Explicit Modeling of Temporal Dynamics within Musical Signals for Acoustical Unit Formation and Similarity

Mathieu Lagrange, Martin Raspaud, Roland Badeau, Gaël Richard

Explicit Modeling of Temporal Dynamics within Musical Signals for Acoustical Unit Similarity

Mathieu Lagrange[*,a], Martin Raspaud[b], Roland Badeau[a], Gaël Richard[a]

[a]*Institut Telecom, Telecom ParisTech, CNRS LTCI, 46 rue Barrault 75634 PARIS Cedex 13 - FRANCE*
[b]*Linköping University, Bredgatan 33, SE-60174 Norrköping - SWEDEN*

Abstract

Timbre is a major cue for the human auditory system to recognize musical sounds. Timbral patterns can be decomposed in the widely used spectral envelope and the less considered temporal dynamics.

In this paper, we present new temporal dynamics similarity measures, which will prove valuable for the recognition of timbral patterns. These similarity measures are evaluated, first alone, then in conjunction with spectral envelope similarity measures, for both single tones and solo recordings. Results are provided, showing that the new temporal dynamics features improve significantly timbral pattern recognition.

Key words: Audio Similarity, Timbre modeling, Audio Analysis, Temporal Dynamics

---

1. Introduction

The timbre, along with the perceived loudness, pitch, and duration, is an important attribute. As reported in McAdams and Bigand (1993), contemporary research in psychological and cognitive acoustics decomposes this attribute into several perceptual dimensions of temporal, spectral, and spectro-temporal nature.

It is worth noticing that in the audio signal processing area, the spectral nature of timbre has received much more interest than the others. One of the best examples is the set of features called Mel-Frequency Cepstral Coefficients

*Corresponding author. Tel +33 1 45 81 73 24; Fax: +33 1 45 81 71 44

(MFCC's), which is widely used in speech and speaker recognition systems presented by Rabiner and Juang (1993) as well as in music classifications system such as the one proposed by Tzanetakis and Cook (2002). From a modeling point of view, the spectral envelope is related to the filter part of a source/filter model of the analyzed sound as described by Markel and Gray (1976). In many cases, the spectral properties of this filter are specific to the vibrating body, *i.e.* the vocal tract or the shape of a musical instrument. This makes the modeling of the spectral envelope particularly interesting for the description of musical sounds. From a technical point of view, the spectral envelope can easily be extracted in a frame-based manner with minimal delay.

However, it is widely known that the temporal dimension of timbre is very important at least from a perceptual point of view as reported in Grey and Moorer (1977). As detailed in Section 2, the temporal dynamics of timbre are implicitly modeled by the frame-to-frame variability of the spectral features. The processing of temporal dynamics thus has to be performed subsequently. One can consider another feature that characterizes the variation over time of a previously computed feature, such as the ΔMFCC's presented by Rabiner and Juang (1993). However, recent work of Joder et al. (2009) shows that the use of such integrators do not improve the classification performance. By contrast, we propose in this paper to model explicitly the temporal dynamics of the spectral parameters in order to build spectro-temporal features. The rationale behind the proposed approach is that, compared to feature-level dynamic modeling, the proposed features take into account finer spectro-temporal modulations, which modulations are useful to characterize the audio scene.

The spectro-temporal features are used – through the definition of similarity metrics – to discriminate audio signals produced by different musical instruments. The discrimination performance of the similarity metric in turn gives us an evaluation of the ability of the spectro-temporal features to describe the analyzed sound in a meaningful way.

The remainder of the paper is organized as follow: related work is reviewed and discussed in Section 2. The proposed approach is motivated in Section 3 within different spectral representations to propose several features that will be considered the definition of similarity metrics between acoustical units in Section 4. In light of those experiments, the benefits of spectro-temporal features proposed in this paper in Section 5.

2. Previous work

Using sinusoids was one of the first attempts to model explicitly the temporal dynamics of sound. The sinusoidal model represents pseudo-periodic sounds as sums of sinusoidal components – so-called partials – controlled by parameters that evolve slowly with time as considered by McAulay and Quatieri (1986); Serra (1997):

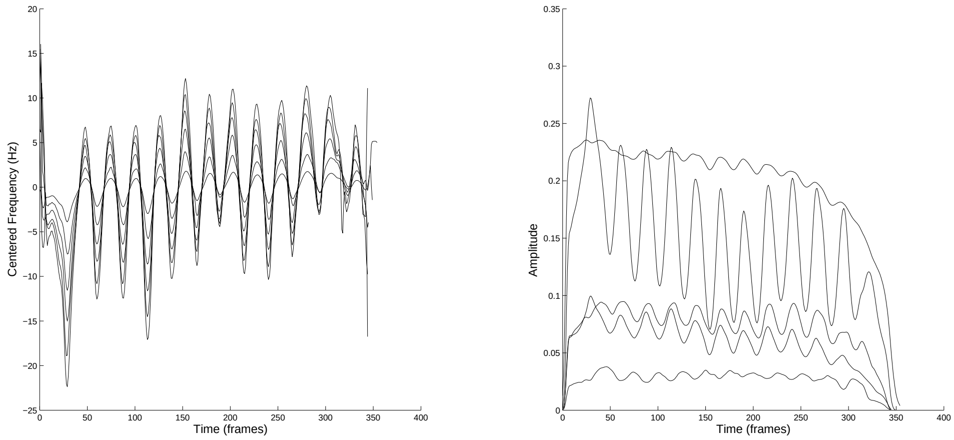
$$p^k(n) = \{f^k(n), a^k(n), \phi^k(n)\} \tag{1}$$

where $f^k(n)$, $a^k(n)$, and $\phi^k(n)$ are respectively the frequency, amplitude, and phase of the partial p^k at frame index n. These parameters are valid for all $n \in [b^k, \cdots, b^k + l^k - 1]$, 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 unit*.

