_1(n - n_0)$, $[0, N-1]$, zero otherwise, then $X\left(e^{j\omega}\right) = e^{-j\omega n_0} X_1\left(e^{j\omega}\right)$. As expected the phase of x is the sum of a linear phase component and the phase of x_1. If ω is the minimum rate $\omega = {}^{n2\pi}/_{NT}$ then $X\left(e^{jn(2\pi/N)}\right) = e^{j2\pi n(n_0/N)} X_1\left(e^{j\omega}\right)$. If $n_0 > {}^{N}/_{2}$ the linear phase component will change by more than π between samples and unless the phase of x_1 counteracts this change, the phase unwrapping algorithm will yield erroneous results. This has been observed in computer experiments when the composite signal is delayed by more than half the record length. As expected this not only reduces the echo detectability in both the phase and complex cepstra, but also severely distorts the recovered wavelet.

C.3 Spectrum Notching

It should also be noted that zeros near the unit circle in the z-transform of the echo sequence result in notches in the spectrum sequence wherein additive noise may dominate. We have seen earlier that one phase unwrapping algorithm requires that the changes in phase between samples must be less than $\pm\pi$, i.e., the derivative of the phase with respect to frequency must be less than $\pm\pi$.

Consider the z-transform of the data sequence evaluated on the unit circle, then

$$X\left(e^{j\omega}\right) = \left|X\left(e^{j\omega}\right)\right| e^{j\angle X\left(e^{j\omega}\right)} = X_{\Re}\left(e^{j\omega}\right) + jX_{\Im}\left(e^{j\omega}\right) \tag{C.5}$$

or

$$\frac{dX\left(e^{j\omega}\right)}{dx_1} = \frac{X_{\Re}\left(e^{j\omega}\right)\left[{}^{dX_{\Im}\left(e^{j\omega}\right)}/_{dx_1}\right] - X_{\Im}\left(e^{j\omega}\right)\left[{}^{dX_{\Re}\left(e^{j\omega}\right)}/_{dx_1}\right]}{\left|X\left(e^{j\omega}\right)\right|^2} \tag{C.6}$$

Thus the change in phase is inversely proportional to the magnitude squared of the spectrum. If a notch occurs in the spectrum, then the change in the phase may be quite large, and, therefore, proper phase unwrapping may be difficult to achieve. Further, the phase may change sign rapidly in these spectrum notches. This represents a serious problem even in the absence of noise as the above example illustrates. Therefore, it is quite possible for the unwrapped phase curve to contain discontinuities (jumps or steps) in the vicinity of a spectrum notch.

As was pointed out earlier spectral nulls can be caused physically by π phase reversals in reflections at boundary interfaces. Nulls in the spectrum may be an important aid to data interpretation. The investigator needs to understand the physical situation under which the data are collected and to model it well.

C.4 Aliasing

Aliasing of the cepstrum is of course an ever present problem since the nonlinear complex logarithm introduces harmonics into $\hat{X}(z)$. The appending of zeros to the input data sequence reduces aliasing as will selecting the data record length NT to be as large as possible. This latter choice is subject to the constraints imposed by the investigator on the number of points that can be analyzed and the minimum sampling rate. If the total data record length exceeds the duration of

the composite signal contained within the record, then it is questionable if the total data record length should be further extended with still more "data." The reason for this doubt is that the spectral samples will increasingly reflect the effect of the noise rather than the signal as the total data record length surpasses that of the composite signal duration.

C.5 Oversampling

Oversampling of the data record when noise is present is also a problem. Outside the signal band, noise dominates the spectrum. This usually presents no problem in ordinary spectrum analysis since these components frequently contain little power but this may not be the case for the cepstrum. Because of the nonlinear logarithmic operation, the regions of low power in the spectrum may contribute as much or more to the cepstrum as the regions which contain the signal in the spectrum. When this occurs it affects both echo detectability and wavelet recovery. Oversampling also aggravates phase unwrapping and aliasing since it shortens the data record (if the total number of data points or samples is fixed), which in turn implies that the samples of the log spectrum are spaced farther apart.

