Spectral similarity metrics for sound source formation based on the common variation cue

Mathieu Lagrange, Martin Raspaud

To cite this version:

Mathieu Lagrange, Martin Raspaud. Spectral similarity metrics for sound source formation based on the common variation cue. Multimedia Tools and Applications, 2010, pp.185-205. hal-01132571

HAL Id: hal-01132571
https://hal.science/hal-01132571v1

Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noname manuscript No.
(will be inserted by the editor)

Spectral Similarity Metrics For Sound Source Formation Based on the Common Variation Cue

Mathieu Lagrange · Martin Raspaud

Received: date / Accepted: date

Abstract Scene analysis is a relevant way of gathering information 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. Consequently, 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 its benefits 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 takes better account of micro-modulations.

Keywords auditory scene analysis, mid-level representation, clustering, common variation cue

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 [2]. 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 [13, 5] 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

M. Lagrange
Telecom ParisTech 46, rue Barrault 75634 PARIS Cedex 13 - FRANCE
Tel.: +33 (0)1 45 81 73 24 Fax: +33 (0)1 45 81 71 44
E-mail: lagrange@telecom-paristech.fr

M. Raspaud
Linköping University Bredgatan 33 SE-60174 Norrköping - SWEDEN
Tel.: +46 (0)11-36 34 53 Fax: +46 (0)11-36 32 70
E-mail: Martin.Raspaud@itn.liu.se

process to a given element of the scene [28]. 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 [21].

In order to extract such representation or segmentation, many cues can be considered [6]. Timbre is one of them. The description of the timbre of monophonic sounds has been widely studied [31] and many descriptors have been proposed [18]. These descriptors or *features* are mainly based on the temporal or spectral observations of the sounds since "Timbre depends primarily upon the spectrum of the stimulus, but it also depends on the waveform, the sound pressure, the frequency location, of the spectrum, and the temporal characteristics of the stimulus.", as stated in the ANSI definition of timbre [19]. Unfortunately, most of these descriptors can not be directly extracted from polyphonic recordings.

If the sounds produced by the instruments can be considered as pseudo-periodic, a monophonic or polyphonic signal may be decomposed into sinusoidal components with parameters that evolve slowly with time, the *partials*. This restriction is not too strong since most classical instruments fit in this category, from strings to brass instruments. In this case, several criteria or psychoacoustical 'cues' proposed in the Auditory Scene Analysis (ASA) literature [6] may then be considered for an automatic evaluation of the timbre of each sounds sources [14]. In particular, it is shown in the work of McAdams [32] that the correlated evolution of the parameters of the partials of a given musical or vocal tone is an important cue for the perception of timbre.

Consequently, in order to ensure the relevance of the approach proposed in this paper, the analysed signals have to be pseudo-periodic in order to be suitable for the sinusoidal model that is the front-end of our method. The signals can be inharmonic. In fact, that is the main motivation of the use of the common variation cue to complement the harmonicity one. They should be best monophonic but in case of weak polyphonies, *i.e.* no unison, some partials are not overlapping and can be assigned to only one of the two different sources active at the same time.

The common variation cue has been used for source separation [9, 12, 46] *i.e.* to determine which partials have been produced simultaneously by the same Producing Sound System (PSS) and therefore automatically extract a high level description of polyphonic sound. This cue is also a musical parameter that describes timbre and therefore also have potential for Musical Information Retrieval (MIR) applications such as musical instrument, instrument class identification, and instrumentalist or locutor recognition.

These applications both rely on the definition of a metric to evaluate how dissimilar two partials are, according to the common variation of their parameters. We will show in this paper that considering the spectrum of these variations allows us to propose a robust dissimilarity metric. 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 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 variation 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, where the database and the criteria that evaluate the ability of the tested metric to discriminate partials produced from different PSS. The results of this evaluation are presented in Section 6.

The timbral discrimination capabilities of the proposed metric, *i.e.* its ability to differentiate partials produced by not only different PSS but also different instruments or different classes of intruments are studied in Section 7 and some potential applications are described in Section 8.

2 High-Level Representation of Polyphonic Sounds

Most of the descriptors used in MIR applications consider temporal features such as mean zero-crossing rate or spectral ones such as Mel-Frequency Cepstrum Coefficients (MFCC), see the work of P. Herrera *et al.* [18] for a deeper review. These descriptors are generally extracted on a frame basis and the frames are usually considered independently, loosing most of the temporal information.

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

2.1 Sinusoidal Model

The sinusoidal model represents pseudo-periodic sounds as sums of sinusoids – so-called partials – controlled by parameters that evolve slowly with time [33,43]. More formally put, the audio signal s can be calculated from the controlling parameters using Equations 1 and 2, where N is the number of partials and the functions f_p, a_p, and ϕ_p are the instantaneous frequency, amplitude, and phase of the p-th partial, respectively. The N pairs (f_p, a_p) are the parameters of the additive model and represent points in the frequency-amplitude plane at time t.

$$s(t) = \sum_{p=1}^{N} a_p(t) \ \cos(\phi_p(t)) \tag{1}$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u) \ du \tag{2}$$

This can also be written from the set point of view:

$$P_k(m) = \{F_k(m), A_k(m), \Phi_k(m)\} \tag{3}$$

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.

