A NEW DISSIMILARITY METRIC FOR THE CLUSTERING OF PARTIALS USING THE COMMON VARIATION CUE

Mathieu Lagrange
SCRIME – LaBRI, Université Bordeaux 1
351, cours de la Libération,
F-33405 Talence cedex, France
lagrange@labri.fr

ABSTRACT

In order to be able to operate relevant musical transformations of a sound, a knowledge of the musical structure of this sound is often mandatory. Therefore, we consider a structured representation of sound composed of acoustical entities on top of the widely used sinusoidal model. From the analysis point of view, these entities are clusters of partials which are perceived as a unique complex sound. To be able to automatically identify these clusters, numerous cues are proposed in the literature. We study in this article a generic one, the common variation of the parameters of the partials. Thanks to an original dissimilarity metric based on the autoregressive modeling of the vectors of frequency or amplitude of the partials, we are able to use even micro-modulations of these parameters for our purpose.

1. INTRODUCTION

Many musical transformations can be processed with a "short-term" sinusoidal model. The audio signal is divided in overlapping frames where some sinusoidal components (often called *peaks*) are identified. The parameters of these components are then modified to achieve meaningful musical transformations. The automatic processing of these transformations should be guided by a knowledge of the musical structure of the sound. Yet, this knowledge can hardly be deduced from a "short-term" representation of the sound.

Alternatively, we consider a structured representation of the sound. We first identify continuities between peaks of successive frames. Peaks are linked from frame to frame to build partials: quasi-sinusoidal oscillators with parameters evolving slowly and continuously with time. A new partial tracking algorithm is used [1] to enhance the identification of the onsets / offsets of partials as well as the variation of the parameters of partials even in polyphonic recordings.

The next step is to identify similarities between partials to identify which ones belong to the same acoustical entity. From the perceptual point of view, 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 [2].

The common onset is an important cue that can be robustly handled using a long-term sinusoidal model since the onset of partials are explicitly modelled. This issue is left for further discussion. In this article, we will consider that the considered partials start together. The spatial location is generally not handled by common sinusoidal model although it may be an interesting issue.

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 [3, 4] and others consider only the harmonic relation of the frequencies of the partials [5, 6, 7]. The main advantage of this cue is to rely on a short-term sinusoidal model. Yet, many musical instruments are not harmonic.

In contrast, the cue that consider the correlated evolutions of the parameters of the partials is generic. Numerous psycho acoustical studies showed that the variations or the micro-modulations are important for perception. Bregman writes: "Small fluctuations in frequency occur naturally in the human voice and in musical instruments. The fluctuations are not often very large, ranging from less than 1 percent for a clarinet tone to about 1 percent for a voice trying to hold a steady pitch, with larger excursions of as much than as 20 percent for the vibrato of the singer. Even the smaller amounts of frequency fluctuation can have potent effects on the perceptual grouping of the components harmonics." According to the work of McAdams [8], a group of partials is perceived as a unique acoustical entity only if these variations are correlated.

The clustering method proposed in this article therefore relies on the definition of a dissimilarity metric to evaluate how "far" are two partials given the correlation of their evolutions int the time / frequency plane or the time / amplitude plane. Some dissimilarity metrics proposed in the literature [9, 6] are reviewed in Section 2. We then introduce a new metric based on the autoregressive (AR) modeling of the evolutions of the parameters of the partials. This new metric is compared in Section 4 to those described in Section 2. The proposed metric is then used to cluster partials thanks to a well-known clustering method presented in Section 5: the Agglomerative Hierarchical Clustering (AHC) [10]. The experiments presented in Section 6 show that the proposed metric allows us to consider strong variations such as vibrato or tremolo but also micro-modulations to cluster partials of the same entity.

2. REVIEW OF EXISTING DISSIMILARITY METRICS

To compare the metrics that will be described in the remainder of this article, we consider a set of partials extracted from various musical sounds: a tone of saxophone with vibrato (entity C_1), a modulated singing voice (entity C_2), a piano tone (entity C_3) and a triangle tone (entity C_4). For each sound, the five partials with the highest amplitude are selected. Five other partials, erroneously extracted from a white noise signal are added to the evaluation set (entity C_0). All the partials are truncated to be of the same duration ($\approx$ 1 second). This testing set is plotted on Figure 1 where the partials are sub-indexed by the number of the entity it belong. Except for the metric d_v proposed by Virtanen in [6] where the amplitudes and frequencies parameters of partials are used as-is, the mean is subtracted before any computation and the amplitudes are normalized.

During the experiments, it appears that the correlated evolutions of the frequency parameter was the most relevant cue. It will then be used for the comparison of the different metrics although the correlated evolutions of the amplitudes can also be considered for the clustering of partials as it will be shown in Section 6.