2.1. The Common Variation Cue

When a sound-generating object changes its properties so that its fundamental frequency gets higher or lower, all the partials of the sound also change synchronously. In several experiments McAdams (1989) studied the influence of this phenomenon in the auditory system. Whether the common variation cue is an important cue for the fusion and segregation capacity of the human auditory system is still an open issue as reported by McAdams and Bigand (1993).

However, from a physical point of view, this phenomenon can be measured and can therefore be used to perform the clustering of the partials of the same acoustical unit, as proposed by Lagrange (2005). Let us consider the case of a harmonic set of partials modulated by a vibrato. The frequencies $f^k(n)$ are periodically modulated at the same rate, and the depth of the vibrato is a function of the rank of the partial in the harmonic set. An extra care should be taken while considering the induced modulation of the amplitude. Indeed, depending on the sign of the spectral envelope slope at the partial frequency location, the modulation phase can be shifted. This phenomenon is illustrated by the Figure 1 where the lowest amplitude partial has its amplitude modulated at the same rate but with a $\pi/2$ delay.



(a) Frequencies (b) Amplitudes

Figure 1: *Mean-centered frequencies and amplitudes of some partials of a saxophone tone with vibrato.*

2.2. Integration of Frequency-Axis Features

As with most model-based approaches, the sinusoidal model tends to be brittle when applied to real-world sounds as studied by Lagrange (2004). Therefore, most practical approaches are based on spectral Fourier representations computed in frame-based manner and summarized by numerous means, one of the most famous being the MFCC's Rabiner and Juang (1993). The MFCC's are coefficients that describe the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. By selecting the first coefficients, one can estimate a "smooth" version of the spectrum, usually considered as an approximate of the spectral envelope. This kind of feature is a static observation of the spectral content of the signal and will therefore be termed a *Frequency-Axis Feature* (FAF).

The temporal aspect of the analyzed sound is not completely neglected by the frame-based approach. Indeed, the temporal dynamics are in this case implicitly encoded by the frame-to-frame variability of the frame-based features. This variability can potentially be captured at later stages by the following two approaches previously studied by Aucouturier and Pachet (2007).

The first one, known as classifier or *late integration*, does not try to explicitly extract feature dynamics, but rather operates at the classifier level, usually a supervised classifier with sequentiality constraints, like Hidden Markov Models considered in Eronen (2003); Kitahara et al. (2006) or Sequence kernel-based Support Vector Machines considered in Scaringella and Zoia (2005); Joder et al. (2009); Shimodaira et al. (2002); Cuturi et al. (2007). As these techniques do not deal with explicit modeling of the temporal dynamics of the sound, they will not be discussed further in this paper.

A second approach, known as *feature-level integration*, refers to the computation of a new feature vector that characterizes the evolution of a given set of features at a larger time scale. The most commonly used feature-integration feature is the ΔMFCC's. Meng et al. (2007) have studied more complex models, like high-order auto-regressive models for genre classification. However, Joder et al. (2009) reported that the use of this kind of feature integration actually degrades the classification performance in an instrument recognition task. The poor performance may be due to the nature of FAFs which tends to smooth away potentially meaningful information about the temporal dynamics of the sound.

We introduce in the next section an alternative approach that is to model the temporal dynamics of the spectrum and is consequently termed *spectral-level integration*.

3. Proposed Approach

The spectral envelope is an important piece of information if we want to characterize the timbre of an audio signal. However, we believe that this should be complemented by *Time-Axis Features* (TAFs) that explicitly model the temporal dynamics of the spectral components of the sound.

Let us consider a simplified version the source/filter model in order to better motivate our approach. The FAFs mainly model the filter part of the sound production chain. In order to complement these features, it is important to avoid any redundancy and therefore focus on a different aspect of the sound production chain. Following the source/filter dichotomy, we propose to root the TAFs on the source part.

We now introduce three types of TAFs based respectively on frequency, amplitude and magnitude parameters of sound.

3.1. Frequency Evolution Features

We design Frequency Evolution Features (FEFs) such that they encode the modulation of the frequency of the main components of the spectrum over time. It is therefore natural to choose the sinusoidal model as a signal representation.

In a first approach, the temporal evolution of the frequency of the partials f^k can be considered directly. Alternatively, since the Fourier transform is based on the periodicity of the input signal, using a spectrum of the evolutions of partial parameters might show common periodicities of the partials. This will be useful 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 as studied by Mellody and Wakefield (2000) and Marchand and Raspaud (2004). It can also be interesting for micro-modulations such as the ones produced by vibrating strings such as the strings of a piano (see Figure 2).

Let us define the following operator, based on the complex modulus of the short-time Fourier transform:

$$X(k) = \left| \sum_{n=0}^{N-1} (x(n) - \bar{x}) h(n) e^{\frac{-2j}{N}\pi k n} \right|^2 \tag{2}$$

$$|\mathcal{F}|_{n_l}^{n_h}(x(n)) = \{X(m) | n_l < m < n_h\} \tag{3}$$

where N is the size of the Fourier Transform, h is an analysis window, and $\bar{x}$ denotes the mean of x. n_l and n_h are respectively the minimal and maximal frequency indexes considered. Thanks to the complex modulus applied to the spectrum, this operator is phase-invariant. We can then define *Frequency Evolution Features* (FEF), as:

$$F^k = |\mathcal{F}|_1^{N-1}(f^k) \tag{4}$$

$$F_c^k = |\mathcal{F}|_1^{n_c}(f^k) \tag{5}$$

where f^k is the frequency of the partial k and n_c is chosen in order to reduce the dimensionality of the feature while keeping low frequency content information.