On a frame basis, the instantaneous frequency, amplitude, and phase of each partials can be estimated using Fourier based approaches like the parabolic methods [1] the phase-based methods [25] and the reassignment one proposed in [3]. In order to go beyond the resolution limitation of the Fourier transform, one can also consider parametric methods like the ESPRIT algorithm [29,4] or maximum likelihood ones, like the matching pursuit [8,10]. Those estimate can be complemented with the estimation of the slope of the frequency and amplitude [1,42] that could be considered at the tracking phase to obtain a more precise modeling of the long term evolution of the frequency and amplitude parameters through time.

The partials can be extracted from the parameters estimated on a frame basis using partial tracking algorithms [33,43,44,27,40,35]. Polyphonic sounds can be considered with dedicated tracking algorithms [11,26]. However, in order to avoid problems due to strong polyphony [13], we only consider in this paper mixtures of entities extracted from monophonic signals.

2.2 Acoustical Entities

These sinusoidal components are called partials because they are only a part of a more perceptively coherent entity that may be called an acoustical entity.

This can be written as:

$$\mathcal{S} = \bigcup_{n=1}^{N} E_n \tag{4}$$

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:

$$E_n = \bigcup_{k=1}^{M_n} P_k^n \tag{5}$$

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

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 [6].

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 [17,15] and others consider only the harmonic relation of the frequencies of the partials [23,46,41]. Yet, many musical instruments are not perfectly harmonic.

In contrast, the cue that consider the correlated evolutions of the parameters of the partials is generic. Also, 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 [32], a group of

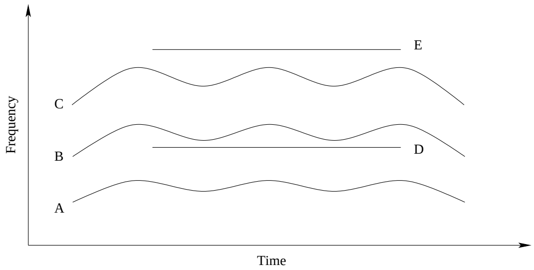


Fig. 1 *Representation of two fictive sounds in the time-frequency domain. Partials A, B, and C (clearly correlated in modulation and starting and ending times, that is common variation) represent the sinusoidal components of the first sound, while D and E represent the sinusoidal components of the second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

partials is perceived as a unique acoustical entity only if these variations 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

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

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.

Cooke uses a distance [9] equivalent to the cosine dissimilarity d_c, also known as *intercorrelation*:

$$d_c(X_1, X_2) = 1 - \frac{c(X_1, X_2)}{\sqrt{c(X_1, X_1)}\sqrt{c(X_2, X_2)}} \tag{6}$$

$$c(X_1, X_2) = \sum_{i=1}^{N} X_1(i)\, X_2(i) \tag{7}$$

where X_1 and X_2 are real vectors of size N. This dissimilarity is scale-invariant.

T. Virtanen *et al.* proposed (in [46]) to use the mean-squared error between the vectors first normalized by their average values:

$$d_v(X_1, X_2) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_1(i)}{\bar{X}_1} - \frac{X_2(i)}{\bar{X}_2} \right)^2 \tag{8}$$

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.

It is proposed in [24] to consider the Auto-Regressive (AR) model as a scale-invariant metric that considers only the predictable part of the evolutions of the parameters:

$$X_l(n) \approx \sum_{i=1}^{n} k_l(i) X_l(n-i) \tag{9}$$

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 [20]:

$$d_{\mathrm{AR}}(X_1, X_2) = \log \int_{-\pi}^{\pi} \frac{|K_1(\omega)|}{|K_2(\omega)|} \frac{d\omega}{2\pi} \tag{10}$$

where

$$K_l(\omega) = 1 + \sum_{i=1}^{n} K_l(i) e^{-ji\omega} \tag{11}$$

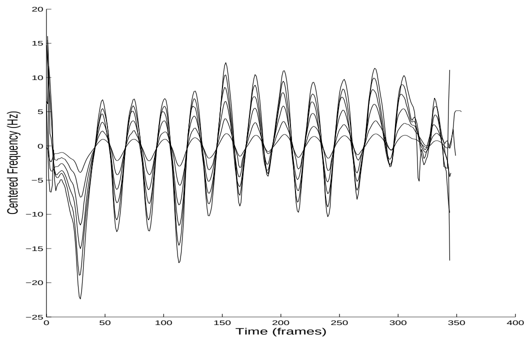
When considering the amplitudes of the partials, a scale-invariant metric is also important. In this context, the normalization proposed by T.Virtanen is no longer motivated since the relative amplitudes of the harmonics depend on the envelope of the sound. For example, on Figure 2(b), the topmost curve (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.