Some widely known metrics are now described. The euclidean distance d_e between two vectors is defined as:

$$d_e(F_1,F_2) = \sqrt{\sum_{i=1}^{N} (F_1(i) - F_2(i))^2} \qquad (1)$$

where F_1 and F_2 are frequency vectors of size N.

Let us consider a harmonic tone modulated by a vibrato of given depth and rate. All the harmonics are modulated at the same rate but their respective depth depends on their harmonic rank. It is then important to consider a dissimilarity metric which is scale-invariant. The euclidean distance is therefore not suitable for our purpose. Alternatively, Cooke uses a distance [9] equivalent to the cosine dissimilarity d_c defined as:

$$\begin{aligned} d_c(F_1,F_2) &= 1 - \frac{c(F_1,F_2)}{\sqrt{c(F_1,F_1)}\sqrt{c(F_2,F_2)}} && (2) \\ c(F_1,F_2) &= \sum_{i=1}^{N} F_1(i)F_2(i) && (3) \end{aligned}$$

where F_1 and F_2 are frequency vectors of size N. Thanks to the normalization, this dissimilarity is scale-invariant.

Virtanen proposed in [6] to use the mean-squared error between the frequency vectors first scaled by their average values:

$$d_v(F_1,F_2) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{F_1(i)}{\overline{F_1}} - \frac{F_2(i)}{\overline{F_2}} \right)^2 \qquad (4)$$

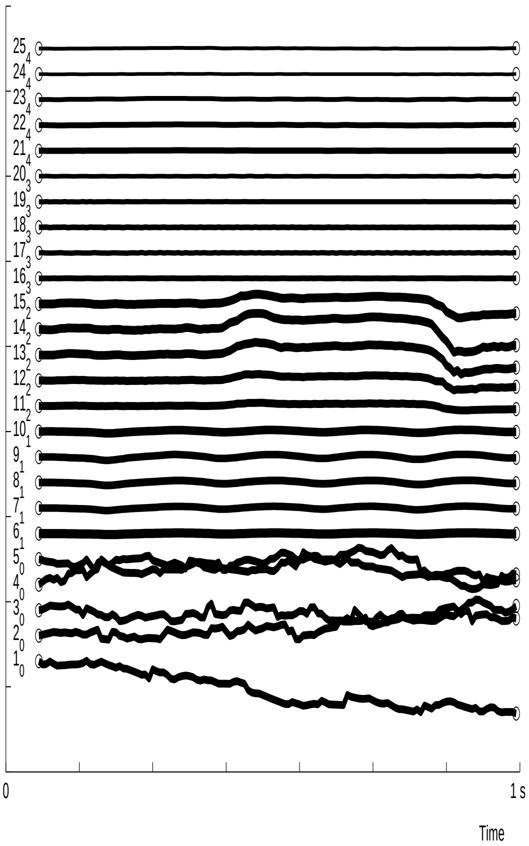


Figure 1. Frequency evolution of five partials extracted from a white noise signal (class C_0), five partials of a saxophone tone with vibrato (entity C_1), five partials of a singing voice (entity C_2), five partials of a piano tone (entity C_3) and five another from a triangle tone (entity C_4). These frequency evolutions are arbitrarily distributed over the ordinate and are indexed by growing index and sub-indexed by number of entity.

where F_1 and F_2 are frequency vectors of size N and $\overline{X}$ denotes the mean of X.

If one considers the evolution in time of the frequencies of a partial as a signal, one can consider a decomposition of this signal in two parts. One part evolves slowly and continuously with time and therefore comply to the requirements of the sinusoidal model. The other part belongs to observation noise due to estimation error of the frequencies of the partial or background noise.

Only the first part should be considered to identify similarities between the parameters of the partials. As a consequence, a relevant dissimilarity for our purposes should be scale-invariant only for the part of the signal that comply to the sinusoidal model. This issue will be studied in the next section to propose an improved dissimilarity metric.

3. PROPOSED DISSIMILARITY METRIC

Let F_l be the frequency vector of the partial l. According to the Auto Regressive (AR) model [11], the sample $F_l(n)$ can be approximated as a linear composition of past samples:

$$F_l(n) = \sum_{i=1}^{k} K_l(i) F_l(n-i) + E_l(n) \qquad (5)$$

where $E_l(n)$ is the prediction error. The coefficients $K_l(i)$ model the predictable part of the signal and it can be shown that these coefficients are scale invariant. On contrary, the non-predictable part $E_l(n)$ is not scale invariant.

We have shown in [12] that AR-modeling of the frequency and amplitudes parameters is relevant to improve the tracking of partials. In this article, we show that the AR modeling is a good candidate for the design of a robust dissimilarity metric.