3.2. Amplitude Evolution Features

In order to encode the temporal dynamics of the amplitude, we can in a similar fashion consider the evolution of the amplitude a^k of the partials p^k

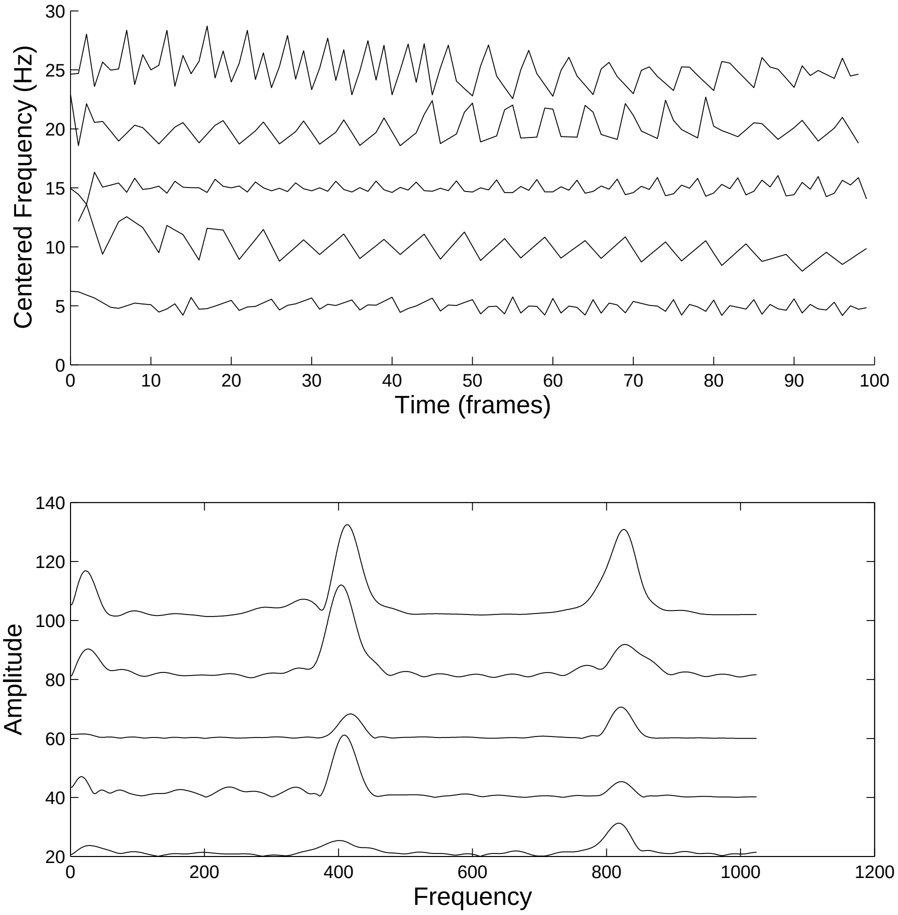


Figure 2: *Centered frequencies (top) of a piano note from the IOWA database and their corresponding spectra (bottom). Each curve is shifted and the spectra are interpolated using zero-padding for clarity sake.*

or the corresponding spectral features, called *Amplitude Evolution Features* (AEF):

$$A^k = |\mathcal{F}|_1^{N-1}(a^k) \tag{6}$$

$$A_c^k = |\mathcal{F}|_1^{n_c}(a^k) \tag{7}$$

In order to consider only the modulated part of the amplitude signal, leaving aside the global envelope, it is relevant to decompose the signal in two components, one being polynomial, and the other being pseudo-periodic as proposed by Raspaud et al. (2005):

$$a^k(t) = \Pi(t) + \sum_i \alpha_i(t)\cos(\psi_i(t)) \tag{8}$$

where $\Pi(t)$ is a polynomial, and $\alpha_i(t)$ and $\psi_i(t)$ are the parameters of sinusoidal components, see Figure 3(a).

Indeed, while subtracting the mean of the signal – as performed by the operator $|\mathcal{F}|$ – is enough to center the oscillations of the evolution of the frequency of partials, it is not the case for the evolution of the amplitude of partials. As studied by Raspaud (2007), the idea behind this polynomial subtraction is that the envelope of a sound (seen as attack, decay, sustain and release) can be approximated by a 9th degree polynomial. Then, we define the following two other AEFs based solely on the oscillating part of the partials amplitudes.

$$a_p^k = a^k - \tilde{\Pi}(a^k) \tag{9}$$

$$A_p^k = |\mathcal{F}|_1^{N-1}(a_p^k) \tag{10}$$

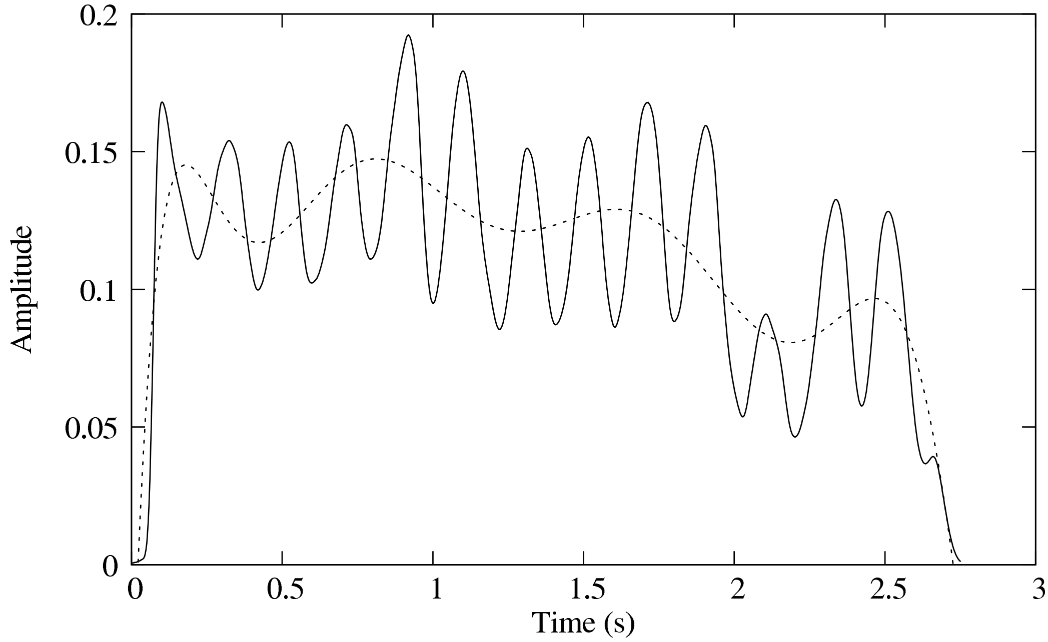
$$A_{p,c}^k = |\mathcal{F}|_1^{n_c}(a_p^k) \tag{11}$$

