SPECTRAL SIMILARITY METRICS FOR SOUND SOURCE FORMATION BASED ON THE COMMON VARIATION CUE

Mathieu Lagrange

Telecom ParisTech
46, rue Barrault
75634 PARIS Cedex 13 - FRANCE
lagrange@telecom-paristech.fr

Martin Raspaud

Linköping University
Bredgatan 33
SE-60174 Norrköping - SWEDEN
martin.raspaud@itn.liu.se

ABSTRACT

Scene analysis is a relevant way of gathering informations about the structure of an audio stream. For content extraction purposes, it also provides prior knowledge that can be taken into account in order to provide more robust results for standard classification approaches.

In order to perform such scene analysis, we believe that the notion of temporality is important. We study in this paper a new way of modeling the evolution over time of the frequency and amplitude parameters of spectral components. We evaluate the benefits of such an approach by considering its ability to automatically gather the components of the same sound source. The evaluation of the proposed metric shows that it achieves good performance and take better account of micro-modulations.

1. INTRODUCTION

Extracting content from polyphonic audio such as musical streams appears to be bounded to moderate performance if the stream is considered “blindly”, *i.e.* processed without any prior knowledge of the structure of the stream. As scene analysis is a relevant way of gathering informations about the structure of an audio stream, performing such operation prior extracting content is a way to address this issue.

On the high end, one can consider a mid-level representation of the polyphony [1, 2] describing polyphonic sounds as a set of coherent spectral regions, where each set can be considered as monophonic. In this case, one can focus the content extraction process to a given element of the scene [3]. On a lower end, one can consider some time segmentation of the audio stream where sections that have similar properties are identified and/or clustered. Based on this representation, the temporal priors are considered to integrate the indexing decision done at each analysis frame to obtain more robust classification results [4].

In order to extract such representation or segmentation, many cues can be considered [5]. As far as timbral cues are considered, the common variation cue [6] is of interest as it encodes temporal dynamics. In this paper, we focus on this cue to propose new metrics for defining the similarity between spectral components using the sinusoidal model.

The paper is organized as follows: after a presentation of the sinusoidal model in Section 2, existing metrics proposed in the literature are reviewed in Section 3 and the requisites of a relevant a metric are also detailed.

The proposed metric is next introduced in Section 4. Motivated by the properties of the evolutions of the frequencies of the partials, a first metric is proposed. We next show that this metric can also be successfully used while considering the evolutions of the amplitudes as soon as the variations of the envelope is removed. The definition of a metric that jointly considers these two cues is next studied.

In order to compare existing metrics to the ones introduced in this article, we use the evaluation methodology presented in Section 5. In particular, the database and the criteria that evaluate the ability of the tested metric to discriminate partials produced from different instruments. The results of this evaluation are presented in section 6.

2. MID-LEVEL REPRESENTATION OF POLYPHONIC SOUNDS

For various applications, one needs a representation of polyphonic sounds where the frequency information as well as its evolution with respect to time for each sound sources can easily be extracted. In this section, we discuss the fact that the well-known sinusoidal model can be a basis for such a representation.

The sinusoidal model represents pseudo-periodic sounds as sums of sinusoidal components – so-called partials – controlled by parameters that evolve slowly with time [7, 8]:

$$P_k(m) = \{F_k(m), A_k(m), \Phi_k(m)\} \tag{1}$$

where $F_k(m)$, $A_k(m)$, and $\Phi_k(m)$ are respectively the frequency, amplitude, and phase of the partial P_k at time index m. These parameters are valid for all $m \in [b_k, \cdots, b_k + l_k -$

1], where the b_k and l_k are respectively the starting index and the length of the partial.

These sinusoidal components are called partials because they are only a part of a more perceptively coherent entity that will be noted in this article an acoustical entity.

Thus, this can be written as:

$$\mathcal{S} = \bigcup_{n=1}^{N} E_n \tag{2}$$

with $\mathcal{S}$ being the mid-level representation of the sound, E being an acoustical entity and N the total number of entities in the sound. Hence each entity is made of a group of partials:

$$E_n = \bigcup_{k=1}^{M_n} P_k^n \tag{3}$$

where M_n is the total number of partials P_k^n in the entity.