C.6 Zero-Padding

It is well known that appending zeros to a data sequence increases the sampling "rate" of its discrete Fourier transform. This benefits the computation of the cepstrum in two ways. First, the increased sampling "rate" in the frequency domain reduces aliasing of the cepstrum. Second, increasing the fineness with which the phase curve is sampled reduces the number of phase unwrapping errors (which result from jumps greater than π between samples). Childers [Childers et al., 1977] indicated that extending the record length with zeros results in a modest improvement in the recovered wavelet even when aliasing and phase unwrapping errors do not appear to be a problem. It should be noted that unless the record length is extended with zeros, then aliasing causes an ambiguity in the determination of the echo epoch (arrival time) and amplitude. This is due to the fact that there is no way to distinguish between an echo of relative amplitude a and delay n_0 and one with amplitude $1/a$ and delay $N - n_0$ where N is the total number of samples.

Mathematically, these statements are verified as follows: let us consider the z-transform of the sequence $x(n)$ where $x(n) = 0$ outside $[0, N-1]$

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \tag{C.7}$$

which when evaluated on the unit circle gives

$$X\left(e^{j\omega}\right) = \sum_{n=0}^{N-1} x(n) e^{-j\omega nT} \tag{C.8}$$

If we sample at uniformly spaced intervals around the unit circle, we obtain

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi k \frac{n}{N}} \tag{C.9}$$

which is just the discrete Fourier transform (DFT) of $x(n)$. It follows that

$$\hat{X}(k) = \log X(k) \tag{C.10}$$

Since the logarithm (which is a zero memory nonlinearity) of a sampled function is equivalent to sampling the logarithm of the function, then with a little additional effort it follows that the complex cepstrum of the DFT of $x(n)$ (or the z-transform of $x(n)$ sampled on the unit circle) is just the periodic extension of the complex cepstrum of the original data sequence. We see that the effect of appending zeros is to increase N. This implies we sample the log spectrum at smaller intervals, since the spacing between these samples is proportional to $1/N$. As described above, the errors introduced by a linear phase component or aliasing are reduced by increasing N through zero-padding.

C.7 Effects of Windowing

Echo detection and extraction are degraded by applying to the data record a window ordinarily used to reduce leakage, e.g., Hamming, Hanning, Tapering (Tukey window), unless the window is relatively constant (flat) over that portion of the data record containing the composite signal. In speech processing this is not the case. Here the data are highly nonstationary. And the investigator is frequently interested in analyzing the speech signal over one pitch period (or at most three pitch periods). In this case windowing is of considerable benefit.

One can see for the single echo case that windowing the input data record normally prevents the logarithmic operation from fully separating the basic wavelet and the echo series as follows:

$$x(n) = w(n)\left[x_1(n) + ax_1(n - n_0)\right] \tag{C.11}$$

or

$$X(z) = W(z) * \left[X_1(z)\left(1 + az^{-n_0}\right)\right] \tag{C.12}$$

For arbitrary $W(z)$, the contributions of the basic wavelet and the echo cannot generally be separated by taking the logarithm of equation C.12 since the term in brackets is convolved with $W(z)$. Fortunately, as will be discussed more fully below, in practice the cepstrum procedure can still be applied with effectiveness even though there is some error.

Schafer [Schafer, 1968] suggested a window which does preserve the separability of the basic wavelet and echo series and which has proven extremely useful in cepstrum analysis. This window denoted as

$$w(n) = \begin{cases} \alpha^{nT}, & 0 \le n \le N-1,\ 0 < \alpha < 1 \\ 0, & \text{otherwise} \end{cases} \tag{C.13}$$

was proposed to reduce the error associated with truncating the echo when it extended beyond the end of the record [Schafer, 1968]. Childers [Childers et al., 1977] indicated that this window is quite useful because it reduces the aliasing of the echo impulse train in the complex cepstrum by imposing an $\left(\alpha^{n_0 T}\right)^n$ weighting on the impulses. This follows directly from a calculation of the z-transform of equation C.11 with equation C.13 used for $w(n)$, i.e., for this case

$$X(z) = X_1\left(\alpha^{-T} z\right)\left[1 + a\alpha^{n_0 T} z^{-n_0}\right] \tag{C.14}$$

provided no truncation error is present and that the basic wavelet begins at $n = 0$.