where $\tilde{\Pi}(x)$ is the envelope polynomial computed from signal x using a simple least-squares method. As an approximation of the polynomial removal, one can consider removing the DC component in the following way:

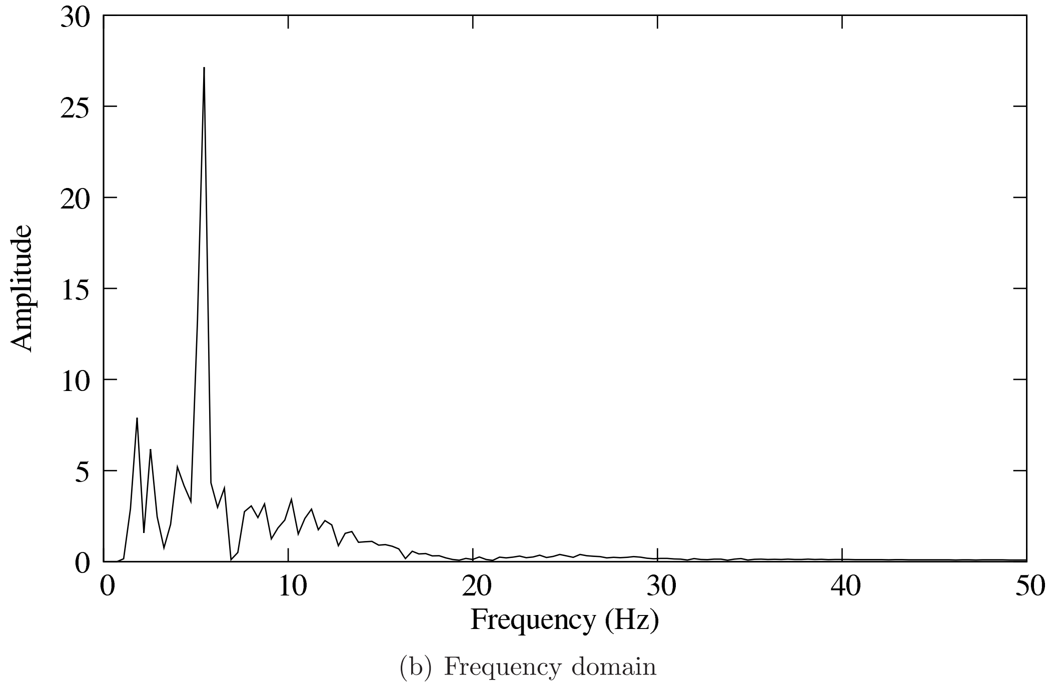
$$A_d^k = |\mathcal{F}|_{n_d}^{N-1}(a_p^k) \tag{12}$$

$$A_{d,c}^k = |\mathcal{F}|_{n_d}^{n_c}(a_p^k) \tag{13}$$

where n_d is chosen so that periodicities like the tremolo are preserved, see Figure 3(b).



(a) Time domain



(b) Frequency domain

Figure 3: *Amplitude of a partial, its estimated polynomial envelope and the corresponding frequency domain representation of the polynomial-removed amplitude evolution.*

3.3. Magnitude Evolution Features

As stated previously, the sinusoidal model provides a meaningful representation for analyzing the temporal dynamics. However, the estimation of this model from complex signals can hardly be done in a fully automatic fashion. As an approximation of the AEFs, we also consider in the second experiment *Magnitude Evolution Features* (MEF) that rely on the spectrogram only. Taking into account the evolution of the magnitude in the spectrogram is a non parametric way to account for the temporal dynamics, *i.e.* without relying on the sinusoidal model.

Let us consider $X(k,n)$ the spectral bin k of the frame n of the spectrogram of the signal x. For a given spectral bin of frequency index k, the MEFs correspond to the magnitude evolution of a frequency line in the spectrogram:

$$m^k \;=\; X(k,n) \tag{14}$$

where n varies within a given horizon of observation. The other MEFs M^k, M_c^k, M_p^k, $M_{p,c}^k$, M_d^k and $M_{d,c}^k$ are computed as described in the previous section.

4. Acoustical Units Similarity

We define an acoustical unit to be a musical tone or a sequence of musical tones performed by a unique musical instrument, within a limited time interval. The task is then to decide whether two acoustical units have been played by the same musical instrument or not. This decision is made according to the information given by the features computed from the acoustical unit. This evaluation task is chosen for two reasons.

From a practical application point of view, there is an increase of interest towards recommendation systems that are not based on an ontology such as genre as used by Tzanetakis and Cook (2002) or instrument type considered by Joder et al. (2009). Alternatively, one can consider a recommendation system that states "show me tunes that are comparable to the ones I like". In this case, one needs to define the similarity between musical audio signals. For the definition of such a similarity, the timbre is an interesting dimension.