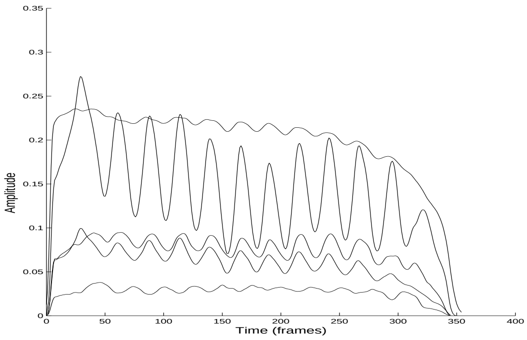
The amplitude evolution of a partial is composed of a temporal envelope and some periodic modulations. Since the envelope of the amplitude of the partials can be very different from partials to partials of the same entity it may be useful to consider only the periodic modulations while computing their similarities. The metric introduced in the next section will cope with these issues.

4 Proposed Metric

We propose 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 10). Since the Fourier transform is based on the fact that the input signal is periodic, using a spectrum of the evolution 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 [30]). It can be also interesting for micro-modulations such as the ones produced 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.



(a) Frequencies



(b) Amplitudes

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

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 6). This gives

$$d_s(f_1, f_2) = d_c(|F_1|, |F_2|) \tag{12}$$

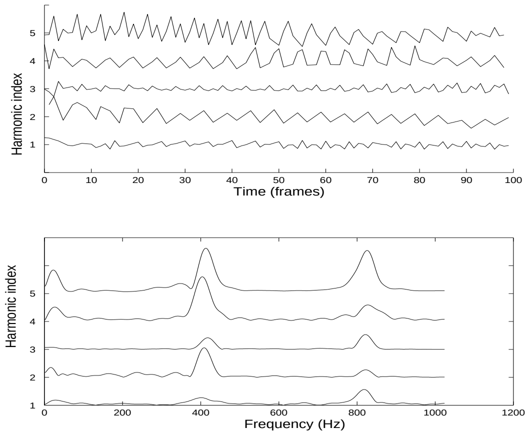


Fig. 3 Centered frequencies (top) of a piano note and their corresponding spectra (bottom). Each curve is shifted for clarity sake.

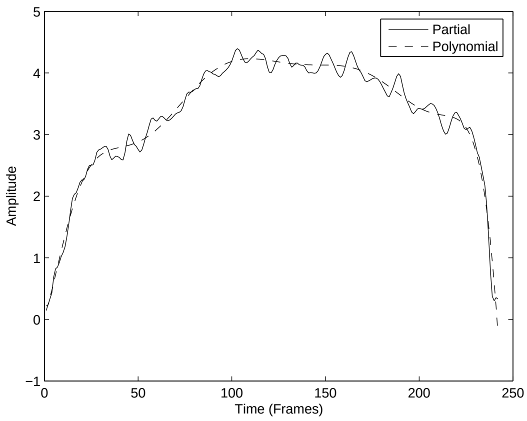
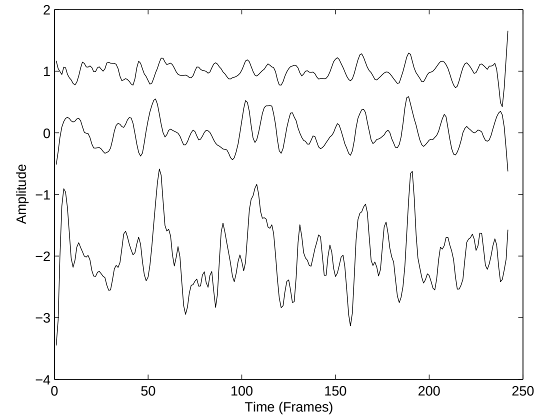


Fig. 4 Amplitudes of a partial of an Bb Clarinet and its polynomial envelope estimation.

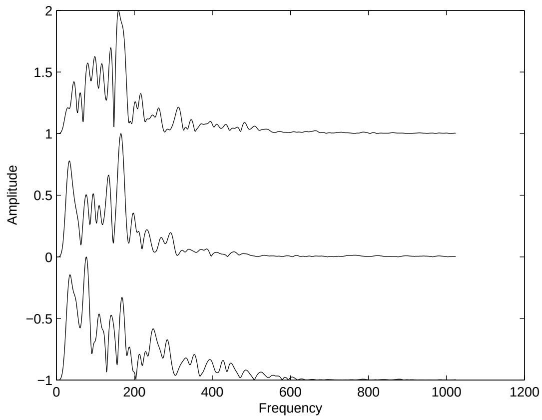
where f_1 and f_2 are the frequency vectors of two partials P_1 and P_2 and F_k is the Fourier spectrum of f_k. Thanks to the absolute value applied to the spectra, this distance is phase-invariant.

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 work [38], subtracting a polynomial is sufficient to center the oscillations around zero, as we see on Figure 4. The idea



(a) Modulations



(b) Corresponding Spectra

Fig. 5 Amplitudes of three partials of an Bb Clarinet when the polynomial envelope is removed (a), and their corresponding spectra (b). The curves have been shifted for clarity sake.