From equation C.12 we see that when no window is used and a is near unity and the echo delay is a substantial portion of the record length NT, the higher order peaks may not decrease rapidly enough to avoid aliasing. This problem can be overcome with the window under consideration. Childers [Childers et al., 1977] suggests that the choice of α is data dependent and α should be chosen as close to unity as possible, consistent with the desired reduction in aliasing. The closer

the data sequence is to a maximum phase sequence, the more one can reduce α, e.g., from 0.99 to 0.98 or 0.96. The choice of α is also dependent on the echo delay time which is discussed more fully later.

The exponential window can introduce some distortion into the recovered wavelet even if the data are unweighted by the inverse window in the recovery process. This is primarily due to the distortions introduced into the data that extend beyond the duration of the wavelet of interest.

In summary the exponential window performs nearly as well as the rectangular window when no noise is present but does introduce some distortion as noted above. Further, the echo arrival time can be determined even when wavelet recovery cannot be effected. Also if rectangular windowing is judiciously applied, then the cepstrum can be used to detect similar but not necessarily identical wavelets. We suspect that if the exponential window is used to make the composite signal minimum phase, then the echo, most probably, will be lost.

Finally, it should be noted that the exponential window may be used to alter the SNR of a data record more effectively than the rectangular window. This can be effected when the composite signal occupies only a portion of the total record. In this case the window may weight the signal more or less heavily than those portions of the record containing the noise. However, caution should be exercised in echo detection and extraction when the signal (wavelet) of interest occurs near the end of the data record and thus will be greatly reduced by the exponential window.

Recently it has been proposed that the exponential window be generalized to include complex exponential weighting, i.e., $\alpha^{nT}e^{j\phi nT}$. It may at first appear that this phase factor will have no significant effect on the complex cepstrum, i.e., it will introduce only a phase shift. However, the procedure can be used to change the phase relation of the echo (multipath reflection) by π. This may make it easier to detect a peak in the cepstrum. The complex exponential factor ϕ can be varied in a prescribed fashion so that it may be used as a hypothesis tester. Thus trial sweeps of the complex weight can be generated to confirm or deny a priori estimates of the echo delay. It appears that this technique may prove to be a powerful investigative tool to assist the researcher in interpreting his data.

C.7.1 The Effect of Windowing the Log Spectrum

One might be motivated to window the log spectrum in order to reduce leakage in the complex cepstrum which could be falsely interpreted as peaks due to echoes. Windowing of the log spectrum will, of course, introduce some loss in time resolution in the cepstrum domain. Then, if the echo contributions can be liftered from the complex cepstrum and if the recovered log spectrum can be corrected (by multiplying by the inverse of the windowing series), we should be able to recover the basic wavelet. Our results have, however, indicated that such windowing of the log spectrum raises the echo detection threshold by around 12 dB and severely distorts the recovered wavelet when additive noise is present [Childers et al., 1977]. This is apparently due to the fact that windowing the log spectrum is equivalent to smoothing the complex cepstrum. Thus, it appears that windowing the log spectrum may smooth out the very peaks one wishes to detect in the complex cepstrum. The distortion introduced into the recovered wavelet is undoubtedly due to this windowing of the log spectrum (or smoothing of the complex cepstrum).

C.7.2 The Effect of Windowing the Complex Cepstrum

Since noise is usually interspersed throughout the data record and the composite signal may occupy only a portion of the record, it seems reasonable that the high quefrency components of the complex cepstrum may frequently contain more noise than signal information. Our results have shown that by judiciously zeroing the high quefrency components of the complex cepstrum we may significantly improve the fidelity of the recovered wavelet in a noisy environment [Childers et al., 1977]. At

low SNR the MSE can be reduced by a factor of 2 by a judicious rectangular windowing of the complex cepstrum. This is essentially short pass liftering [Kemerait, 1971, Oppenheim et al., 1968, Schafer, 1968] in which the aim is not to eliminate the echo peaks (which are generally notch filtered prior to the windowing) but rather to eliminate the high quefrency noise dominated sections of the complex cepstrum. This concurs with the results of Kemerait [Kemerait, 1971] in which it is reported that a Hanning smoothing of the log spectrum (which is equivalent to Hanning windowing of the complex cepstrum) improves wavelet recovery. It appears that there is little to choose between the rectangular or Hanning window of the complex cepstrum to improve the fidelity (MSE) of the recovered wavelet. We mention once again that these observations are probably data dependent and are influenced by the duration of the window as well.

