SOUND MORPHING BY FEATURE INTERPOLATION

Marcelo Caetano, Xavier Rodet

Analysis/synthesis Team, IRCAM

ABSTRACT

The goal of sound morphing by feature interpolation is to obtain sounds whose values of features are intermediate between those of the source and target sounds. In order to do this, we should be able to resynthesize sounds that present a set of predefined feature values, a notoriously difficult problem. In this work, we present morphing techniques to obtain hybrid musical instrument sounds whose feature values correspond as close as possible to the ideal interpolated values. When the features capture perceptually relevant information, the morphed sound whose features are interpolated is perceptually intermediate. The features we use are acoustic correlates of salient timbre dimensions derived from perceptual studies, such that sounds whose feature values are intermediate between two would be placed between them in the underlying timbre space. We measure the perceptual impact of the morphed sounds directly by the feature values, using them as an objective measure with which to evaluate the results. Thus we consider that the morphed sounds change perceptually linearly when the corresponding feature values vary linearly.

Index Terms— Sound morphing, sonic features, musical instrument sound, timbre space

1. INTRODUCTION

Sound morphing has been used in music compositions [1], [2], [3], in synthesizers [4], and even in psychoacoustic experiments, notably to study timbre spaces [5]. When morphing musical instrument sounds, we usually want to obtain hybrid sounds that are perceptually intermediate across timbre dimensions, such that the intermediate sounds would correspond to hybrid instruments between source and target. A challenging aspect is to control the transformation with a single parameter α, called morphing or interpolation factor [6], as illustrated in Fig. 1 for images. Figure 1 shows that, ideally, we want the morphing factor α to control perceptually related features of the transformation, such that the morph should be perceptually halfway when α = 0.5, for instance. 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 features [4], [7], [8]. In this work, parameter refers to coefficients from which we can 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. The basic idea behind the interpolation principle depicted in Fig. 2 is that if we can represent different sounds by simply adjusting the parameters of a model, we should obtain a somewhat smooth transition between two (or more) sounds by interpolating between these parameters These authors often conclude that the linear interpolation of the parameters of known models does not correspond to linearly varying perceptually relevant features [7]. If we want the result to sound perceptually intermediate, we need to develop techniques to interpolate perceptually motivated features. For such, we adopted the morphing by feature interpolation principle, depicted in Fig. 2. Ideally, we would like to be able to interpolate in the feature space and retrieve the set of parameters that correspond to the interpolated feature values. Unfortunately, this is a notoriously difficult problem to solve since most features commonly used do not allow direct inversion for resynthesis, particularly when the features are correlated to perceptual characteristics of sounds [9], [10], [11]. In this work, we describe techniques to obtain morphed sounds whose values of features are as close as possible to the interpolated feature values. Ideally, we want the feature values to vary linearly when the morphing factor varies linearly. Most features we use (log attack time, temporal centroid, spectral centroid, spread, skewness, kurtosis [12]) are acoustic correlates of timbre dimensions obtained by perceptual studies, such that sounds whose feature values are intermediate between two would be placed between them in the underlying timbre space used as guide. We measure the perceptual impact of the morphed sounds directly by the feature values, using them as an objective measure with which to evaluate the results. Thus we consider that the morphed sounds change perceptually linearly when the corresponding feature values vary linearly.

The next section introduces timbre spaces obtained in psychoacoustic experiments and the most salient dimensions of timbre perception unveiled. Then, we present the sound model we

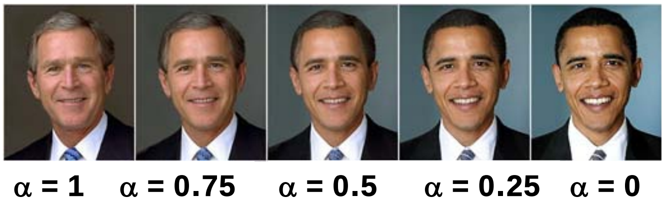

Fig.1. Depiction of image morphing to exemplify the aim of sound morphing. Original image from http://i40.tinypic.com/11tqy52.jpg.

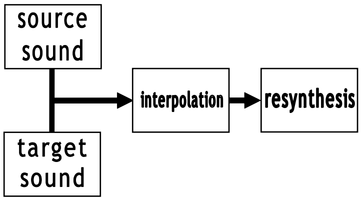

Fig. 2. Morphing using the interpolation principle.