The partials can be extracted from polyphonic sounds with dedicated tracking algorithms [9]. However, in order to avoid problems due to strong polyphony [1], we only consider here mixtures of already tracked entities.

To extract these entities from a sinusoidal representation of a sound, similarities between partials should be considered in order to gather the ones belonging to the same acoustical entity. From the perceptual point of view, some partials belong to the same entity if they are perceived by the human auditory system as a unique sound. There are several cues that lead to this perceptual fusion: the common onset, the harmonic relation of the frequencies, the correlated evolutions of the parameters and the spatial location [5].

The earliest attempts at acoustical entity identification and separation consider harmonicity as the sole cue for group formation. Some rely on a prior detection of the fundamental frequency [10, 11] and others consider only the harmonic relation of the frequencies of the partials [12, 13, 14]. Yet, many musical instruments are not perfectly harmonic.

According to the work of McAdams [6], a group of partials is perceived as a unique acoustical entity only if the variations of these partials are correlated. Therefore, the correlated evolutions of the parameters of the partials is a generic cue since it can be observed with any vibrating instruments. As an example, see Figure 1.

In order to define a dissimilarity metric that considers the common variation cue, we will study in the next section the physical properties of the evolutions of the frequency and amplitude parameters of the partials.

3. THE COMMON VARIATION CUE

Let us consider a harmonic tone modulated by a vibrato of given depth and rate. All the harmonics are modulated at the same rate and phase but their respective depth is scaled by a factor equal to their harmonic rank, see Figure 2(a). It is then important to consider a metric which is scale-invariant.

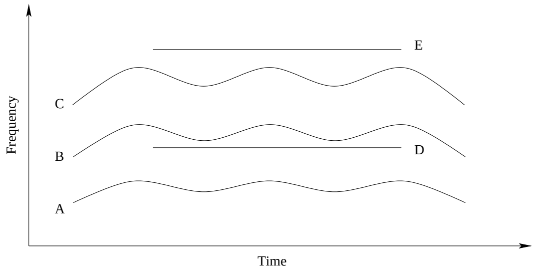


Fig. 1. *Representation of two fictive sounds in the time-frequency domain. Partials A, B, and C (clearly correlated in modulation and starting and ending times, that is common variation) represent the sinusoidal components of the first sound, while D and E represent the sinusoidal components of the second sound.*

M. Cooke uses a distance [15] equivalent to the cosine dissimilarity d_c, also known as *intercorrelation*:

$$\begin{aligned} d_c(X_1, X_2) &= 1 - \frac{c(X_1, X_2)}{\sqrt{c(X_1, X_1)}\sqrt{c(X_2, X_2)}} && (4) \\ c(X_1, X_2) &= \sum_{i=1}^{N} X_1(i)\, X_2(i) && (5) \end{aligned}$$

where X_1 and X_2 are real vectors of size N. In this article, X_1 and X_2 will be the frequency and amplitude of a partial over time. This dissimilarity is scale-invariant.

T. Virtanen *et al.* proposed (in [13]) to use the mean-squared error between the vectors first normalized by their average values:

$$d_v(X_1, X_2) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_1(i)}{\bar{X}_1} - \frac{X_2(i)}{\bar{X}_2} \right)^2 \tag{6}$$

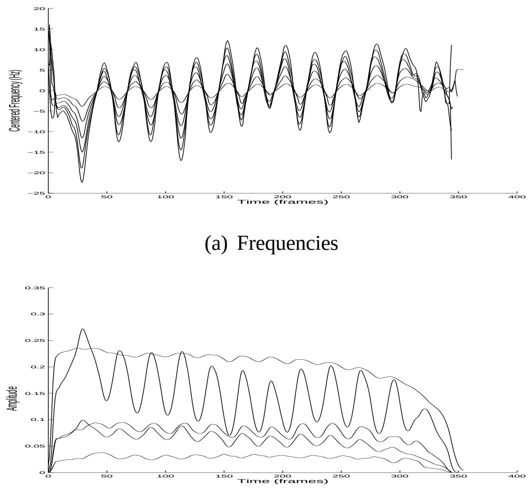
where X_1 and X_2 are vectors of size N and $\bar{X}$ denotes the mean of X. This normalization is particularly relevant while considering the frequencies since the ratio between the mean frequency of a given harmonic and the one of the fundamental is equal to its harmonic rank.