Appendix D

Discrete Cepstrum

In this appendix we will discuss the regularized estimation of the discrete cepstrum proposed by Cappé [Cappé et al., 1995, Cappé and Moulines, 1996].

D.1 Regularized Estimation of the Discrete Cepstrum

Cappé revisited the problem of estimation of a cepstrum based spectral envelope from a set of discrete frequency points and proposed many improvements not only in notation and formalization, but also to the method itself. Firstly, they reformulate the problem as obtaining a set of cepstral coefficients c_k such that the log amplitude envelope $\log|P(\omega)|$ evaluated at frequencies ω_k is maximally close to the desired amplitudes of $\log|X(\omega_k)|$. Using this formulation, the source $S(\omega)$ can be neglected and equation 7.82 can be expressed in matrix form as

$$\epsilon = h\left\|a - Mc\right\|^2 = (x - Mc)^T H (x - Mc) \tag{D.1}$$

where

$$x = [\log x_1 \cdots \log x_N]^T \tag{D.2}$$

and

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \dots & 2\cos(2\pi f_1 L) \\ \vdots & \vdots & & \vdots & \\ 1 & 2\cos(2\pi f_N) & 2\cos(2\pi f_N 2) & \dots & 2\cos(2\pi f_N L) \end{bmatrix} \tag{D.3}$$

and c is the vector of unknown cesptral coefficients $c = [c_0 \cdots c_L]^T$ that represent the parameters that minimize the error. H is a diagonal matrix whose elements are $[h_1 \dots h_N]$. Then, the least squares solution is easily found to be

$$c = \left(M^T H M\right)^{-1} M^T H x \tag{D.4}$$

such that, under this formulation, the cepstral coefficients are obtained by a simple matrix inversion, provided that the matrix is invertible (nonsingular), that is, we must have more equations than unknowns, expressed by the condition $L < N$.

When used as described above, the standard discrete cepstrum method is known to yield meaningless results because the matrix $\left(M^T H M\right)$ is frequently poorly conditioned [Cappé et al., 1995]. This means that non-significant perturbations of the data such as machine rounding errors can

induce very large variations of the estimated cepstrum coefficients and of the log-amplitude envelope $\log|P(\omega)|$. To overcome the problems associated with the standard discrete cepstrum method, Galas [Galas and Rodet, 1990] suggested to increase the number N of frequency points by replacing the original ω_k by clusters of neighboring points. Cappé [Cappé et al., 1995, Cappé and Moulines, 1996] proposed an alternative solution based on a well-known regularization technique which consists of imposing additional constraints on the log-amplitude envelope. The idea consists in seeking an envelope which, in addition to minimizing the least-squares criterion in equation D.1, is also smooth, in a sense that will be formulated below. The least-squares criterion is modified as follows

$$\epsilon_r = \sum_{n=1}^{N} h_n \left\| \log(x_n) - \log|P(\omega_n)| \right\|^2 + \lambda\Gamma\left[\log|P(\omega_n)|\right] \tag{D.5}$$

where $\Gamma\left[\log|P(\omega_n)|\right]$ is a penalty functional: Γ is small if the envelope is smooth, and large otherwise and λ is the regularization parameter which controls the relative importance of the smoothness constraint in the criterion to be minimized. As indicated by equation D.5, the new criterion favors envelopes that are close to the specified frequency points (first term in the right member of D.5) while exhibiting some degree of smoothness (second term in the right member of D.5).