For each frequency vector F_l, we compute a vector K_l of 4 AR coefficients with the Burg method [13, 14]. Since the direct comparison of the AR coefficients computed from the two vectors F_1 and F_2 is generally not relevant, the spectrum of these coefficients may be compared. The Itakura distortion measure [15], issued from the speech regognition community can therefore be considered:

$$d_{\mathrm{AR}}(F_1, F_2) = \log \int_{-\pi}^{\pi} \frac{|K_1(\omega)|}{|K_2(\omega)|} \frac{d\omega}{2\pi} \qquad (6)$$

where

$$K_l(\omega) = 1 + \sum_{i=1}^{k} K_l(i) e^{-ji\omega} \qquad (7)$$

An other approach may be considered. The amount of error done by modeling the vector F_1 by the coefficients computed from vector F_2 indicate the proximity of these two vectors. Let us introduce a new notation E_1^2, the crossed prediction error defined as the residual signal of the filtering of the vector F_1 with K_2:

$$E_1^2(n) = F_1(n) - \sum_{i=1}^{k} K_2(i) F_1(n-i) \qquad (8)$$

The principle of the dissimilarity d_σ is to combine the two anti-symmetrical dissimilarities $|E_1^2|$ and $|E_2^1|$ to obtain a symmetrical one:

$$d_\sigma(F_1, F_2) = \frac{1}{2}\left(|E_1^2| + |E_2^1|\right) \qquad (9)$$

Given two vectors F_1 and F_2 to be compared, the coefficients K_1 and K_2 are computed to minimize the power of the respective prediction errors E_1 and E_2. If the two vectors F_1 and F_2 are similar, the power of the crossed predictions errors E_1^2 and E_2^1 will be as weak as those of E_1 and E_2.We can consider an other dissimilarity d'_σ defined as the ratio between the sum of the crossed prediction errors and the sum of the direct prediction errors:

$$d'_\sigma(F_1, F_2) = \frac{|E_1^2| + |E_2^1|}{1 + |E_1| + |E_2|} \qquad (10)$$

These dissimilarity metrics based on AR modeling will now be compared to the ones presented in Section 2 using two criteria described in the next section.

4. COMPARISON OF DISSIMILARITY METRICS

A relevant dissimilarity between two elements (the partials) is a dissimilarity which is low for elements of the same class (acoustical entity) and high for elements that do not belong to the same class. The intra-class dissimilarity should then be minimal and the inter-class dissimilarity as high as possible. Let U be the set of elements of cardinal $\#\,U$ and $\mathcal{C}_i$ the class of index i between N_c different classes. An estimation of the relevance of a given dissimilarity $d(x,y)$ for a given class is:

$$\begin{aligned} \mathrm{intra}(\mathcal{C}_i) &= \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} d(\mathcal{C}_i(j), \mathcal{C}_i(k)) && (11) \\ \mathrm{inter}(\mathcal{C}_i) &= \sum_{j=1}^{n_i} \sum_{l=1}^{\#\,U-n_i} d(\mathcal{C}_i(j), \mathcal{F}_i(l)) && (12) \\ Q_d(\mathcal{C}_i) &= \frac{\mathrm{inter}(\mathcal{C}_i)}{\mathrm{intra}(\mathcal{C}_i)} && (13) \end{aligned}$$

where n_i is the number of elements of $\mathcal{C}_i$ and $\mathcal{F}_i = U - \mathcal{C}_i$. The overall quality Q_d is then defined as:

$$Q_d(U) = \frac{\sum_{i=1}^{N_c} \mathrm{inter}(\mathcal{C}_i)}{N_c \sum_{i=1}^{N_c} \mathrm{intra}(\mathcal{C}_i)} \qquad (14)$$

For example, let $\mathcal{A}_1 = \{\{1.1, 5.1\}, \{1.0, 5.2\}, \{1.0, 5.3\}\}$ and $\mathcal{A}_2 = \{\{1.0, 1.1\}, \{1.1, 0.9\}\}$ be two classes of points $e_i = \{x_i, y_i\}$ in a two dimensional space. One dissimilarity considers the abscissa $d_x(e_i, e_j) = |x_i - x_j|$ and the other considers the ordinate $d_y(e_i, e_j) = |y_i - y_j|$. If we study the data, the most relevant dissimilarity is d_y which is verified by the quality measure Q_d : $Q_{d_x}(U) = 0.75 < Q_{d_y}(U) = 32$.

The criterion defined in Equation 13 is first used to evaluate the capability of the metrics proposed in the last two sections to discriminate partials of a given class from the others. Next, the criterion defined in Equation 14 is used to globally evaluate this criterion for each metric. The results, summarized in Table 1 will be further detailed in the remaining of the section. It can however be noticed that this criterion is highly dependant of the scale of the studied dissimilarity metric.