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. An example of such a subtraction is shown on Figure 5.

This gives us the distance d_{sp}:

$$d_{sp}(a_1, a_2) = d_c(|\widetilde{A_1}|, |\widetilde{A_2}|) \tag{13}$$

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 [34].

4.3 Metric Combination

In order to exploit both the frequency and amplitude parameters, we need a way to combine the measures of amplitude and frequency distances.

T.Virtanen *et al.* proposed to combine frequency and amplitude parameters distances by means of adding the two distance measures while considering an harmonicity factor. In their work [46], each distances are weighted before performing the addition. For comparison purposes, we consider the following distance:

$$d_{v+v}(P_1, P_2) = \frac{d_v(f_1, f_2) + d_v(a_1, a_2)}{2} \tag{14}$$

where f_k and a_k are respectively the frequencies and amplitude of partials P_k. Since the weights are not supplied and no harmonicity information is available it is only an approximation of the combination scheme proposed by T. Virtanen.

Since our proposed distances d_s and d_{sp} are normalized, if we want to give the same weight to the two distances, we can combine the frequency and amplitude distances by performing a simple mean. This would then yield :

$$d_{+}(P_1, P_2) = \frac{d_s(f_1, f_2) + d_{sp}(a_1, a_2)}{2} \tag{15}$$

In order to take into account the best result on part of one of the measures, a method would be to take the minimum of the two distances:

$$d_m(P_1, P_2) = \min(d_s(f_1, f_2), d_{sp}(a_1, a_2)) \tag{16}$$

As it will be presented in Section 6, better results are achieved when we multiply amplitude and frequency parameter distances. This combination, however less robust to errors, seems to take better account of the performance of each distance measure independently. In order to keep the metrics in the same scale, a square root is applied to the combination:

$$d_{\times}(P_1, P_2) = \sqrt{d_s(f_1, f_2) d_{sp}(a_1, a_2)} \tag{17}$$

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.

The objective of the evaluation presented in the remaining of the paper is to study if the proposed similarity metrics are good candidates for implementing a clustering of the partials of the same acoustical entity. In Section 7, we extend this study by considering the statistical properties of one of the proposed metric while considering not only the entity level but also larger sets such as all the partials played by a given instrument or a class of instruments.

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 [16] whose instrument hierarchy is plotted in Figure 7, globally fit to this condition even though some samples have to be removed. The "pizzicato" tones, *i.e* plucked-string tones with strong attack and weak resonating phase as well as the "pianissimo" tones *i.e* tones with very low amplitude are discarded.

In order to extract the partials for each tone, each file of the IOWA database is split into a series of audio files, each containing only one tone. The spectral parameters at each frames are estimated using the phase derivative method studied in [25] with the following parameters: the window size is 2048 samples long, the hop size is 512 samples long at a sampling rate of 44100 Hz. An implementation of the algorithm proposed by McAuly and Quatieri in [33] is used with a frequency tolerance of 50 Hz. Since we consider only the prominent partials of a given tone, only the extracted partials lasting for at least 2 seconds are retained. For each entity, only the 20 partials with the highest amplitude are retained.

5.2 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. For the two entities of the considered couple, the median values of the starting/ending time index of the partials t_s and t_e are computed. Only the partials existing before and after $t_s + \epsilon_s$ and $t_e - \epsilon_e$ are kept (see Figure 6). The values ϵ_s and ϵ_e are arbitrarily small constants.

Then, the partials of the two entities are gathered. Only the common part defined as the time interval where all the partials are active is considered to evaluate the tested metric. For example, the common part of the partials represented in Figure 6 is between c_s and c_e.

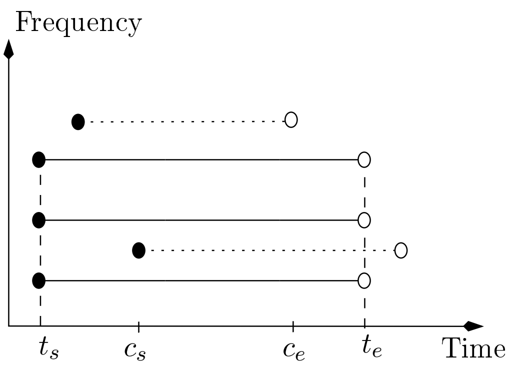


Fig. 6 Selection of the common parts of the partials of the two acoustical entities. A partial start is represented with a black filled dot and its end with a white filled dot. Only the partials existing before and after t_s and t_e are kept, represented with solid lines. The indexes c_s and c_e delimit the common part of all the partials.

5.3 Performance Criteria

Once the evaluation database and the evaluation methodology are defined, some criteria have to be defined that reflect if, by considering the evaluated metric, two partials are "close" if they actually belong to the same acoustical entity and "far" otherwise.