A possible smoothness criterion is

$$\Gamma\left[\log|P(\omega_n)|\right] = \int_{-1/2}^{1/2} \left[\frac{d}{df}\log|P(\omega_n)|\right]^2 df \tag{D.6}$$

which is null when $\log|P(\omega_n)|$ is constant. This smoothness criterion can be expressed as a quadratic form of the cepstral coefficients by substituting equation 7.79 into equation D.6. The result is

$$\Gamma\left[\log|P(\omega_n)|\right] = c^T R c \tag{D.7}$$

where R is the following diagonal matrix

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1^2 & 0 & 0 & 0 \\ 0 & 0 & 2^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & L^2 \end{bmatrix} \tag{D.8}$$

Finally, the solution to the modified criterion expressed in equation D.5 is given by

$$c = \left(M^T H M + \lambda R\right)^{-1} M^T H a \tag{D.9}$$

Cappé [Cappé and Moulines, 1996] states that the use of the penalty functional D.6 induces virtually no additional computational cost; it amounts to adding to the diagonal elements of the matrix to be inverted a term which is proportional to the square of the cepstrum rank. However, in order to impose a certain tilt to the spectral envelope, they [Cappé et al., 1995] propose to use an alternative penalty functional of the following form:

$$\Gamma\left[\log|P(\omega_n)|\right] = 2\int_{0}^{1/2} \left[\frac{d}{df}\log|P(\omega_n)| - g_\alpha(\omega_n)\right]^2 df \tag{D.10}$$

where $g_\alpha(\omega_n)$ is a function defined for positive frequencies which forces the envelope to vary with slope $\alpha \propto 6\text{dB}$ per octave.

A possible choice of $g_\alpha(\omega_n)$ is α/ω which necessitates a modification of the definition in equation D.10 due to the diverging behavior of the integral term in $\omega = 0$. Truncated versions of α/ω (such as $g_\alpha(\omega_n) = \alpha/\omega$ if $\omega \geq \omega_c$ and $g_\alpha(\omega_n) = \alpha/\omega_c$ if $\omega < \omega_c$) can also be considered although they generally lead to more complex solutions. Moreover, it is preferable to use more regular functions (at least derivable) for $g_\alpha(\omega_n)$. Other choices of $g_\alpha(\omega_n)$ include $w(\omega)\,\alpha/\omega$ where $w(\omega)$ is a properly chosen window function. In practice, a convenient choice is $g_\alpha(\omega_n) = -\alpha\,(2/\log 2)\log\omega$, which behaves approximately like α/ω only in the high-frequency part (for normalized frequencies above 0.25). Minimization of equation D.10 with this choice of the penalty functional yields

$$c = \left(M^T H M + \lambda R\right)^{-1}\left(M^T H a + \lambda r_\alpha\right) \tag{D.11}$$

where vector $r_\alpha = \alpha\,(16\pi/\log 2)\left[\begin{array}{cccc} 0 & J(1) & J(2) & \cdots \end{array}\right.$ $\left.\begin{array}{c} J(L) \end{array}\right]^T$ and R is defined as in equation D.8. The $J(i)$ are integral terms given by

$$J(i) = \begin{cases} \int_0^{i/2} \sin(2\pi f)\log(f)\,df & \text{if } i \text{ is even} \\ \int_0^{i/2} \sin(2\pi f)\log(f)\,df - \frac{1}{\pi}\log i & \text{if } i \text{ is odd} \end{cases} \tag{D.12}$$

that need to be precomputed by numerical integration.

Appendix E

Line Spectral Frequencies

In this appendix we will see the fundamental theorem of palindromic polynomials, which is the theoretical bias for the line spectral pair (LSP) representation.

E.1 The Fundamental Theorem of Palindromic Polynomials

Given a polynomial $a(x)$ of degree M, let $a_r(x)$ denote its reciprocal, i.e., $a_r(x) = x^M a(1/x)$.

$$a(x) = \sum_{m=0}^{M} a_m x^m = a_0 + a_1 x + a_2 x^2 + \cdots + a_{M-2} x^{M-2} + a_{M-1} x^{M-1} + a_M x^M \tag{E.1}$$

$$a_r(x) = \sum_{m=0}^{M} a_m x^{M-m} = a_M + a_{M-1} x + a_{M-2} x^2 + \cdots + a_2 x^{M-2} + a_1 x^{M-1} + a_0 x^M \tag{E.2}$$