We proposed in [16] to consider the Auto-Regressive (AR) model as a scale-invariant metric that considers only the predictable part of the evolutions of the parameters:

$$X_l(n) \approx \sum_{i=1}^{k} K_l(i) X_l(n-i) \tag{7}$$

where the $K_l(i)$ are the AR coefficients. Since the direct comparison of the AR coefficients computed from the two vectors X_1 and X_2 is not relevant, the spectrum of these coefficients is compared as proposed by Itakura [17]:

$$d_{\text{AR}}(X_1, X_2) = \log \int_{-\pi}^{\pi} \frac{|K_1(\omega)|}{|K_2(\omega)|} \frac{d\omega}{2\pi} \tag{8}$$



(a) Frequencies

(b) Amplitudes

Fig. 2. *Mean-centered frequencies and amplitudes of some partials of a saxophone tone with vibrato.*

where

$$K_l(\omega) = 1 + \sum_{i=1}^{k} K_l(i) e^{-ji\omega} \tag{9}$$

When considering the amplitudes of the partials, a scale-invariant metric is also important. In this context, the normalization proposed by T. Virtanen is no longer motivated since the relative amplitudes of the harmonics depend on the envelope of the sound. For example, on Figure 2(b), the topmost curve (with small modulations) represents the amplitudes of the fundamental partial, while the second to the top curve with broad oscillation represents the first harmonic.

Moreover the envelope is globally decreasing as the frequency grows, but it can appear that the amplitude of the envelope is also ascending due to the specific shape of the envelope around formants. Therefore, when the frequency of a partial is modulated, the amplitude may be modulated with a phase shift, see the bottom curve of Figure 2(b). Therefore, a metric that is phase-invariant should be considered.

The amplitude evolution of a partial is composed of a temporal envelope and some periodic modulations. Since the envelope of the amplitude of the partials can be very different from partials to partials of the same entity it may be useful to consider only the periodic modulations while computing their similarities.

The metric introduced in the next section will cope with these issues.

4. PROPOSED METRIC

The aim of proposing a new metric is to go beyond temporal domain by taking the parameters to the spectral domain. There was already an attempt at this, using AR models (see equation 8).

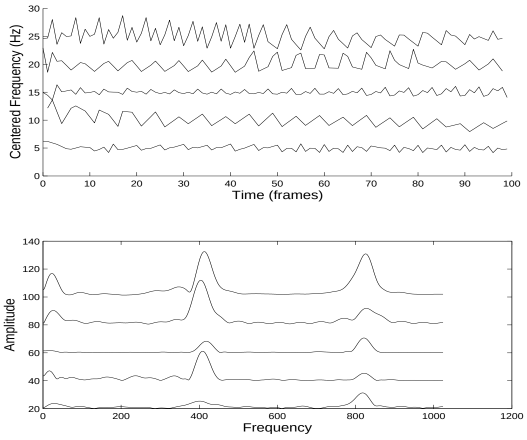


Fig. 3. *Centered frequencies (top) of a piano note and their corresponding spectra (bottom). Each curve is shifted and the spectra are smoothed using zero-padding for clarity sake.*

Since the Fourier transform is based on the fact that the input signal is periodic, using a spectrum of the evolutions of the partials might show common periodicities of the partials. This will be handy for the modulations of the partials created by vibrato and tremolo, since we can assimilate these modulations to sinusoidal ones over a short period of time (see [18, 19]). It can be also interesting for micro-modulations such as the ones produces by vibrating strings such as the strings of a piano (see Figure 3). Hence, the spectrum of the evolutions in frequency and amplitude of the sound are relevant from the point of view of the correlation of evolutions.

In this section, we explain how we compute the correlation of evolutions in order to obtain our new metric, first for the frequency parameters of the sound, second for the amplitude parameters of the sound (since two slightly different methods are used).

4.1. Using the Frequencies of the Partials

The first step in the calculation of our new metric is to correlate the evolutions of the frequencies of the partials. As we said before, a good description of these evolutions is given by the spectra of these evolutions.

The way to compute the spectra of the frequency evolutions of the signal from a partial is to take off the mean value of this frequency and then compute the Fourier transform of the resulting signal. Indeed, in order to have a clean spectrum relevant to the evolutions, it is necessary to have the evolutions centered around zero.