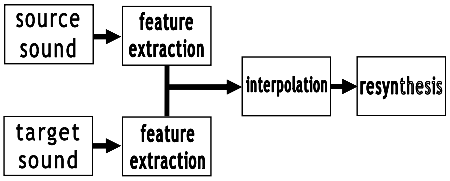

Fig. 3. Morphing by feature interpolation. Ideally we want to interpolate the feature values and retrieve the set of parameters that correspond to the interpolated features.

developed, which includes amplitude and spectral envelope estimation, temporal segmentation, and sinusoidal and residual modeling. Finally, we explain the feature interpolation techniques that we developed specifically for musical instrument sounds.

2. ACOUSTIC CORRELATES OF TIMBRE SPACES

Multi Dimensional Scaling (MDS) techniques figure among the most prominent when trying to quantitatively describe timbre. McAdams [11] and Handel [13] independently propose comprehensive reviews of the early timbre space studies. Grey [5] investigated the multidimensional nature of the perception of musical instrument timbre, constructed a three-dimensional timbre space, 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 [14] conducted a similar study using synthesized sounds and also found three dimensions related to attack, synchronicity and brightness. Krimphoff [10] 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 [11] 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 [9] 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. Here we should notice that most MDS techniques suppose that the underlying space is orthogonal and metric, which means that the dimensions are independent and the notion of distance is defined. Therefore, shifting linearly from a source to a target sound in such space would correspond to a perceptually linear change across all dimensions and would also result in linear variation of the values of the correlates of each dimension.

2.1. The Features Used as Guides

We included temporal and spectral features related to the acoustic correlates of timbre spaces presented earlier [12] and are supposed to capture the most perceptually salient dimensions of the timbre spaces proposed, namely, the attack time and the distribution of spectral energy. The temporal features we use are the log attack time and the temporal centroid. The spectral shape features we take into consideration are spectral centroid, spread, skewness and kurtosis.

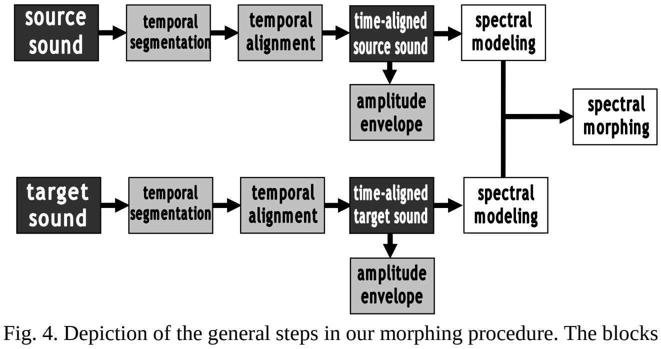

Fig. 4. Depiction of the general steps in our morphing procedure. The blocks represent temporal and spectral feature extraction and processing.

3. MODELING PERCEPTUALLY RELEVANT FEATURES

The morphing technique we developed is depicted in Figure 4, and consists of two basic steps, temporal segmentation and consequent temporal alignment followed by spectral modeling and spectral morphing. The temporal processing steps are represented by light grey blocks and spectral processing by blocks with white background. In general terms, the sound model consists in a temporal segmentation step, amplitude envelope estimation, sinusoidal plus residual modeling and spectral envelope extraction.

3.1 Temporal Modeling

Temporal modeling comprises the temporal segmentation step and amplitude envelope estimation. The temporal segmentation, shown in Fig. 5, consists in automatically estimating the boundaries of four perceptually important regions, namely, the attack, the transition, the steady-state or sustain, and the release using an automatic segmentation technique we proposed elsewhere [15]. From these estimations, we calculate the length of each region. The most salient descriptor here is log attack time. Amplitude envelope estimation is performed with the true amplitude envelope technique we developed [15], based on cepstral smoothing. The interpolated amplitude envelope modulates the spectral frames of the morphed sound, so we use the temporal centroid as its feature.

3.2 Spectral Modeling

The basic spectral modeling technique used is harmonic sinusoidal modeling plus noise residual [16]. In order to independently manipulate the spectral shape, we model the amplitude of the partials using a spectral envelope model instead of the amplitudes output by the sinusoidal analysis. Fig. 6 illustrates both representations of the sinusoidal model. In part a) we show the traditional sinusoidal representation, where each partial has a frequency and associated amplitude tied together. In part b), however, the spectral envelope controls the global spectral shape independently from the frequencies of the partials. For every frame of both the sinusoidal and noise residual we calculate the spectral envelope using true envelope [17]. The sinusoidal component is modeled as shown in part b) of Fig. 6, and the noise residual is modeled as white noise filtered with the spectral envelope estimated from each frame of the noise residual.

4. MORPHING STRATEGY AND EVALUATION WITH FEATURE VALUES

The morphing steps are as follows: temporal alignment, spectral envelope morphing and amplitude envelope morphing. We use the values of features as a guide to measure the perceptual impact of each step. Ideally, we want the features to vary linearly when the morphing factor varies linearly.