We then also consider an other criterion, noted ζ which is more independent of the scale the evaluated dissimilarity metric than the previous one. Given a set of elements X, $\zeta(X)$ is defined as the ratio of couples (a,b) so that b is the closest element to a and a and b belong to the same class.

Given a function named “cl” defined as:

$$\begin{array}{llll} \mathrm{cl}: & X & \rightarrow & \mathbb{N} \\ & a & \mapsto & i \end{array}$$

where i is the index of the class of a. We get:

$$\zeta(X) = \frac{\#\,\{(a,b)\,|\,d(a,b) = \min_{c\in X} d(a,c) \wedge \mathrm{cl}(a) = \mathrm{cl}(b)\}}{\#\,X} \qquad (15)$$

where X can be either a class $\mathcal{C}_i$ or the set of elements U and $\#\,x$ denotes the cardinal of x.

As shown on the first column of Tables 1 and 2, the dissimilarity d_e obtain bad marks for the saxophone tone and

Q_d	d_e	d_c	d_v	d_{AR}	d_σ	d'_σ
$\mathcal{C}_0$	3.9	4.3	0	57.3	8.9	16.8
$\mathcal{C}_1$	10.5	806.6	47634	37.4	97.3	46.9
$\mathcal{C}_2$	3.1	1586.6	37.4	33.6	23.1	29.5
$\mathcal{C}_3$	147.9	4.8	57.1	22.8	251.8	81.3
$\mathcal{C}_4$	49.5	43.9	83866	23.3	72.1	22.6
U	5.5	10.6	5.1	27.8	21.2	36.2

Table 1. Quality estimation according to the criterion Q_d defined in equations 13 for the first five lines and 14 for the last line. The evaluated dissimilarities are: d_e the euclidean distance, d_c the cosine distance, d_{AR} the Itakura distortion measure, d_σ the crossed prediction error dissimilarity and d'_σ the normalized cross prediction error dissimilarity. The frequency vectors used for the experiment are plotted on Figure 1. The metrics based on AR modeling obtain the better overall results.

ζ	d_e	d_c	d_v	d_{AR}	d_σ	d'_σ
$\mathcal{C}_0$	0.6	0.6	0.6	1	0	1
$\mathcal{C}_1$	0	1	1	1	0.8	0.8
$\mathcal{C}_2$	0	1	1	1	0.2	1
$\mathcal{C}_3$	1	0.6	0.4	1	1	1
$\mathcal{C}_4$	0	1	1	1	1	1
U	0.3	0.84	0.8	1	0.6	0.96

Table 2. Quality estimation according to the criterion ζ defined in Equations 15 for dissimilarity d_e the euclidean distance, d_c the cosine distance, d_{AR} the Itakura distortion measure, d_σ the crossed prediction error dissimilarity and d'_σ the normalized cross prediction error dissimilarity. The frequency vectors used for the experiment are plotted on Figure 1. The d_c and d_v metrics obtain comparable results while d_{AR} and d'_σ obtain the best overall results.

the modulated voice because this dissimilarity is not scale-invariant. The dissimilarity d_c gets better results in case of modulations, as shown by the second column of Tables 1 and 2. The dissimilarity d_v shows disparate results. Some entities are easily discriminated like the saxophone tone $\mathcal{C}_1$ and the triangle one $\mathcal{C}_4$. On contrary, the marks are not as satisfying for others entities like the piano one and the “noisy” partials.

The dissimilarity d_{AR} gets very good marks as it can be noticed in the fourth column of Tables 1 and 2. The dissimilarity d_σ offers good performances in case of predictable evolutions of the frequency. On contrary, the correlations between the partials of the voice tone or the noisy partials are not clear, see the fifth column of Table 2. The dissimilarity d'_σ offers more homogeneous results for the classes of partials we want to handle, see the last column of Table 1. This homogeneity is crucial for the clustering method described in the following section.

5. AGGLOMERATIVE HIERARCHICAL CLUSTERING

As stated before, dissimilarity-vector based classification involves calculating a dissimilarity metric between pairwise combinations of elements and grouping together those for which the dissimilarity metric is small according to a given clustering method.

One employed by Cooke [9] selects a seed element from the data set and computes the distance between this seed and all other elements in the data set in order to clusters the elements whose distance from the seed is below a given threshold. If any elements remain in the data set after this search, a new seed is selected from the remaining elements and the grouping procedure is repeated. If no element is found in the data set that is within the threshold of the seed, the seed is considered to belong to a singleton group. The entire process is repeated until no elements remain in the data set.

The method of Virtanen [6] is a slight variation to this approach where the initial seeds are not individual tracks but rather small groups of tracks that have been formed by matching onset times. The distance between the seed group and each of the remaining tracks in the data set is then defined as the average distance between each of the tracks in the seed and the track under consideration. Virtanen adopted this method as a mean for computational complexity reduction, over performing an exhaustive minimisation of the distance between tracks within each group.