Then, we apply the previously exposed process to the frequencies of all the partials from which we want to measure evolution correlation. Once we have these frequencies expressed in terms of spectra, the way to compute the distance between two partial signals is to intercorrelate their spectra (see equation 4). This gives

$$d_s(f_1, f_2) = d_c(|F_1|, |F_2|) \tag{10}$$

where f_1 and f_2 are the frequency vectors of two partials P_1 and P_2 and F_k is the Fourier spectrum of f_k, f_k being the frequencies of partial P_k.

4.2. Using the Amplitudes of the Partials

In the case of the amplitudes of the partials, the problem is slightly more complicated. Indeed, in order to center the oscillating part of the signal around zero subtracting the mean will not be sufficient. As presented in other works [20], subtracting a polynomial is sufficient to center the oscillations around zero.The idea behind this polynomial subtraction is that the envelope of a sound (seen as attack, decay, sustain and release) can be roughly approximated by a 9th degree polynomial.

This gives us the distance d_{sp}:

$$d_{sp}(a_1, a_2) = d_c(|\widetilde{A_1}|, |\widetilde{A_2}|) \tag{11}$$

where $\widetilde{A_k}$ is the Fourier spectrum of $\widetilde{a_k}$ with

$$\widetilde{a_k} = a_k - \Pi(a_k)$$

where a_1 and a_2 are the amplitudes of two partials, $\Pi(x)$ is the envelope polynomial computed from signal x using a simple least-squares method.

5. EVALUATION

In this section, we present the methodology used for evaluating the performance of the different metrics reviewed in Section 3 and proposed in Section 4. The evaluation database is first described. Next, several criteria are presented, each one evaluating a specific property of the evaluated metric.

5.1. Database

In this study, we focus on a subset of musical instruments that produce pseudo-periodic sounds and model them as a sum of partials (see Section 2). The instruments of the IOWA database [21] globally fit to this condition even though some samples have to be removed.

The evaluation database is created as follows. Each file of the IOWA database is split into a series of audio files, each containing only one tone. The partials are then extracted for each tone using common partials tracking algorithms [7, 8, 22]. Since we consider only the prominent partials of a given tone, only the extracted partials lasting for at least 1 second are retained.

5.2. Criteria

Once the evaluation database is defined, one need some criteria to evaluate the capability of a given metric to determine that two partials are “close” if they actually belong to the same acoustical entity and “far” otherwise.

5.2.1. Fisher criterion

A relevant dissimilarity metric between two partials is a metric which is low for partials of the same entity – the class from the statistical point of view – and high for partials that do not belong to the same entity. The intra-class dissimilarity should then be minimal and the inter-class dissimilarity as high as possible. The Fisher criterion $\mathcal{F}(U)$ described in [16] is loosely based on the fisher discriminant commonly used in statistical analysis to reflect this property. It provides a first evaluation of the discrimination quality of a given metric. It can however be noticed that this criterion is dependent of the scale of the studied dissimilarity metric.

5.2.2. Density criterion

Dissimilarity-vector based classification involves calculating a dissimilarity metric between pair-wise combinations of elements and grouping together those for which the dissimilarity metric is small according to a given classification algorithm.

The density criterion $\mathcal{D}$ intends to evaluate a property of the tested metric that should be fulfilled in order to be relevantly used in combination with common classification algorithms such as hierarchical clustering or K-means. Indeed, many classification algorithms iteratively cluster partials which relative distance is the smallest one. The density criterion, mathematically described in [16] verifies that these two partials actually belong to the same acoustical entity.

5.2.3. Classification criterion

For this criterion, the quality the tested metric is evaluated by considering the quality of a classification done using the tested metric and a classification algorithm.

We consider an agglomerative hierarchical clustering procedure [23]. This algorithm produces a series of partitions of the partials: $(G_n, G_{n-1}, \ldots, G_1)$.

The first partition G_n consists of n singletons and the last partition G_1 consists of a single class containing all the partials. At each stage, the method joins together the two cluster of partials which are most similar according to the chosen dissimilarity metric.

Here, for the classification criterion, the acoustical entities are identified by simply cutting the dendrogram at the highest levels to achieve the desired number of entities. If the desired number of entities is 2, only the highest level is cut.