4.1 Temporal Alignment

The first important step in morphing is the temporal alignment of perceptually different regions, such as the attack, characterized by fast transients, and the sustain part, much more stable. We cannot

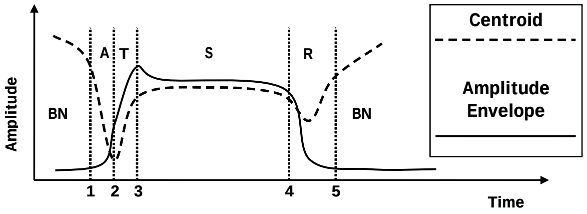

Fig. 5. Amplitude/Centroid Trajectory model. The figure shows the regions indicated by letters and their boundaries indicated by numbers.

expect to attain good results if we combine a sound that has a long attack with another sound with a short one regardless of the differences. The region where attack transients are combined with more stable partials will not sound natural. So, with this in mind, to achieve a more perceptually seamless morph, we need to temporally align these regions so that their boundaries coincide.

For each sound, we measure the length of each region (labeled with letters) by computing the time difference using the markers (numbers). The length of the attack is represented in logarithmic scale. For the other regions the representation is linear. Then, we interpolate between the lengths of the regions according to equation (1) to obtain their corresponding lengths in the morphed sound. The interpolated lengths are represented by a letter that stands for the region and subscripts indicating both sounds, e.g. S_{12} for the sustain as shown in equation (1)

$$S_{12} = \alpha S_1 + [1-\alpha] S_2 \qquad (1)$$

where S_1 represents the length of the sustain of the first sound and S_2 of the second. The stretch/compress factors R_{S1} for the first sound and R_{S2} for the second are calculated as in equation (2)

$$R_{S1} = \frac{S_1}{S_{12}} \qquad R_{S2} = \frac{S_2}{S_{12}} \qquad (2)$$

Finally, we simply time stretch/compress each region by the corresponding ratio, for instance, S_1 by R_{S1}, etc. Since we interpolate the attack time logarithmically, the attack times (perceived on a logarithmic scale [9], [11]) will be perceived as varying linearly under the time-alignment transformation. The other regions will also be properly aligned and ready to be morphed in the spectral domain.

4.2 Spectral Envelope Morphing

Following the morphing by feature interpolation principle, the objective of the spectral envelope morphing step is to obtain a morphed spectral envelope that has intermediate formant peaks and intermediate values of spectral shape features. In order to achieve this, the ideal would be to be able to interpolate the spectral feature values and invert this representation to obtain spectral envelope parameters corresponding to the interpolated feature values. One difficulty this approach poses is that there is no known analytic inversion from the chosen features. Instead, we will study which spectral envelope representation leads to linearly varying values of spectral shape features when its parameters are linearly interpolated. The representations investigated are the envelope curve (ENV) [4], [8], line spectral frequencies (LSF) [18], cepstral coefficients (CC) [7] and dynamic frequency warping (DFW) [19]. These correspond to the main sound morphing methods proposed in the literature.

Figure 7 shows the source and target envelopes in solid lines and nine intermediate envelopes corresponding to linearly varying the interpolation factor α by 0.1 steps in dashed and dotted lines; on the right, we see the associated values of the spectral shape descriptors for each step. Figure 7 confirms that for this case interpolating envelope curves does not account for formant shifting and most spectral shape descriptors do not vary in a straight line. Figure 7 also shows that the linear interpolation of cepstral based envelope representations like Slaney [7] proposes neither shifts the formants nor results in linear variation of descriptors. The same applies for the DFW based spectral envelope morphing proposed by Ezzat [19]. On the other hand, LSFs behave fairly well under both constraints in this case just like Paliwal [18] states for LSFs. In conclusion, in our model we adopt LSFs as the most suitable parameters to represent and interpolate the spectral envelopes.

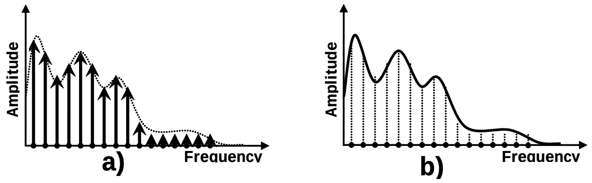

Fig. 6. Spectral representation of partials. The figure shows the traditional sinusoidal representation with the frequency values and amplitudes tied to each other in part a). Part b) depicts our representation, where the amplitudes of the partials are represented independently with a spectral envelope model.