These two approaches rely on a good initialization, and would failed if the threshold or the seed are not relevant for the considered data set. Alternatively, we propose to cluster partials by means of the agglomerative hierarchical clustering (AHC) method which requires no initialization.

An agglomerative hierarchical clustering procedure produces a series of partitions of the data: $(P_n, P_{n-1}, \ldots, P_1)$. The first partition P_n consists of n singletons and the last partition P_1 consists of a single class containing all the elements. At each stage, the method joins together the two classes which are most similar according to the chosen dissimilarity metric. At the first stage, of course, this amounts to joining together the two elements that are closest together, since at the initial stage each class has one element.

Hierarchical clustering may be represented by a two dimensional diagram known as “dendrogram” which illustrates the fusions made at each successive stage of clustering, see Figure 2(b) where the length of the vertical bar that links two classes is calculated according to the distance between the two joined classes. The classes are then found by “cutting” the dendrogram at levels where the difference between the distance of this level and those of the previous level is above a given threshold.

Differences between methods arise because of the different ways of defining dissimilarity between classes. For computing efficiency, the elements properties (amplitude of frequency vector of partials) should not be considered to compute the distance between the union of two joined classes $\mathcal{C}_i$ and $\mathcal{C}_j$ and the remaining classes set $\mathcal{C}_k, \forall k \neq (i, j)$.

A first method, known as the minimal linkage method, consists in choosing the smaller distance between $d(\mathcal{C}_i, \mathcal{C}_k)$ and $d(\mathcal{C}_j, \mathcal{C}_k)$. The distance between two classes is then the distance between the two nearest elements of these classes. Another method, proposed by Ward [16], minimises the intra-class inertia during the aggregation process. Given a partition of K classes, the intra-class inertia considers the class homogeneity:

$$I = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(\mathcal{C}_k(i), \overline{\mathcal{C}_k}) \tag{16}$$

where $\mathcal{C}_k$ is a class with n_k elements and $\overline{\mathcal{C}_k}$ its barycenter. The distance between the union of two classes $\mathcal{C}_i$ and $\mathcal{C}_j$ and another class $\mathcal{C}_k$ is computed as follow:

$$\begin{aligned} d(\mathcal{C}_i \cup \mathcal{C}_j, \mathcal{C}_k) = \frac{1}{n_k + n_j + n_i} [(n_k + n_i)\, d(\mathcal{C}_i, \mathcal{C}_k) + \\ (n_k + n_j)\, d(\mathcal{C}_j, \mathcal{C}_k) + (n_i + n_j)\, d(\mathcal{C}_i, \mathcal{C}_j)] \end{aligned} \tag{17}$$

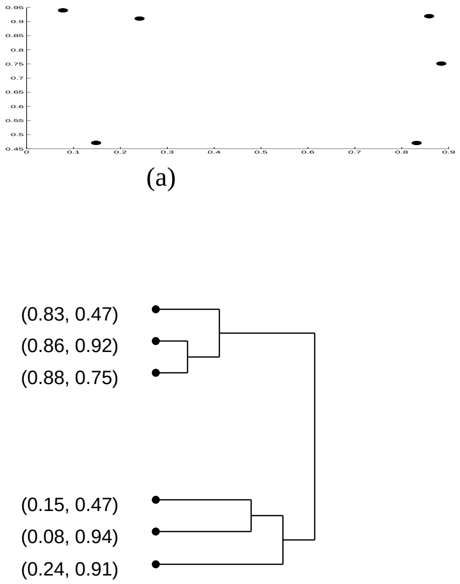
where n_i is the number of elements of the class $\mathcal{C}_i$. This method is designed for the analysis of scalar data vectors and gives good results for our applications. As an example, let the data set be 12 points which attributes are their coordinates, as shown in Figure 3(a). The Figures 3(b) and 3(c) are respectively the dendrograms computed using the minimal link method and the Ward method with the euclidean distance as dissimilarity metric. The first exhibit a "chain" effect not suitable for our purpose. The second leads to a more balanced dendrogram, easing the identification of classes.

6. EXPERIMENTS

Several experiments were conducted to evaluate the relevance of the dissimilarity metrics described in this article if the HAC algorithm is used to cluster partials. Only the hierarchies obtained with the normalized crossed prediction error dissimilarity d'_σ defined in Equation 10 are reported here since it gave the most relevant results. The hierarchies were computed with the Ward method.