If a polynomial has real coefficients and is equal to its reciprocal, we call it a palindromic polynomial since the coefficients are the same when read backwards or forwards. In other words, if $a(x)$ has real coefficients $\{a_m\}$, it is called palindromic if $a_m = a_{M-m}$ and antipalindromic if $a_m = -a_{M-m}$. It is not hard to show that the product of two palindromic or two antipalindromic polynomials is palindromic, while the product of an antipalindromic polynomial with a palindromic one is antipalindromic. It is possible to prove that every polynomial with real coefficients that has all of its zeros on the unit circle is either palindromic or antipalindromic. The simplest cases are $x+1$ and $x-1$, which are obviously palindromic and antipalindromic, respectively. Next consider a second degree polynomial with a pair of complex conjugate roots on the unit circle

$$\left(x - e^{j\phi}\right)\left(x - e^{-j\phi}\right) = x^2 - \left(e^{j\phi} + e^{-j\phi}\right)x + e^{j\phi}e^{-j\phi} = x^2 - 2\cos\phi x + 1 \tag{E.3}$$

which is palindromic. Any polynomial with real coefficients that has k pairs of complex conjugate roots will be the product of k palindromic polynomials, and thus palindromic. If a polynomial has k pairs of complex conjugate roots and the root $x = 1$ it will also be palindromic, while if it has the root $x = -1$ it will be antipalindromic. The converse of this statement is not necessarily true; not every palindromic polynomial has all its zeros on the unit circle [Konvalina and Matache, 2004]. The idea behind the LSFs is to define palindromic and antipalindromic polynomials that do obey the converse rule.

Any arbitrary polynomial $a(x)$ can be written as the sum of a palindromic polynomial $p(x)$ and an antipalindromic polynomial $q(x)$

$$a_m = \frac{1}{2}(p_m + q_m) \tag{E.4}$$

where

$$\begin{cases} p_m = a_m + a_{M-m} \\ q_m = a_m - a_{M-m} \end{cases} \tag{E.5}$$

(if M is even the middle coefficient appears in p_m only). When we are dealing with polynomials that have their constant term equal to unity, we would like the polynomials p_m and q_m to share this property. To accomplish this we need only pretend for a moment that a_m is a polynomial of order $M+1$ and use the above equation with $a_{M+1} = 0$.

$$a_m = \frac{1}{2}(p_m + q_m) \tag{E.6}$$

$$\begin{cases} p_m = a_m + a_{M+1-m} \\ q_m = a_m + -a_{M+1-m} \end{cases} \tag{E.7}$$

Now $a_0 = p_0 = q_0 = 1$ but p_m and q_m are polynomials of degree $M+1$.
Formally we can write the relationships between the polynomials

$$a(x) = \frac{1}{2}(p(x) + q(x)) \tag{E.8}$$

where

$$\begin{pmatrix} p(x) \\ q(x) \end{pmatrix} = a(x) \pm x^{M+1} a\left(x^{-1}\right) \tag{E.9}$$

and it is not hard to show that if all the roots of $a(x)$ are inside the unit circle, then all the roots of $p(x)$ and of $q(x)$ are on the unit circle [Soong and Juang, 1984]. Furthermore, the roots of $p(x)$ and $q(x)$ are intertwined, i.e., between every two roots of $p(x)$ there is a root of $q(x)$ and vice versa. Since these roots are on the unit circle they are uniquely specified by their angles. For the polynomial in the denominator of the LPC frequency response these angles represent frequencies, and are called the line spectral frequencies. Why are the LSFs a useful representation of the all-pole filter? The LPC coefficients are not a very homogeneous set, the higher-order being more sensitive than the lower-order ones. LPC coefficients do not quantize well; small quantization error may lead to large spectral distortion. Also the LPC coefficients do not interpolate well; we can't compute them at two distinct times and expect to accurately predict them in between. The errors of the LPC polynomial are a better choice, since they all have the same physical interpretation. However, finding these zeros numerically entails a complex two-dimensional search, while the zeros of $p(x)$ and $q(x)$ can be found by simple one-dimensional search techniques. In speech applications it has been found empirically that the LSP frequencies quantize well and interpolate better than all other parameters that have been tried.

Appendix F

Perceptual Similarity for Musical Instrument Sound

The aim of this listening test is to compare the perceptual similarity between the original recording and a model of musical instrument sounds.