The classification criterion $\mathcal{H}$ is then defined as the number of partials correctly classified versus the number of partials classified:

$$\mathcal{H}(X) = \frac{1}{\# X} \# \left\{ a | a \in \hat{E}_n \wedge E(a) = i \right\} \qquad (12)$$

where $\hat{E}_n$ is an acoustical entity extracted from the hierarchy.

5.3. Methodology

To compare the metrics proposed in Section 4 and those reviewed in Section 3, we use the following methodology to compute the three evaluation criteria. First, a number of acoustical entities is randomly selected in the database. Then, for each couple of entities between this selection, the following procedure is operated.

For the two entities of the considered couple (E_i, E_j), we compute t_s and t_e, the median values of the starting/ending time index of the partials. Only the partials existing before $t_s + \epsilon_s$ and after $t_e - \epsilon_e$ are kept.The values ϵ_s and ϵ_e are arbitrarily small constants.

Then, the partials of the two entities are gathered to obtain the tested sinusoidal representation of the mixture $S = E_i + E_j$. Only the common part defined as the time interval where all the partials are active is considered to evaluate the tested metric.

6. RESULTS

Each distances reviewed in Section 3 and proposed in Section 4 are now compared using the evaluation methodology described in the last section. The correlation distance d_c of Equation 4 and the distance d_v proposed by Virtanen (see Equation 6) requires no parameterization.

The distance based on AR modeling d_{ar} considers AR vectors of 4 coefficients computed with the Burg method.The distance d_s of Equation 10 considers spectra computed with the Fast Fourier Transform (FFT) using vectors windowed by the periodic Hann window. The computation of the distance d_{sp} (see Equation 11) is similar except that a 9^{th} order polynomial is first estimated and removed before the FFT computation.

300 acoustical entities were considered for all the experiments detailed in the remainder of this section. The results are presented as mean values for criterion, and the bracketed values are the standard deviations (not shown for $\mathcal{F}$ since the value is already normalized).

6.1. Frequency Parameter

The distances between partials based on the frequency parameter is showed on Table 4(a). The d_s distance we proposed gives the best results for the three criteria. It should be noted that the correlation distance (d_c) gives also good results for the two last criteria. We can also see that removing the polynomial from the frequencies of the partials does not contribute to the quality of the metric since frequencies of the partials of the sounds in the IOWA database are quasi-stationary. The performance is even worse because of the modulations that the polynomial might take away from the frequency evolutions.

	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{H}$
d_c	2.909	0.938 (0.216)	0.929 (0.137)
d_v	1.763	0.929 (0.230)	0.881 (0.172)
d_{ar}	1.863	0.712 (0.326)	0.757 (0.166)
d_s	**3.488**	**0.944** (0.210)	**0.940** (0.130)
d_{sp}	2.909	0.936 (0.219)	0.931 (0.133)

(a) Frequencies

	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{H}$
d_c	1.304	**0.818** (0.300)	0.786 (0.162)
d_v	1.298	0.784 (0.316)	0.773 (0.159)
d_{ar}	**1.938**	0.664 (0.331)	0.733 (0.156)
d_s	1.452	0.778 (0.301)	0.781 (0.163)
d_{sp}	1.366	0.796 (0.297)	**0.803** (0.171)

(b) Amplitudes

Fig. 4. *Three criteria (Fisher, density, hierarchical classification) results for distances presented in this paper, applied on (a) the frequencies of the partials, (b) the amplitudes of the partials. The density and hierarchical criteria (two last columns) are presented as scores between 0 and 1, 1 being a perfect result.*

6.2. Amplitude Parameter

As presented on Table 4(b), the performance of the distance measures for the amplitude parameter are globally worse than those obtained for the frequency parameter, lowering from 94% to 80% correct classifications at best. However, the polynomial removal slightly enhances the results.

The metric d_c performs best for the density criterion since it is generally very low for very similar partials. The metric d_{ar} gives a good result for the Fischer criterion while it performs badly for the two other criteria. This metric was tested in another work [16], but only on a very limited database. On a larger database such as one the one of the IOWA, we can see that this metric does not seem very stable on the three criteria. In this mater, the spectral metrics d_s and d_{sp} perform best.