We first consider the data set plotted on Figure 1. If we consider the similarity between the frequency vector of the partials, the resulting hierarchy is almost perfect, see Figure 4. The partials of musical tones are correctly clustered and the noisy ones are inserted in the hierarchy at a high level, easing their elimination by the use of the cutting threshold (set to 0.1 here). If we consider the similarity between the amplitude vector of the partials, the hierarchy is not as satisfying, see Figure 5 because partials from entities 1 (saxophone tome with vibrato) 2 (voice tone) and 4 (piano tone) are mixed.

The testing material used above consider a wide range of modulations. We now focus on the micro modulations of the parameters of some piano tones to clusters partials. The partials of three piano tones from the IOWA database with different pitches and similar intensity are used as testing material. Five partials per tone are considered. The hierarchies are computed with the same algorithm.



(a)

(b)

Figure 2. On top are plotted six points on a plane with arbitrary axes to be clustered by the AHC method. At bottom is plotted the dendrogram representing the hierarchy obtained using an euclidean distance as the dissimilarity metric between points. We can clearly distinguish two classes, one composed of points with abscissa close to 0 and the other composed of points with abscissa close to 1.

Even if the resulting hierarchies are not perfect, see Figures 6 and 7, some correlations are clear especially for the tone with the highest pitch (cluster 3). It shows that the d'_σ dissimilarity is able to discriminate between micro-modulations and observations noise even for steady pitch tones of musical instruments like piano.

7. CONCLUSION

We have shown in this article that the long-term sinusoidal model allows us to consider a generic cue for partials clustering: the common variation cue. The autoregressive modeling of the evolutions of the parameters of the partials appears to be relevant for the design of a robust dissimilarity metric exploiting this cue.

The experiments showed that thanks to the dissimilarity proposed in this article, not only the large modulations such as vibrato or tremolo but also the micro-modulations are relevant for clustering partials into acoustical entities. The analysis of these modulations may be of interest for the description of acoustical entities with applications to instruments recognition. This topic should be explored in a near future.

8. REFERENCES

[1] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Improving the Tracking of Partials

for the Sinusoidal Modeling of Polyphonic Sounds," in *IEEE ICASSP*, March 2005, vol. 4, pp. 241–244.

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

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

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

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

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

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

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

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

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

[11] Steven M. Kay, *Modern Spectral Estimation*, chapter Autoregressive Spectral Estimation : Methods, pp. 228–231, Signal Processing Series. Prentice Hall, 1988.

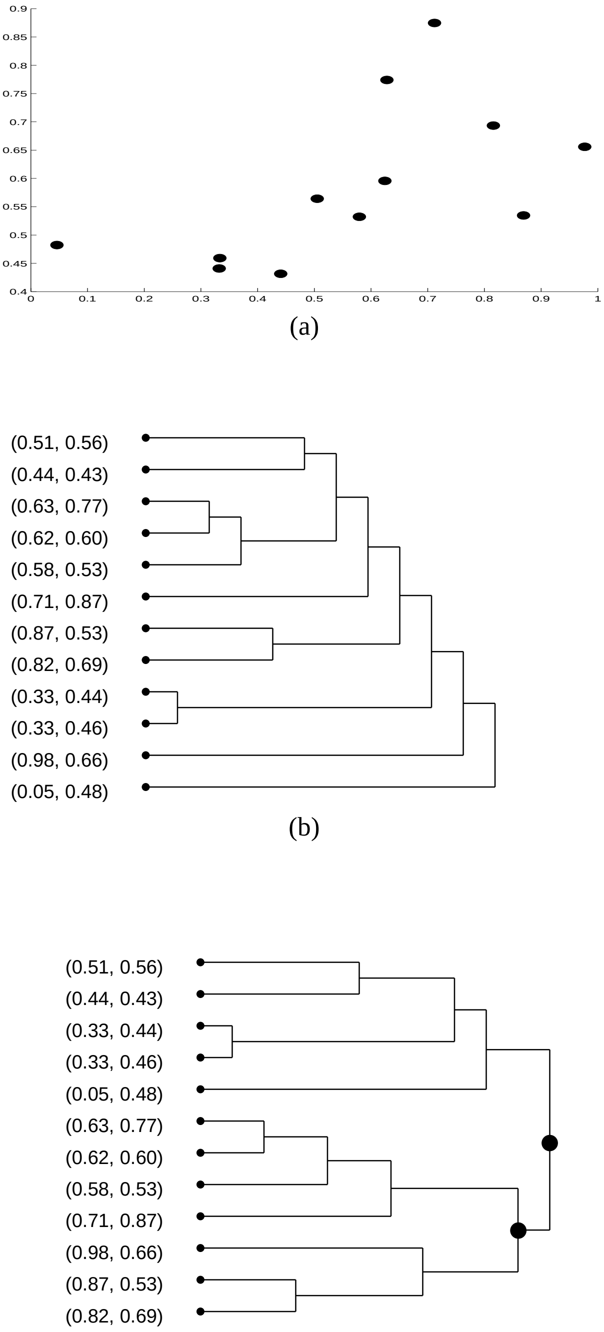
[12] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault, "Enhanced Partial Tracking Using Linear Prediction," in *Proc. DAFx*. Queen Mary, University of London, September 2003, pp. 141–146.