From a scientific point of view, this allows us to propose a much simpler evaluation framework than the one required by classification-based systems (such as the previously cited genre and instrument type). Indeed, the latter rely on training complex classifiers which may have constraints over the statistical properties of the features and can consequently introduce a bias over the evaluated features.

4.1. Features Integration

One issue with the proposed scheme is that the dimensionality of the problem is the square of the number of elements to be sorted. Consequently, we are interested in efficiently describing longer acoustical units than those that are usually considered in standard classification systems. Hence, two fixed sizes (or integration intervals) are considered, a "texture" duration of about 22 ms and an "event" duration of about 1 second. When considering isolated notes, the acoustical unit duration is adapted to the actual duration of the note. Within this acoustical unit, the features are extracted from a spectrogram computed in the following way:

1. the input signal is first filtered with a DC block filter and a pre-emphasis filter.
2. The spectra are next computed with a frame size of 40 ms and a hop size of 10 ms.

Concerning the FAFs features, the MFCCs are computed with the Matlab implementation of Ellis (2005) within each analysis frame with 40 Mel subbands and only the 12 DCT coefficients after the zeroth coefficient are considered. The ΔMFCCs are computed using a 5-point derivation filter. Finally, those coefficients are summed over the integration interval, as proposed in Joder et al. (2009).

The TAFs features are extracted within each acoustical unit in the following way:

1. the spectrogram is first integrated over time and split in 20 Mel subbands.
2. Within each subband, the bin with maximal amplitude is selected, indicating the potential frequency location of a partial.

The magnitude evolution within each of those bins is considered for the computation of the MEFs. In order to extract the FEFs and the AEFs, partials are tracked in a frequency interval around each of the maximal amplitude bins using a standard partial tracker proposed by Ellis (2003). The width of this interval is set to $\approx$ 220 Hz. The partial with maximal cumulative amplitude within each frequency interval is then selected to compute the TAFs.

The TAFs are associated to 20 partials belonging to an acoustical unit. However, in this task, we need to define the similarity between acoustical units. Consequently,

- the TAFs describing each partial of a given acoustical unit are integrated by summation.
- The time domain features f^k, a^k, and m^k are resampled to have the same length (20 in our experiments) prior to summation.
- The base-10 logarithm is applied to the integrated spectral features which are computed using an $|\mathcal{F}|$ operator with $N = 64, n_c = 16$ and $n_d = 2$.

4.2. Similarity Metrics

The features described above are next considered to build similarity metrics after being normalized to have zero mean and unity variance. In order to demonstrate the ability of the proposed TAFs features to complement a FAF feature like the MFCCs, the combination with the MFCCs is also considered. For that purpose, the radial basis function is considered:

$$s(U_i(k), U_j(l)) = e^{-||V_i(k) - V_j(l)||} \tag{15}$$

where $||x||$ is the Euclidean norm of x and $V_i(k)$ is a feature vector or a stacking of feature vectors describing the acoustical unit $U_i(k)$ played by the instrument i. In order to account for the discrepancy between feature dimensionalities, the features are divided by the square root of their dimensions prior to stacking.

4.3. Evaluation Metrics

In order to evaluate a set of features describing an acoustical unit, the value of the similarity metric applied to this set of features is first computed for every pair of acoustical units. To compute performance indicators, we propose to cast our multi-class problem into a binary one, *i.e.* even though the databases comprise acoustical units played by multiple instruments, we evaluate a classifying task which corresponds to the binary question "Are those acoustical units played by the same instrument?". Therefore, at a given classifying threshold T_c, we can define the *False-Alarm Rate* (FAR) as:

$$\text{FAR} = \frac{\#\{s(U_i(k), U_j(l)) > T_c\}}{\#U.(\#U - 1)} \tag{16}$$

where $\#X$ denotes the cardinal of X and U is the overall set of acoustical units in the evaluation database. Similarly, the *Miss Detection Rate* (MDR) is defined as:

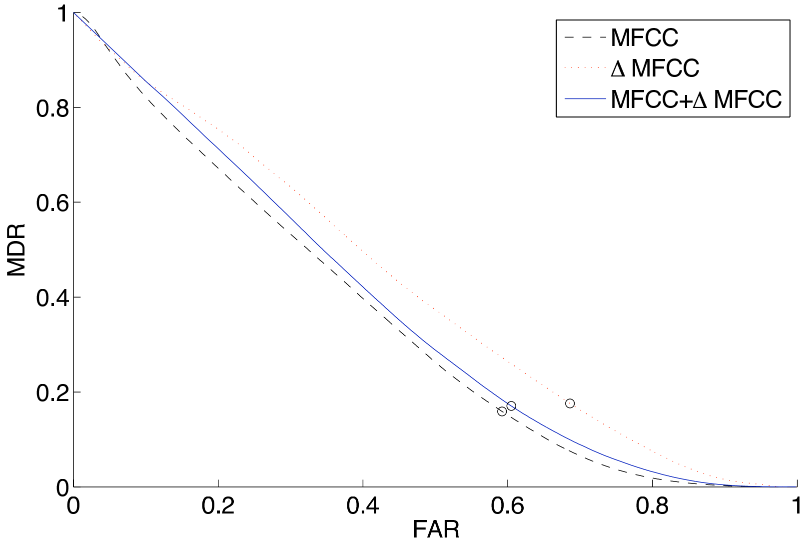
$$\text{MDR} = \frac{\#\{s(U_i(k), U_i(l)) < T_c\}}{\sum_i \#U_i.(\#U_i - 1)} \tag{17}$$

The *Detection Error Trade-Off* (DET) curve proposed by Martin et al. (1997) is used to visualize the performance of the classifier corresponding to the evaluated feature or combination of features at a varying classification threshold, see Figure 4. In order to summarize the behavior of the classifier depending on the chosen threshold, we consider three criteria. The first one called *Equal Error Rate* (EER) corresponds to the crossing of the DET curve with a line that starts from $(0, 0)$ coordinates with unity slope. The second one called *Minimum Cost Point* (MCP) indicates the performance of the classifier for an optimal error trade-off between a balanced weighting of the MDR and FAR as described in Martin et al. (1997). The MCP for each feature is plotted on Figure 4 with a small circle. The last criterion, called AREA, computes the area under the curve and indicates the general behavior of the classifier. For all of those criteria, a lower value indicates a better performance.