5.3.1 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. Let U be the set of elements of cardinal $\# U$ and C_i the entity of index i between N_c different entities. An estimation of the relevance of a given dissimilarity $d(x, y)$ for a given acoustical entity is:

$$\mathrm{intra}(C_i) = \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} d(C_i(j), C_i(k)) \tag{18}$$

$$\mathrm{inter}(C_i) = \sum_{j=1}^{n_i} \sum_{l=1}^{\# U - n_i} d(C_i(j), \overline{C}_i(l)) \tag{19}$$

$$\mathcal{F}(C_i) = \frac{\mathrm{inter}(C_i)}{\mathrm{intra}(C_i)} \tag{20}$$

where n_i is the number of partials in C_i and $\overline{C}_i = U \backslash C_i$. The overall quality $\mathcal{F}(U)$ is then defined as:

$$\mathcal{F}(U) = \frac{\sum_{i=1}^{N_c} \mathrm{inter}(C_i)}{\sum_{i=1}^{N_c} \mathrm{intra}(C_i)} \tag{21}$$

This last criterion $\mathcal{F}(U)$ is loosely based on the fisher discriminant commonly used in statistical analysis. 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.

5.3.2 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 verifies that these two partials actually belong to the same acoustical entity.

More formally, given a set of elements X, $\zeta(X)$ is defined as the ratio of couples (a, b) so that b is the closest to a and a and b belong to the same acoustical entity.

Given a function named cl defined as:

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

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

$$\mathcal{D}(X) = \frac{\#\ \{(a,b) \,|\, d(a,b) = \min_{c \in X} d(a,c) \wedge \mathrm{cl}(a) = \mathrm{cl}(b)\}}{\#\ X} \tag{22}$$

where X can be either an acoustical entity C_i or the universe U and $\#\ x$ denotes the cardinal of x.

5.3.3 Classification criterion

For this criterion, the quality of 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 (AHC) procedure [22]. This algorithm produces a series of partitions of the partials: $(P_n, P_{n-1}, \ldots, P_1)$.

The first partition P_n consists of n singletons and the last partition P_1 consists of a single class containing all the partials. At each stage, the method joins together the two cluster of partials which are most similar according to the chosen dissimilarity metric. At the first stage, of course, this ends in joining together the two partials that are closest together, since at the initial stage each cluster has only one partial. At each stage, the dissimilarity between the new cluster and the other ones is computed using the method proposed by Ward [47].

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 7 where the length of the vertical bar that links two classes is calculated according to the distance between the two joined clusters.

The acoustical entities can then be found by “cutting” the dendrogram at relevant levels. 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 (see Figure 7).

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

$$\mathcal{H}(X) = \frac{\#\ \{a | a \in \hat{C}_i \wedge cl(a) = i\}}{\#\ X} \tag{23}$$

where $\hat{C}_i$ is an acoustical entity extracted from the hierarchy.

6 Results

Each metrics reviewed in Section 3 and proposed in Section 4 are now compared using the evaluation methodology described in the previous section. The correlation metric d_c of Equation 6 and the metric d_v proposed by T.Virtanen (see Equation 8) requires no parameterization.

The metric d_{ar} considers AR vectors of 4 coefficients computed with the Burg method [7]. The metric d_s of Equation 12 considers spectra computed with the Fast Fourier Transform (FFT) using vectors windowed by the periodic Hann window. The

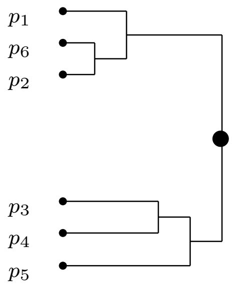


Fig. 7 Dendrogram representing the hierarchy obtained using the AHC algorithm with 6 partials. The cut at the highest level of the hierarchy represented by a dot identify two acoustical entities $C_1 = \{p_1, p_6, p_2\}$ and $C_2 = \{p_3, p_4, p_5\}$.

	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{H}$
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)

Table 1 Three criteria (Fisher, density, hierarchical classification) results for the five metrics presented in this paper, applied on the frequencies of the partials. The density and hierarchical criteria (two last columns) are presented as scores between 0 and 1. For every criteria, a higher value means better performance.

computation of the metric d_{sp} (see Equation 13) is similar except that a 9^{th} order polynomial is first estimated and removed before the FFT computation. The results are presented as mean values for each 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 metrics between partials based on the frequency parameter is showed on Table 1. The d_s metric we proposed gives the best results for the three criteria. It should be noted that the correlation metric (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.

6.2 Amplitude Parameter

As presented on Table 2, the performance of the metrics 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.

	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{H}$
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)

Table 2 Three criteria (Fisher, density, hierarchical classification) results for the five metrics presented in this paper, applied on the amplitudes of the partials. The density and hierarchical criteria (two last columns) are presented as scores between 0 and 1. For every criteria, a higher value means better performance.