[13] John P. Burg, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Stanford University, 1975.

[14] Florian Keiler, Daniel Arfib, and Udo Zölzer, "Efficient Linear Prediction for Digital Audio Effects," in *Proc. DAFx*. Università degli Studi di Verona and COST, December 2000.

[15] Fumitada Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[16] Joe H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, pp. 238 – 244, 1963.



(a)

(b)

(c)

Figure 3. 12 points on a plane with arbitrary axes are clustered by the AHC method. Dendrograms of the hierarchies obtained with either the minimal link method (b) and the Ward's method (c). In both hierarchies, the 2-dimensional euclidean distance is used as a dissimilarity metric between the elements to be classified. The first hierarchy exhibits a "chain" effect not suitable for our purpose. The second one is more balanced, easing the identification of classes.

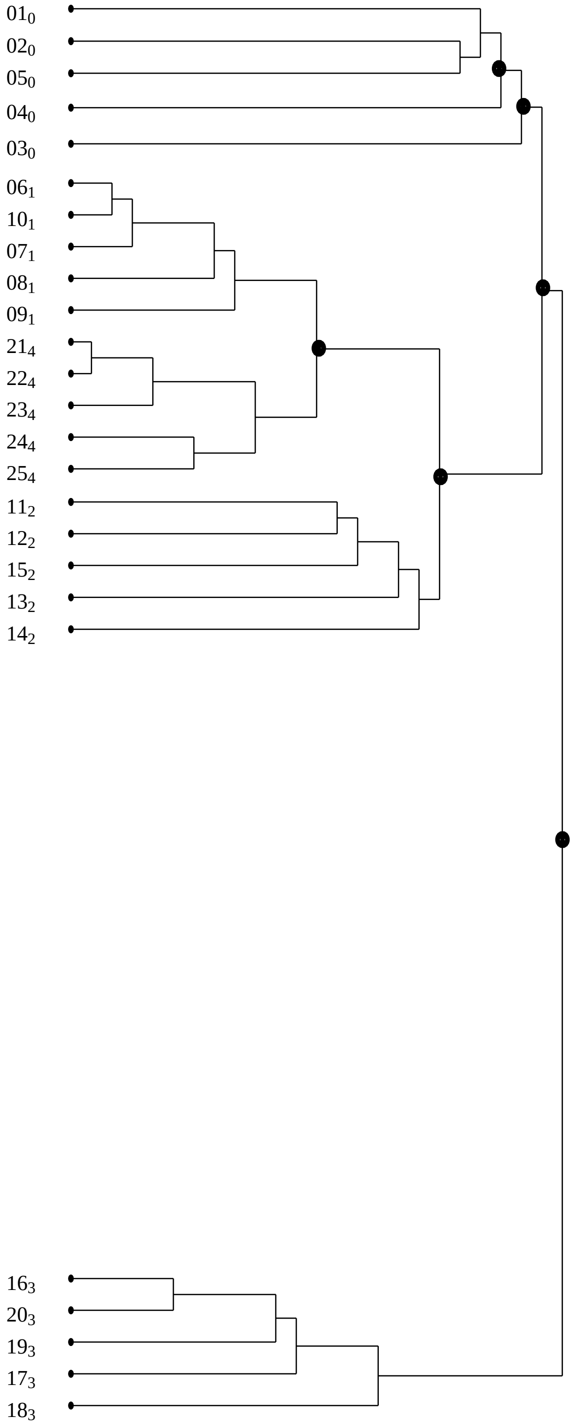


Figure 4. Dendrogram obtained while clustering the partials of the testing set plotted on Figure 1 according to the dissimilarity of their frequency vectors computed with the d'_σ metric. The partials are indexed by growing index and sub-indexed by number of entity. The cuts, represented with dots allows us to identify classes that clusters partials from the same entity. Using the frequency variation cue, all partials are correctly clustered. Additional cuts split the cluster of partials erroneously extracted from noise (class 0). This is not a major disadvantage since these partials does not explicitly belong to an acoustical entity.