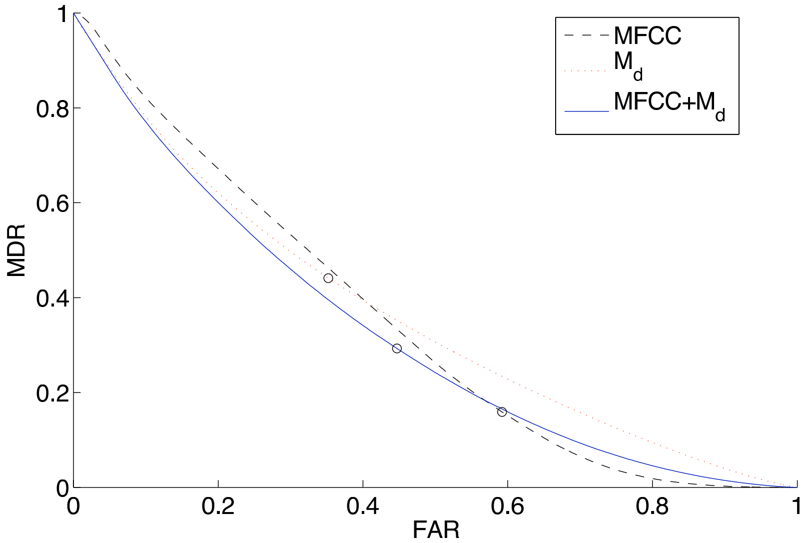
4.4. Isolated Musical Tones

We consider in a first experiment, two databases of isolated tones, the IOWA database and the RWC database proposed by Goto et al. (2003). The IOWA database features 3637 tones of mean duration 4.3 seconds and standard deviation 7.5 seconds from 15 Musical Instruments for a total of 4.3 hours. The RWC database features 6138 tones of mean duration 2.6 seconds and standard deviation 0.8 seconds for a total of 4.7 hours.

Considering isolated notes is helpful to study the performance of the evaluated features in a controlled environment. Indeed, in this case, the observation interval is at an optimal location and duration and no frequency resolution issues may arise except for very low pitched tones, making the tracking of partials relatively non-ambiguous.



(a)



(b)

Figure 4: *DET curves of selected features computed over the IOWA database. On both axis, a lower value indicates better performance.*

We first evaluate the features solely, see Table 1. For each of the TAFs, the spectral features perform best. For the amplitude and magnitude ones, removing the polynomial before the spectral transformation is helpful. Even though there should be no tracking issues, the magnitude features clearly dominate the other TAFs, leading to comparable results with the MFCCs. It should be noticed that there is a large correlation between the different evaluation criteria (EER, AREA, and MCP) demonstrating a good behavior of the evaluated classifiers.

To fully understand the behavior of the TAFs when considered jointly with the MFCCs, we plot on Figure 4 the DET curves for several features computed over the IOWA database. On Figure 4(a), the DET curves of the MFCCs, ΔMFCCs are considered solely or jointly. The curve of the MFCCs show a typical behavior with a linear slope on the high MDR range. The DET curve of the ΔMFCCs shows similar evolution properties with worst performance. As they are highly correlated, considering them jointly leads to an averaging of their performance.

On the contrary, the DET curve of the M_d feature shows a different behavior, almost symmetrical towards the unity slope. Considering them jointly leads to a very balanced classifier with good properties. Empirical experiments show that a better joint feature can be obtained by weighting the features prior to similarity calculation.

Considering the proposed TAFs jointly with the MFCCs leads to an improvement with respect to the performance of the TAFs. However, only the use of the MEFs leads to an improvement with respect to the sole use of the MFCCs, see Table 2.

4.5. Solo Performance Music

We consider in a second experiment a database of solo recordings used in several musical instruments classification experiments done by Essid et al. (2004) and Joder et al. (2009). The SOLOS database features 505 solos recordings where each one is of mean duration 110 seconds and standard deviation 162 seconds performed by 20 different instruments for a total of 15.56 hours.

As the tracking is more difficult when considering less constrained sounds, the performance of the FEFs and the AEFs do not improve compared to the previous experiment. Consequently, only the spectral MEF is considered here. Also, the use of the DC or polynomial removal does not lead to a significant difference. As shown on Figure 5, the properties of the different

Criterion	Database	f	F	F_c	a	A	A_c	a_p	A_p	$A_{p,c}$
EER	IOWA	0.699	**0.6**	0.621	0.659	0.655	0.656	0.69	0.639	**0.618**
	RWC	0.72	**0.687**	0.69	0.696	0.68	0.677	0.726	0.677	**0.675**
AREA	IOWA	0.491	**0.401**	0.418	0.458	0.451	0.45	0.484	0.434	**0.411**
	RWC	0.512	**0.48**	0.482	0.491	0.474	0.47	0.518	0.472	**0.465**
MCP	IOWA	0.928	**0.61**	0.629	0.66	0.656	0.656	0.701	0.645	**0.619**
	RWC	0.928	**0.687**	0.691	0.723	0.684	0.677	0.977	0.679	**0.675**

Criterion	Database	$MFCC$	$\Delta MFCC$	m	M	M_c	m_p	M_p	$M_{p,c}$	M_d	$M_{c,d}$
EER	IOWA	**0.56**	0.62	0.71	0.56	0.62	0.71	**0.54**	0.61	0.56	0.63
	RWC	**0.58**	0.65	0.73	0.69	0.69	0.72	**0.68**	0.69	0.69	0.7
AREA	IOWA	**0.34**	0.41	0.49	0.35	0.41	0.49	**0.34**	0.40	0.36	0.42
	RWC	**0.37**	0.44	0.52	0.48	0.49	0.51	**0.48**	**0.48**	0.48	0.49
MCP	IOWA	**0.61**	0.70	0.92	0.56	0.64	0.87	**0.54**	0.61	0.56	0.65
	RWC	**0.67**	0.69	0.96	0.80	0.79	0.96	0.75	**0.75**	0.80	0.77

Table 1: Results for single tones and single features. Best results for each feature group are displayed in bold characters.