	$\mathcal{F}$	$\mathcal{D}$	$\mathcal{H}$
d_{v+v}	1.298	0.784 (0.316)	0.773 (0.159)
d_+	2.040	0.923 (0.230)	0.928 (0.137)
d_m	**3.303**	0.934 (0.216)	0.943 (0.122)
$d_\times$	2.702	**0.937** (0.217)	**0.951** (0.116)

Table 3 Three criteria (Fisher, density, hierarchical classification) results for the four combined metrics we defined. The density and hierarchical criteria (two last columns) are presented as scores between 0 and. For every criteria, a higher value means better performance.

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 [24], 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.

6.3 Combination

In order to jointly take into account the common variation cue of the frequency and amplitude parameters, we considered all possible combinations of preceding metrics (d_c, d_v, d_{ar}, d_s, d_{sp}) for each spectral paramter with the three operators we proposed ($+$, $\times$, min). Only the most relevant ones are presented on Table 3 for clarity sake.

The metric d_m is given best for the Fischer criterion while the metric $d_\times$ shows best results for both density and hierarchical classification criteria (the classification performance is enhanced by 1% over the obtained results with the frequency cue only). Hence the metric $d_\times$ will be kept for timbral discrimination presented in the next Section.

7 Instruments Class discrimination

In the previous section, we used the evaluation database globally in order to compare the different metrics. We study in this section a detailed evaluation of the behavior of the proposed metric by considering several levels in the instruments hierarchy of the IOWA database. Two groups of entities are considered at each experiment to compute the intra-class and inter-class dissimilarities, noted *intra* and *inter* in the remainder of

Instruments		intra(a)			intra(b)			inter(a, b)		
a	b	mean	σ	max	mean	σ	max	mean	σ	min
Ob	Ob	0.018	0.020	0.099	0.018	0.020	0.099	0.101	0.087	0.004
Ob	Sx	0.018	0.021	0.092	0.062	0.072	0.652	0.314	0.225	0.007
Tu	To	0.021	0.033	0.334	0.012	0.015	0.131	0.277	0.152	0.011
BW	WW	**0.015**	0.022	0.295	**0.083**	0.102	0.667	0.315	0.184	0.016
BS	SS	0.127	0.119	0.905	**0.479**	0.3	1.157	**0.5**	0.265	0.012
S	W	0.237	0.216	0.946	0.059	0.11	0.928	0.373	0.204	0.024

Table 4 Evaluation of the discrimination capabilities of the proposed metric for different instruments such as Oboe (Ob), Saxophone (Sx), Trumpet (Tu) and Trombone (To) as well as sets of instruments of the IOWA database such as Brass Winds (BW), Wood Winds (WW), Bowed Strings (BS), and Strucked Strings (SS). The values in the table are respectively the mean, standard deviation and maximal values of the $d_\times$ metric.

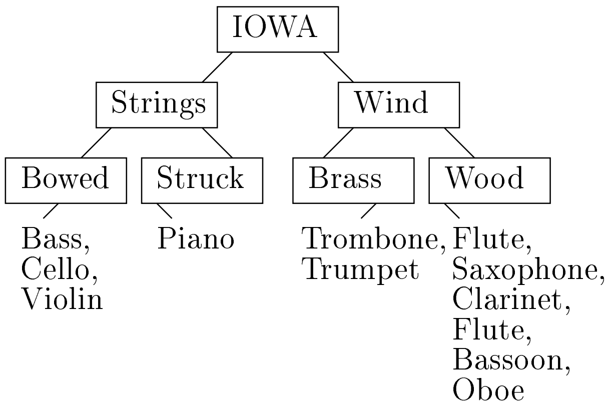


Fig. 8 The IOWA database hierarchy.

this section. Each group corresponds to a node at a given level of the hierarchy showed in Figure 7.

The methodology used for these experiments is the one described in Section 5. For each experiment, we randomly select 100 entities of each considered group and the *intra* and *inter* are computed for each couple of entities, each entity belonging to one group. Only couples with different entities are considered. In order to improve the clarity of the results, the *intra* and *inter* values are not averaged over all couples. Instead, the mean and the standard deviation is computed, as well as the maximum value respectively for the *intra* and the *inter*.

In the first experiment, which results are reported in the first line of Table 4, we consider acoustical entities produced by the Oboe only. Since the same group is considered on both sides, the *intra* values are equal. However, the *inter* is not equal to the *intra* since the computation of the *intra* involves only the partials of one entity, while the computation of the *inter* always involves partials of different entities.

In order to separate perfectly two entities of the Oboe, we would need to have the minimum value of the *inter* greater than the maximum value of the *intra*. It is clearly not the case, since $0.0043 < 0.0996$. However, the average of the *inter* is greater than the maximum value of the *intra*, thus we could achieve good classifications.

Let us now consider two instruments of the Wood Wind family, the Oboe and the Saxophone and two instruments of the Brass Wind family, the Trumpet and the Trombone. Since the set of entities is different from the previous experiment with Oboe only, the *intra* is slightly different. By considering two different instruments, the *inter* is increased to a value that remains almost stable in the higher levels of the hierarchy. It shows that the difference between instruments is the most salient level of the hierarchy, as far as the proposed metric is considered.