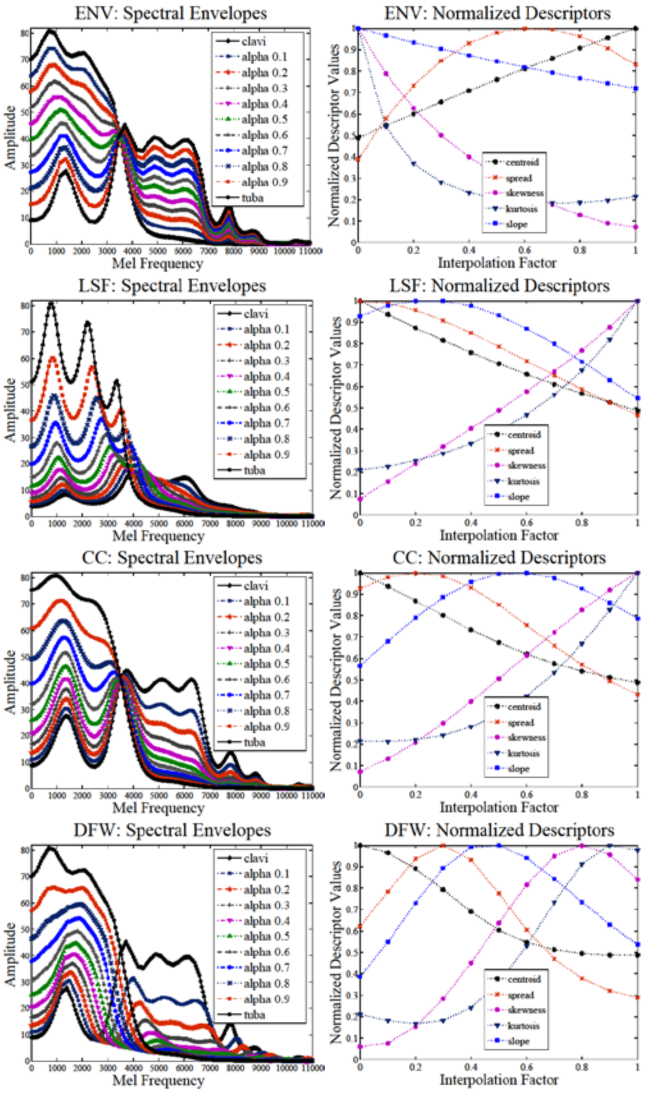

Fig. 7. Spectral envelope morphing guided by high-level spectral shape descriptors. The figure shows the perceptual impact of interpolating between the parameters of several spectral envelope models. The curves are shown on the left and the corresponding feature variation on the right. We want the spectral envelope representation whose linear interpolation of parameters leads to linear variation of spectral shape feature values.

4.3 Interpolation of Partial Frequencies

Since we are morphing quasi-harmonic musical instrument sounds, there exists a direct one to one correspondence between the partials of both sounds. All we have to do is make sure we will be able to

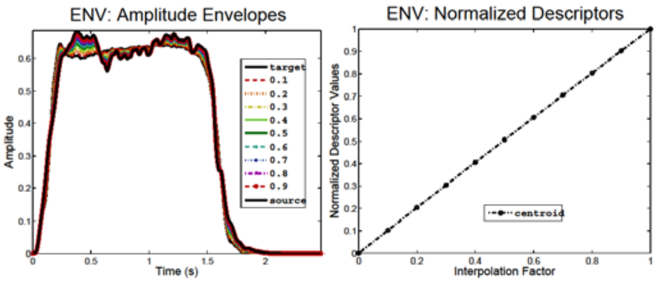

Fig. 8. Amplitude envelope morphing guided by high-level temporal shape descriptors. The figure shows the variation of the temporal centroid value when interpolating between the parameters of the amplitude envelope. The curves are shown on the left and the corresponding descriptor variation on the right.

fuse them into a single percept without losing the inherent spectral variations. Here we propose to morph quasi-harmonic musical instrument sounds with the same pitch, so that the partials have a one to one correspondence and no pitch shift is required. Since the spectral shape and form are morphed separately with the spectral envelope, we simply interpolate the partials frequency values to account for frequency fluctuations (jitter, shimmer), inharmonicity and other temporal features that are encoded in the frequency variation with time.

4.4 Morphing the Amplitude Envelope

Morphing the amplitude envelope is similar to the spectral envelope because the techniques we developed for estimating the amplitude envelope are inspired by spectral envelope estimation techniques [15]. Also, the temporal centroid is the time-domain analogous of the spectral centroid [12], and as such, its values behave in the same fashion under the same transformations. Fig. 8 confirms that the temporal centroid varies fairly linearly even when we interpolate the amplitude envelope curves directly.