F.1 Sound Representations

There are many different possible representations of sounds. Some representations sound different from the original recording, depending on the model. The mp3 compression is a popular example of a lossy encoding that may sound different from the original. One important aspect of sound representations is the perceptual similarity between the original sound and its representation.

F.2 The Test

In this listening test you will be asked to compare the perceptual similarity between the original recording and a representation for musical instrument sounds. Below, you will find a table with 20 lines. Each line contains

- the original sound;
- the model.

Notice that each line has different versions of the same sound. These versions might sound different from one another, but they come essentially from the same recording.

F.3 Framework

For each line of the table, you will hear two sounds. The original recording and the model.

Listen once to all the sounds to get used to the range of differences between them. Only after listening to all of the sounds once should you start the test.

F.4 Perceptual Similarity

When we judge the perceptual similarity of sounds, we are trying to assess how close the sounds are on the perceptual plane. In other words, we want to determine if they sound the same or different. Naturally, when they sound different we can additionally judge how different, just a little or a lot.

F.5 Your Task

Your task is to listen to the original sound and to the model and rate their perceptual similarity using the scale given

- identical;
- slightly different;
- fairly different;
- very different;
- significantly different.

Before taking the test, listen once to all the sounds to get used to the range of differences.

F.6 Recommendations

- Check that the Flash plug-in works correctly and the sound level is properly set;
- Use headphones;
- Do the test in a quiet place;
- Before running the test, do not hesitate to send me an e-mail if you have questions.

Thank you for participating! This experiment won't take you more than 5 minutes.

F.7 Listening conditions

Did you use headphones?

- yes
- no

Did you listen to all the sounds once before taking the test?

- yes
- no

Expertise (are you familiar with a domain related to music, such as acoustic signal processing, music technology, ... ?)

- yes
- no

Appendix G

Evaluation of the Spectral Smoothness of Sound Morphing Algorithms

The aim of this listening test is to compare the smoothness of two different morphing algorithms applied to musical instrument sounds.

G.1 Sound Morphing

The principle of sound morphing is to gradually transform a source sound to become more and more similar to a target sound. One important factor when judging the quality of a sound morphing algorithm is the smoothness of the transformation.

G.2 The Test

In this listening test you will be asked to compare the smoothness of two different morphing algorithms. Below, you will find a table with 11 pairs of morphing transformations. Each pair is a morph between the same source and target sounds using one of the algorithms. The transformation on the lefthand side is always labeled Morph A, and the transformation on the righthand side is always labeled Morph B. However, sometimes Morph A corresponds to one algorithm, sometimes to the other, in order not to bias the results.

G.3 Framework

For each morphing algorithm, you will hear eleven sounds. The source sound labeled 0.0, nine intermediate versions labeled 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and the target sound labeled 1.0. The intermediate versions are a morphing transformation from the source sound labeled 0.0 to the target sound labeled 1.0, and they should start similar to the first sound and become more and more similar to the second sound as the labels increase from 0.1 to 0.9.

G.4 Smoothness

A smooth morphing algorithm should produce intermediate versions that become gradually or smoothly similar to the target sound. In other words, the intervals between successive sounds should be the same, without bumps or sudden discontinuities.

As an example of a bumpy morph, listen to the following transformation paying careful attention to the difference between the intervals. In this example, all the intervals sound very different from one another.

- Bumpy Morph

Now listen to an example of a smoother morph between the same sounds.

- Smoother Morph

G.5 Your Task

Your task is to listen to both morphs on each row of the table below and compare their smoothness. Click on the button corresponding to the morph that you judge to be the smoother between the two, Morph A or Morph B. If they both sound as smooth to you then click on the no preference button.

Remember that sometimes Morph A corresponds to one morphing algorithm being compared, sometimes to the other in order not to bias the results. In other words, if you always prefer Morph A this does NOT mean that you always prefer one morphing algorithm.

Notice that you should try to judge only the smoothness of the morph, not the synthesis quality (possible synthesis artifacts must NOT affect your judgment).

G.6 Recommendations

- Check that the Flash plug-in works correctly and the sound level is properly set;
- Use headphones or earphones;
- Do the test in a quiet place;
- Before running the test, do not hesitate to send me an e-mail if you have questions.

Thank you for participating! This experiment won't take you more than 10 minutes.