7. CONCLUSION AND DISCUSSION

In this article, we have proposed a new metric that allows to gather partials of different acoustical entities by considering the evolutions of their frequency and amplitude parameters.

Considering the correlation of the spectrum of these evolutions lead to more reliable results than the ones obtained with the AR modelling approach proposed in previous works [16]. According to the experiments, the modulations of the frequency appear to be the most relevant cue. However, the modulations of the amplitude can also be considered as relevant especially when the amplitude envelope of the partial is removed.

This new metric may be used for the classification of partials into acoustical entities. It has to be noted that the hierarchical classification used as a quality criterion in our study, even though very naive, yields to very good results, about ninety five percent of correct classifications. The use of more sophisticated classification methods will certainly lead to better performance. It would also be of interest to cope with the problem of contaminated partials when dealing with the more realistic case of acoustical entities mixed in the time domain.

8. ACKNOWLEDGMENTS

This work has been partly funded by the Quaero project within the task 6.4: "Music Search by Similarity" and the French GIP ANR DESAM under contract ANR-06-JCJC-0027-01.

9. REFERENCES

[1] Dan Ellis and David Rosenthal, "Mid-level representations for Computational Auditory Scene Analysis," in *IJCAI - Workshop on Computational Auditory Scene Analysis*, August 1995.

[2] Juan P. Bello and Jeremy Pickens, "A Robust Mid-level Representation for Harmonic Content in Music Signals," in *ISMIR*, October 2005.

[3] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," *IEEE TASLP*, vol. 16, no. 2, pp. 278–290, 2008.

[4] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification with Application to Musical Instrument Classification," *IEEE TASLP*, 2009.

[5] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, 1990.

[6] Stephen McAdams, "Segregation of Concurrrents Sounds : Effects of Frequency Modulation Coherence," *JAES*, vol. 86, no. 6, pp. 2148–2159, 1989.

[7] Robert J. McAulay and Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE TASSP*, vol. 34, no. 4, pp. 744–754, 1986.

[8] Xavier Serra, *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[9] M. Lagrange, S. Marchand, and J.B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE TASLP*, 2007.

[10] Stephen Grossberg, *Pitch Based Streaming in Auditory Perception*, Cambridge MA, Mit Press, 1996.

[11] Paulo Fernandez and Javier Casajus-Quiros, "Multi-Pitch Estimation for Polyphonic Musical Signals," in *IEEE ICASSP*, April 1998, pp. 3565–3568.

[12] Anssi Klapuri, "Separation of Harmonic Sounds Using Linear Models for the Overtone Series," in *IEEE ICASSP*, 2002.

[13] Tuomas Virtanen and Anssi Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in *IEEE ICASSP*, April 2000, vol. 2, pp. 765–768.

[14] Julie Rosier and Yves Grenier, "Unsupervised Classification Techniques for Multipitch Estimation," in *116th Convention of the AES*, May 2004.

[15] Martin Cooke, *Modelling Auditory Processing and Organization*, Cambridge University Press, New York, 1993.

[16] Mathieu Lagrange, "A New Dissimilarity Metric For The Clustering Of Partials Using The Common Variation Cue," in *Proc. ICMC*, Barcelona, Spain, September 2005, ICMA.

[17] Fumitada Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE TASSP*, vol. 23, no. 1, pp. 67–72, 1975.

[18] M. Mellody and G. Wakefield, "The time-frequency characteristic of violin vibrato: modal distribution analysis and synthesis," *JASA*, vol. 107, pp. 598–611, 2000.

[19] Sylvain Marchand and Martin Raspaud, "Enhanced Time-Stretching Using Order-2 Sinusoidal Modeling," in *Proc. DAFx*, Naples, Italy, October 2004, pp. 76–82.

[20] Martin Raspaud, Sylvain Marchand, and Laurent Girin, "A Generalized Polynomial and Sinusoidal Model for Partial Tracking and Time Stretching," in *Proc. DAFx*, Madrid, Spain, September 2005, pp. 24–29.

[21] "The IOWA Music Instrument Samples," Online. URL: `http://theremin.music.uiowa.edu`.

[22] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Using Linear Prediction to Enhance the Tracking of Partials," in *IEEE ICASSP*, May 2004, vol. 4, pp. 241–244.

[23] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika*, , no. 2, pp. 241–254, 1967.