4.5 Morphing the Noise Residual

We morph the spectral envelopes of the residual noise signal and synthesize a morphed residual by filtering white noise with it and mixing it into the morphed sinusoidal component. The noise residual is obtained simply by filtering white noise with a time-varying filter obtained from the interpolation of the spectral envelopes extracted from the residual of the original sounds. We interpolate the LSFs used to represent the spectral envelope of the noise residual just like for the sinusoidal part.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, we describe techniques to morph salient timbral dimensions of quasi-harmonic musical instrument sounds using features as guides. The features we use are acoustic correlates of timbre dimensions obtained in psychoacoustic studies, such that sounds whose feature values are intermediate between two would be placed between them in the underlying timbre space. So, interpolating the feature values becomes the goal itself to render the results more perceptually linear. For such, we introduced the concept of morphing by feature interpolation and developed a model and techniques to apply it. The sound model aims at allowing independent manipulation of perceptually meaningful features of sounds related to salient timbre dimensions, such as spectral and amplitude envelopes, attack time, among others. We measured the perceptual impact of the morphed sounds directly by the feature values. We consider that the morphed sounds change perceptually linearly when the corresponding feature values vary linearly. Future perspectives could include vibrato and tremolo modeling and treatment, and extending the technique to inharmonic sounds, which would probably require a technique to find correspondences between the partials of both sounds. Sound examples on http://recherche.ircam.fr/anasyn/caetano/icassp2011.html.

6. ACKNOWLEDGEMENTS

This work is supported by the Brazilian Governmental Research Agency CAPES (process 4082-05-2).

7. REFERENCES

[1] T. Wishart, *On Sonic Art*.: Simon Emerson: Harwood Academic Publishers, 1998.

[2] M. McNabb, "Dreamsong: The Composition," *Computer Music Journal*, vol. 5, no. 4, pp. 36-53, 1981.

[3] J. Harvey, "Mortuos Plango, Vivos Voco: A Realization at IRCAM," *Computer Music Journal*, vol. 5, no. 4, pp. 22-24, 1981.

[4] L. Haken, B. Holloway E. Tellman, "Timbre Morphing of Sounds with Unequal Numbers of Features," *J. Audio Eng. Soc.*, vol. 43, no. 9, pp. 678-689, 1995.

[5] J.A. Moorer J.M. Grey, "Perceptual Evaluations of Synthesized Musical Instrument Tones," vol. 62, no. 2, pp. 454-462, 1977.

[6] X. Rodet M. Caetano, "Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors," in *Proc. ICMC*, 2010.

[7] M. Covell, B. Lassiter M. Slaney, "Automatic Audio Morphing," in *Proc. ICASSP*, 1996.

[8] L. Haken K. Fitz, "Sinusoidal Modeling and Manipulation Using Lemur," *Computer Music Journal*, vol. 20, no. 4, pp. 44-59, 1996.

[9] S. McAdams, B.K. Smith, S. Winsberg A. Caclin, "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 471-482, 2005.

[10] S. McAdams, S. Winsberg J. Krimphoff, "Caractérisation du Timbre des sons Complexes. II: Analyses Acoustiques et Quantification Psychophysique," *Journal de Physique*, vol. 4, no. C5, pp. 625-628, 1994.

[11] S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff S. McAdams, "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specifities and Latent Subject Classes," *Psychol. Res.*, vol. 58, pp. 177-192, 1995.

[12] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, Project Report 2004.

[13] S. Handel, *Timbre perception and auditory object identification*, B.C.J. Moore, Ed. New York: Academic Press, 1995.

[14] C.L. Krumhansl, *Why is Musical Timbre So Hard to Understand?*, Structure and Perception of Electroacoustic Sound and Music ed., S. Nielzén and O. Olsson, Ed. Amsterdam: Excerpta Medica, 1989.

[15] J. J. Burred, X. Rodet M. Caetano, "Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues," in *Proc. DAFx*, 2010.

[16] X. Serra, *Musical Sound Modeling with Sinusoids Plus Noise*, Musical Signal Processing ed.: Swets & Zeitlinger, 1997.

[17] X. Rodet A. Röbel, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation," in *Proc. DAFx*, 2005.

[18] K. Paliwal, "Interpolation Properties of Linear Prediction Parametric Representations," in *Proc. Eurospeech*, 1995, pp. 1029-1032.

[19] E. Meyers, J. Glass, T. Poggio T. Ezzat, "Morphing Spectral Envelopes using Audio Flow," in *Proc. ICASSP*, 2005, pp. 1029-1032.