Next, the Brass Wind and the Wood Wind family achieve very low *intra*, meaning that partials of the same entity of these two families are dense according to the proposed metric. The fifth line of Table 4 presents the results while considering the Bowed Strings and Struck Strings families, that appear to be very dissimilar. The high *inter* value may be explained by the different types of excitations lead to very different timbre.

The partials of the acoustical entities produced by the Piano (unique instrument of the struck string family in the database) are spread over the feature space. Even though the new metric considers spectral information which does improve the performance over the temporal information in case of micro-modulations, see Figure 3, it appears that the micro-modulations are not as salient as larger modulations such as vibrato or tremolo.

8 Applications

In this section, we describe some applications where such description of the spectro-temporal content of audio streams can be helpful.

8.1 Binaural Scene Analysis

The current paper deals with the common variation of partials. However, two more cues are important for the perceptual gathering of partials: the common direction of arrival, and the harmonicity among partials [6].

The common direction of arrival can be determined in the case of multichannel audio. In the case of binaural sounds (stereo sounds recorded at the entrance of the auditory channels), it is possible to obtain an overall good estimation of the direction of arrival of sound sources. As studied in [37], where it is shown that the direction of arrival of partials, although not a perfect criterion can be used as a partial clustering cue. The harmonicity cue has been used for the gathering of partials too, such as in [46]. By determining the harmonic relationship between partials, it is possible to determine gather the partials by sources of the one hand, and point out the overlapping partials.

These three cues work very differently from each other. Hence, by combining them, we think that we may be able to enhance the robustness and precision of the partial gathering process as the diversity added by the different cues shows interesting perspectives.

8.2 Acoustical Entities Similarity

In this task we are interested in estimating the similarity between two acoustical entities that are whether represented as a segment of audio or its sinusoidal representation.

We are interested in this type of application since there is an increased interest towards recommendation systems that are not based on an ontology such as genre [45] or instrument type [21]. Alternatively, one can consider a recommendation system that states "show me tunes that are similar to the ones I like". In this case, one needs to define the similarity between musical audio signals and the timbre is an interesting dimension to consider.

We are currently investigating a generalized version of the descriptors described in this paper for such a purpose. Preliminar evaluations show that on continuous musical solos, the use of those descriptors combined with standard segmental descriptors like the MFCC's significantly improve the performances.

8.3 Singing Voice Detection

As the proposed descriptors capture the modulations over time of the spectral parameters, they model efficiently the modulations of the singing voice, such as vibrato or tremolo. Assuming that the singing voice is almost always modulated [39], one can consider that the proposed descriptors can be considered to estimate whether a singing voice is active or not. Preliminar experiments show competitive performance compared to state-of-the-art statistical approaches using standard descriptors like the MFCC's [36]. As the proposed descriptors and the MFCC's model different aspects of the audio stream, it is expected that a combination of both approaches will provides a significant improvement.

9 Conclusion

In this article, we have proposed a new metric that discriminate 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 stable results than the one obtained with the AR modeling approach proposed in previous work [24]. According to the experiments, the modulations of the frequency appear to be the most relevant cue, however a slight improvement can be gained concerning the amplitude if the envelope is removed. We also demonstrated that considering the combination of metrics of frequencies and the amplitudes enhanced the classification results as far as the density and hierarchical criteria are concerned.

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 95 percents of correct classifications. Even better performance could certainly be obtained using more sophisticated classification methods, which could be of interest for many MIR applications.

Acknowledgements This work has been initiated when the authors were at the LaBRI (UMR-Cnrs 5800, University of Bordeaux 1) and has been partly funded by the OSEO project Quaero within the task 6.4: "Music Search by Similarity" and the French GIP ANR DESAM under contract ANR-06-JCJC-0027-01.