Criterion	Database	MFCC+								
		f	F	F_c	a	A	A_c	a_p	A_p	$A_{p,c}$
EER	IOWA	0.609	**0.567**	0.568	0.604	0.586	0.586	0.613	0.579	**0.56**
	RWC	0.649	**0.613**	0.619	0.634	0.607	**0.606**	0.637	0.608	0.609
AREA	IOWA	0.398	**0.357**	0.36	0.392	0.377	0.376	0.4	0.369	**0.349**
	RWC	0.441	**0.409**	0.413	0.426	0.404	**0.401**	0.43	0.406	**0.401**
MCP	IOWA	0.645	0.592	**0.578**	0.614	0.611	0.614	0.695	0.602	**0.577**
	RWC	0.726	**0.627**	0.644	0.651	0.632	**0.626**	0.75	0.63	0.63

Criterion	Database	MFCC+								
		ΔMF	m	M	M_c	m_p	M_p	$M_{p,c}$	M_d	$M_{c,d}$
EER	IOWA	**0.578**	0.608	0.526	0.551	0.617	**0.522**	0.558	0.526	0.557
	RWC	**0.621**	0.651	**0.6**	0.613	0.644	0.604	0.622	0.601	0.613
AREA	IOWA	**0.364**	0.395	0.32	0.345	0.404	**0.316**	0.346	0.321	0.35
	RWC	**0.411**	0.443	**0.398**	0.41	0.435	0.399	0.414	**0.398**	0.41
MCP	IOWA	**0.629**	0.663	0.534	0.556	0.677	**0.53**	0.573	0.534	0.565
	RWC	**0.696**	0.73	0.625	0.639	0.73	0.636	0.668	**0.624**	0.637

Table 2: Results for single tones for joint features. Best results for each feature group are displayed in bold characters.

Criterion	Integration type	MFCC	ΔMFCC	M	M_c	MFCC+ ΔMFCC	MFCC+ M	MFCC+ M_c
EER	Texture	0.523	0.661	0.552	0.584	0.579	**0.519**	0.526
	Event	**0.538**	0.665	0.615	0.623	0.59	0.56	0.57
AREA	Texture	0.314	0.456	0.354	0.378	0.357	**0.297**	0.304
	Event	**0.317**	0.45	0.408	0.425	0.363	0.334	0.341
MCP	Texture	0.509	0.658	0.552	0.579	0.557	**0.501**	0.509
	Event	**0.515**	0.651	0.61	0.622	0.563	0.536	0.544

Table 3: Results for solo recordings. Best results are displayed in bold characters.

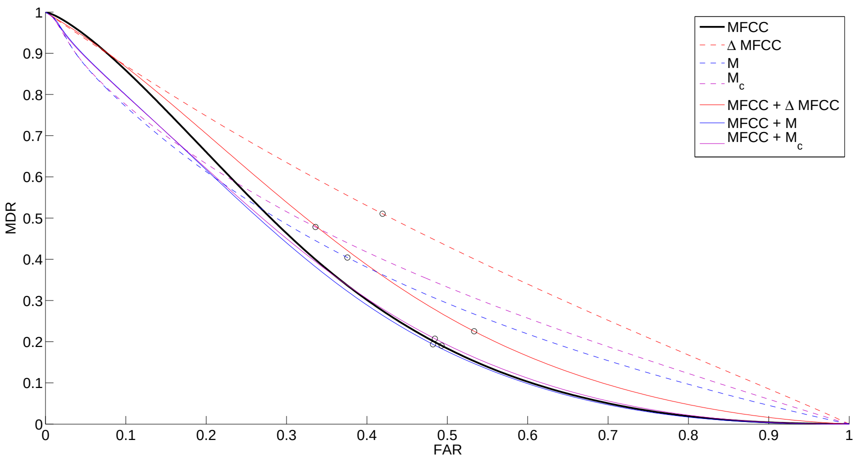


Figure 5: *DET Curves for selected features computed over the SOLOS database with a texture observation interval.*

features discussed in the previous section are rather similar when considering real-world sounds. The use of the dimensionality reduction over the M feature leads to a slight decrease in the performance. The longer duration of the acoustical unit leads to a general performance decrease as shown by Table 3. This decrease is small concerning the MFCC. On contrary, it is significant for the MEFs. This demonstrates that the length of the observation interval is crucial for capturing meaningful periodicities in the variations of the spectral parameters.

5. Conclusion

We have proposed in this paper several approaches for extracting the evolution of the spectral parameters over time and for modeling them in a meaningful way. We show that the spectral description of the evolution of the parameters of the partials is relevant as well as the removal of the polynomial part of the amplitude evolution when considering isolated notes databases. When facing less constrained scenarios, such as the solos database, the removal of the polynomial part does not improve the results whereas the spectral description is still in favor, leading to a relevant feature set when combined with the MFCCs. The loss of relevance of the polynomial removal may be due to the fact that the boundaries of the acoustical units considered in the last experiment are arbitrary set.