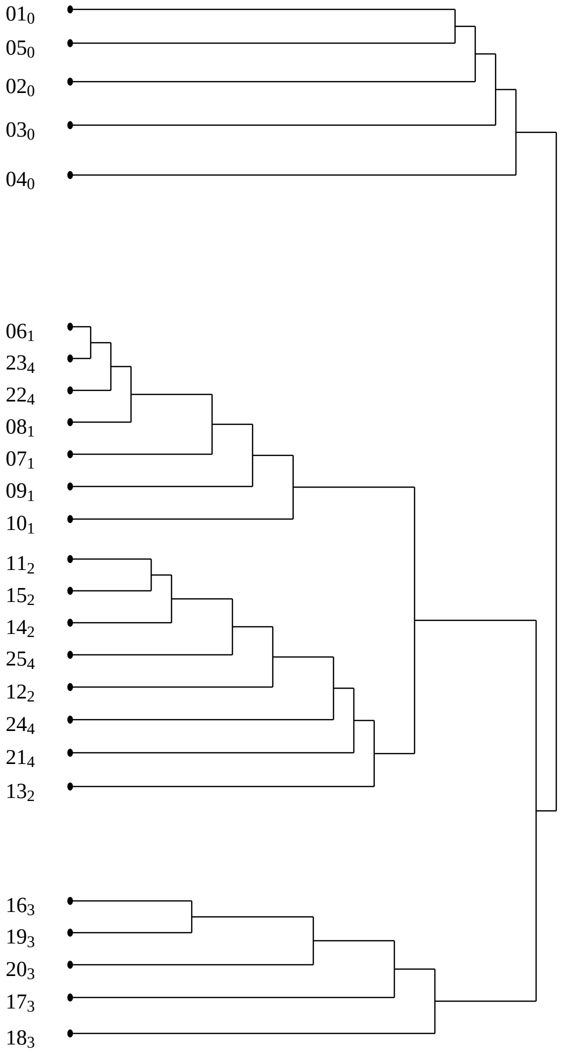


Figure 5. Dendrogram obtained while clustering the partials of the testing set plotted on Figure 1 according to the dissimilarity of their amplitude vectors computed with the d'_σ metric. The hierarchy is not perfect because partials from entities 1 (saxophone tone with vibrato) 2 (voice tone) and 4 (piano tone) are mixed.

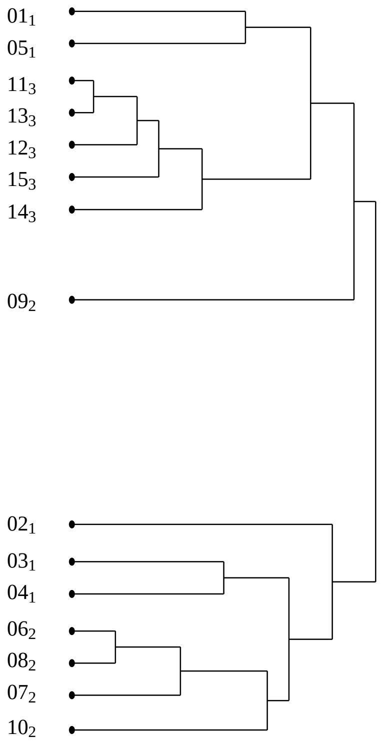


Figure 6. Dendrogram obtained according to the frequency vectors of 15 partials of 3 piano tones with different pitches. The partials are indexed by growing index and sub-indexed by number of entity. The two higher entities (sub-indexed 2 and 3) are well identified in the hierarchy whereas the lower one (sub-indexed 1) is split. Even if the piano has an almost steady pitch, some correlations between the frequency vectors of the same tone can be exploited.

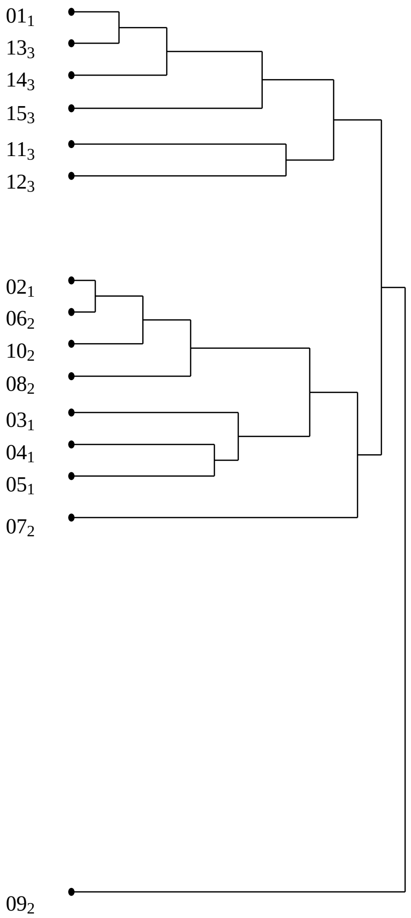


Figure 7. Dendrogram obtained according to the amplitude vector of three piano tones (of five partials each) with different pitches. Only the highest tone (sub-indexed 3) is clearly identified on the hierarchy. The correlation of the amplitude of the partials appears to be a less relevant cue for the clustering of partials than the one that considers the frequencies of partials.