References

1. M. Abe and I. Smith, J. O. Am/fm rate estimation for time-varying sinusoidal modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 3, pages iii/201–iii/204, 18–23 March 2005.
2. J.-J. Aucouturier and F. Pachet. The influence of polyphony on the dynamical modelling of musical timbre. *Pattern Recognition Letters*, 28(5):654–661, 2007.
3. F. Auger and P. Flandrin. Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method. *IEEE Transactions on Signal Processing*, 43:1068–1089, May 1995.
4. R. Badeau, G. Richard, and B. David. Performance of esprit for estimating mixtures of complex exponentials modulated by polynomials. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 56:492–504, 2008.
5. J. P. Bello and J. Pickens. A Robust Mid-level Representation for Harmonic Content in Music Signals. In *ISMIR*, October 2005.
6. A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.
7. J. P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, 1975.
8. M. G. Christensen and S. H. Jensen. On perceptual distortion minimization and nonlinear least-squares frequency estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):99–109, Jan. 2006.
9. M. Cooke. *Modelling Auditory Processing and Organization*. Cambridge University Press, New York, 1993.
10. L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1808–1816, Sept. 2006.
11. P. Depalle, G. Garcia, and X. Rodet. Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 225–228, April 1993.
12. D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Department. of Electrical Engineering & Computer Science, M.I.T, 1996.
13. D. Ellis and D. Rosenthal. Mid-level representations for Computational Auditory Scene Analysis. In *International Joint Conference on Artificial Intelligence (IJCAI) - Workshop on Computational Auditory Scene Analysis*, August 1995.
14. D. Ellis and B. Vercoe. A perceptual representation of sound for auditory signal separation. In *123rd meeting of the Acoustical Society of America*, May 1992.
15. P. Fernandez and J. Casajus-Quiros. Multi-Pitch Estimation for Polyphonic Musical Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3565–3568, April 1998.
16. L. Fritts. The IOWA Music Instrument Samples. Online. URL: `http://theremin.music.uiowa.edu`, 1997.
17. S. Grossberg. *Pitch Based Streaming in Auditory Perception*. Cambridge MA, Mit Press, 1996.
18. P. Herrera, G. Peeters, and S. Dubnov. Automatic Classification of Musical Sounds. *"Journal of New Musical Research"*, 32(1):3–21, 2003.
19. A. N. S. Institute. USA Standard Acoustical Terminology, 1960.
20. F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):67–72, 1975.
21. C. Joder, S. Essid, and G. Richard. Temporal Integration for Audio Classification with Application to Musical Instrument Classification. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):174–186, 2009.
22. S. C. Johnson. Hierarchical Clustering Schemes. *Psychometrika*, 2(2):241–254, 1967.
23. A. Klapuri. Separation of Harmonic Sounds Using Linear Models for the Overtone Series. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.
24. M. Lagrange. A New Dissimilarity Metric For The Clustering Of Partials Using The Common Variation Cue. In *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, September 2005. International Computer Music Association (ICMA).

25. M. Lagrange and S. Marchand. Estimating the instantaneous frequency of sinusoidal components using phase-based methods. *Journal of the Audio Engineering Society*, 2007.
26. M. Lagrange, S. Marchand, and J. Rault. Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds. *IEEE Transactions on Audio, Speech and Language Processing*, 28:357–366, Aug. 2007.
27. M. Lagrange, S. Marchand, and J.-B. Rault. Using Linear Prediction to Enhance the Tracking of Partials. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 241–244, May 2004.
28. M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized Cuts for Predominant Melodic Source Separation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):278–290, 2008.
29. J. Laroche. The use of the matrix pencil method for the spectrum analysis of musical signals. *The Journal of the Acoustical Society of America*, 94(4):1958–1965, 1993.
30. S. Marchand and M. Raspaud. Enhanced Time-Stretching Using Order-2 Sinusoidal Modeling. In *Proc. DAFx*, pages 76–82. Federico II University of Naple, Italy, October 2004.
31. K. D. Martin and Y. E. Kim. Musical Instrument Recognition: a pattern-recognition approach. In *136th meeting of the Acoustical Society of America*, October 1998.
32. S. McAdams. Segregation of Concurrrents Sounds : Effects of Frequency Modulation Coherence. *Journal of the Audio Engineering Society*, 86(6):2148–2159, 1989.
33. R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.
34. A. Nealen. An as-short-as-possible introduction to the least squares, weighted least squares and moving least squares methods for scattered data approximation and interpolation. URL: `http://www.nealen.com/projects/`, May 2004.
35. L. Nunes, R. Merched, and L. Biscainho. Recursive least-squares estimation of the evolution of partials in sinusoidal analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.
36. M. Ramona and G. Richard. Vocal detection in music with support vector machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.
37. M. Raspaud and G. Evangelista. Binaural partial tracking. In *Proc. DAFx*, pages 123–128, Espoo, Finland, September 2008.
38. M. Raspaud, S. Marchand, and L. Girin. A Generalized Polynomial and Sinusoidal Model for Partial Tracking and Time Stretching. In *Proc. DAFx*, pages 24–29. Universidad PolitÃ©cnica de Madrid, September 2005. ISBN: 84-7402-318-1.
39. L. Régnier and G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.
40. A. Röbel. Adaptive additive modeling with continuous parameter trajectories. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 14(4):1440–1453, 2006.
41. J. Rosier and Y. Grenier. Unsupervised Classification Techniques for Multipitch Estimation. In *116th Convention of the Audio Engineering Society*. Audio Engineering Society (AES), May 2004.
42. A. RÃ¶bel. Frequency-slope estimation and its application to parameter estimation for non-stationary sinusoids. *Computer Music Journal*, 32:68–79, 2008.
43. X. Serra. *Musical Signal Processing with Sinusoids plus Noise*, chapter 3, pages 91–122. Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.
44. A. Sterian and G. H. Wakefield. A Model-Based Approach to Partial Tracking for Musical Transcription. SPIE annual meeting, San Diego, California, 1998.
45. G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 10(5):293–302, 2002.
46. T. Virtanen and A. Klapuri. Separation of Harmonic Sound Sources Using Sinusoidal Modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 765–768, April 2000.
47. J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:238 – 244, 1963.