As a conclusion, the proposed features are found to be more adapted to the tasks considered in this paper than the standard feature-level temporal dynamic features usually considered, the $\Delta MFCCs$.

A simple scheme for reducing the dimensionality of the spectral features has been considered. We feel that dedicated dimensionality reduction schemes should be researched for as they could maintain similar performance while reducing the computation cost. Also, the results obtained by the FEFs and AEFs may be limited by the partial tracker considered in the experiments. The use of more advanced algorithm like those described in Lagrange et al. (2007); Robel (2006) may reduce the gap between model-based features (FEF and AEF) and transform-based features (MEF).

6. Acknowledgments

The authors want to thank Dr. Ellis for providing the community with the code for MFCC computation and partial tracking. This work has been partly funded by the Quaero project within the task 6.4: “Music Search by Similarity” and the French GIP ANR under contract ANR-06-JCJC-0027-01, “Décompositions en Éléments Sonores et Applications Musicales” - DESAM.

References

Aucouturier, J.-J., Pachet, F., 2007. The influence of polyphony on the dynamical modelling of musical timbre. Pattern Recognition Letters 28 (5), 654–661.

Cuturi, M., Vert, J.-P., Birkenes, O., Matsui, T., April 2007. A kernel for time series based on global alignments. In: Proc. of IEEE ICASSP. Vol. 2. pp. 413–416.

Ellis, D. P. W., 2003. Sinewave and sinusoid+noise analysis/synthesis in Matlab. Online web resource.
URL http://www.ee.columbia.edu/~dpwe/resources/matlab/sinemodel/

Ellis, D. P. W., 2005. PLP and RASTA (and MFCC, and inversion) in Matlab. Online web resource.
URL http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

Eronen, A., 2003. Musical instrument recognition using ica-based transform of features and discriminatively trained hmms. In: 7th International Sumposium on Signal Processing and its Applications.

Essid, S., Richard, G., David, B., 2004. Musical instrument recognition based on class pairwise feature selection. In: Proc. of ISMIR.

Goto, M., Hashiguchi, H., Nishimura, T., Oka, R., 2003. RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003).

Grey, J. M., Moorer, J. A., 1977. Perceptual Evaluations of Synthesized musical instrument tones. Journal of Acoustical Society of America 62 (3), 454–462.

Joder, C., Essid, S., Richard, G., 2009. Temporal Integration for Audio Classification with Application to Musical Instrument Classification. IEEE Transactions on Acoustics, Speech and Language Processing 17 (1), 174–186.

Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H. G., 2006. Musical instrument recognizer "instrogram" and its application to music retrieval based on instrumentation similarity. In: Proceedings of the IEEE International Symposium on Multimedia.

Lagrange, M., December 2004. Sinusoidal Modeling of Polyphonic Sounds. Ph.D. thesis, University of Bordeaux 1, LaBRI, in french.

Lagrange, M., September 2005. A New Dissimilarity Metric For The Clustering Of Partials Using The Common Variation Cue. In: Proc. ICMC. ICMA, Barcelona, Spain.

Lagrange, M., Marchand, S., Rault, J., 2007. Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds. IEEE Transactions on Acoustics, Speech and Language Processing 28, 357–366.

Marchand, S., Raspaud, M., October 2004. Enhanced Time-Stretching Using Order-2 Sinusoidal Modeling. In: Proc. DAFx. Federico II University of Naple, Italy, pp. 76–82.

Markel, J. D., Gray, A. M., 1976. Linear Prediction of Speech. Berlin: Sringer-Verlag.

Martin, A., Doddington, G., Kamm, T., ordowski, M., Przybocki, M., 1997. The DET curve in Assessment of Detection Task Performance. In: Proceedings of EuroSpeech.

McAdams, S., 1989. Segregation of Concurrrents Sounds : Effects of Frequency Modulation Coherence. JAES 86 (6), 2148–2159.

McAdams, S., Bigand, E., 1993. Thinking in Sound. Oxford Science Publications, Ch. 2.5.

McAulay, R. J., Quatieri, T. F., 1986. Speech Analysis/Synthesis Based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech and Signal Processing 34 (4), 744–754.

Mellody, M., Wakefield, G., 2000. The time-frequency characteristic of violin vibrato: modal distribution analysis and synthesis. Journal of the Acoustical Society of America 107, 598–611.

Meng, A., Ahrendt, P., Larsen, J., Hansen, L., 2007. Temporal Feature Integration for Music Genre Classification. IEEE Transactions on Acoustics, Speech and Signal Processing 15 (5), 1654–1664.

Rabiner, L., Juang, B. H., 1993. Fundamentals of Speech Recognition. Prentice Gall Signal Processing Series.

Raspaud, M., 2007. Hierarchical spectral models for sound and applications. Ph.D. thesis, University of Bordeaux 1.

Raspaud, M., Marchand, S., Girin, L., September 2005. A Generalized Polynomial and Sinusoidal Model for Partial Tracking and Time Stretching.

In: Proc. DAFx. Universidad Politecnica de Madrid, pp. 24–29, iSBN: 84-7402-318-1.

Robel, A., 2006. Adaptive additive modeling with continuous parameter trajectories. IEEE Transactions on Acoustics, Speech and Language Processing 14 (4), 1440–1453.

Scaringella, N., Zoia, G., 2005. On the modelling of time information for automatic genre recognition systems in audio signals. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR).

Serra, X., 1997. Musical Signal Processing. Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, Ch. Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122.

Shimodaira, H., Noma, K., Nakai, M., Sagayama, S., 2002. Dynamic time-alignment kernel in support vector machine. Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press.

Tzanetakis, G., Cook, P., 2002. Musical genre classification of audio signals. IEEE Transactions on Acoustics, Speech and Language Processing 10 (5), 293–302.