Dynamic Spectral Envelope Modeling for Timbre Analysis of Musical Instrument Sounds

Juan José Burred, *Member, IEEE*, Axel Röbel, and Thomas Sikora, *Senior Member, IEEE*

***Abstract*—We present a computational model of musical instrument sounds that focuses on capturing the dynamic behavior of the spectral envelope. A set of spectro–temporal envelopes belonging to different notes of each instrument are extracted by means of sinusoidal modeling and subsequent frequency interpolation, before being subjected to principal component analysis. The prototypical evolution of the envelopes in the obtained reduced-dimensional space is modeled as a nonstationary Gaussian Process. This results in a compact representation in the form of a set of prototype curves in feature space, or equivalently of prototype spectro–temporal envelopes in the time–frequency domain. Finally, the obtained models are successfully evaluated in the context of two music content analysis tasks: classification of instrument samples and detection of instruments in monaural polyphonic mixtures.**

***Index Terms*—Gaussian processes, music information retrieval (MIR), sinusoidal modeling, spectral envelope, timbre model.**

I. Introduction

We address the development of a novel computational modeling approach for musical instrument sounds focused on capturing the temporal evolution of the spectral envelope. We intend the models to be used not only as a mid-level feature in classification tasks, but also as source of *a priori* knowledge in applications requiring not only model discrimination, but also a reasonable degree of model accuracy, such as detection of instruments in a mixture, source separation, and synthesis applications. In this contribution, we present in detail the design guidelines and evaluation procedures used during the development of such a modeling approach, as well as performance evaluations of its application to the classification of individual instrumental samples and to the recognition of instruments in monaural (single-channel) polyphonic mixtures.

The temporal and spectral envelopes are two of the most important factors contributing to the perception of timbre [1]. The temporal envelope, usually divided into Attack, Decay, Sustain, and Release (ADSR) phases, is a valuable feature to distinguish, for instance, between sustained (bowed strings, winds) and constantly decaying instruments (plucked or struck strings). The spectral envelope can be defined as a smooth function of frequency that approximately matches the individual partial peaks of each spectral frame. The global shape of the frame-wise evolution of the individual partial amplitudes (and consequently of the spectral envelope) corresponds approximately to the global shape of the temporal envelope. Thus, considering the spectral envelope and its temporal evolution makes it unnecessary to consider the temporal envelope as a separate entity. We will use the term "spectro–temporal envelope" to globally denote both the frame-wise spectral envelope and its evolution in time. We emphasize that the present method considers timbre (a perceptual sensation) to be mainly affected by the spectro–temporal envelope (a physical aspect). It should be noted, however, that there are other factors that can have an important influence on timbre, such as harmonicity, noise content, transients, masking effects, and auditory and neural processes.

An early work thoroughly and systematically assessing the factors that contribute to timbre was the 1977 work by Grey [2]. He conducted listening tests to judge perceptual similarity between pairs of instrumental sounds, and applied *multidimensional scaling* (MDS) to the results for reducing the dimensionality. In the cited work, MDS was used to produce a three-dimensional *timbre space* where the individual instruments clustered according to the evaluated similarity.

In later works, similar results were obtained by substituting the listening tests by objectively measured sound parameters. Hourdin, Charbonneau, and Moussa [3] applied MDS to obtain a similar timbral characterization from the parameters obtained from sinusoidal modeling. They represented trajectories in timbre space corresponding to individual notes, and resynthesized them to evaluate the sound quality. Similarly, Sandell, and Martens [4] used *principal component analysis* (PCA) as a method for data reduction of sinusoidal modeling parameters.

De Poli and Prandoni [5] proposed their *sonological models* for timbre characterization, which were based on applying either PCA or *self organizing maps* (SOM) to a description of the spectral envelope based on *Mel frequency cepstral coefficients* (MFCCs). A similar procedure by Loureiro, de Paula, and Yehia [6] has recently been used to perform clustering based on timbre similarity.

Jensen [7] developed a sophisticated framework for the perceptually meaningful parametrization of sinusoidal modeling parameters. Different sets of parameters were intended to describe in detail the spectral envelope, the mean frequencies, the ADSR envelopes with an additional "End" segment, and amplitude and frequency irregularities.

Manuscript received December 31, 2008; revised October 26, 2009. Current version published February 10, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Laurent Daudet.

J. J. Burred and A. Röbel are with the Analysis/Synthesis Team, IRCAM-CNRS STMS, 75004 Paris, France (e-mail: burred@ircam.fr).

T. Sikora is with the Communication Systems Group, Technical University of Berlin, 10587 Berlin, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2036300

In Leveau *et al.* [8], timbre analysis is addressed from the perspective of sparse signal decomposition. A musical sound is approximated as a linear combination of *harmonic atoms*, where each atom is a sum of harmonic partials whose amplitudes are learned *a priori* on a per-instrument basis. A modified version of the *Matching Pursuit* (MP) algorithm is then used in the detection stage to select the atoms that best describe the observed signal, which allows single-voice and polyphonic instrument recognition.

A great variety of spectral features have been proposed in the context of audio content analysis, first in fields such as *automatic speech recognition* (ASR) or sound analysis and synthesis, later in *music information retrieval* (MIR). Most of them are basic measures of the spectral shape (centroid, flatness, rolloff, etc.), and are too simple to be considered full models of timbre. More sophisticated measures make use of psychoacoustical knowledge to produce a compact description of spectral shape. This is the case of the very popular MFCCs [9], which are based on a Mel-warped filter bank and a cepstral smoothing and energy compaction stage achieved by a *discrete cosine transform* (DCT). However, MFCCs provide a rough description of spectral shape and are thus unsuitable for applications requiring a high level of accuracy.

The MPEG-7 standard includes spectral basis decomposition as feature extraction [10]. The extraction is based on an estimation of a rough overall spectral shape, defined as a set of energies in fixed frequency bands. Although this shape feature is called Audio Spectrum Envelope, it is not a spectral envelope in the stricter sense of matching the partial peaks.

Our approach aims at combining an accurate spectral feature extraction front-end with a statistical learning procedure that faithfully captures dynamic behavior. To that end, we first discuss the general criteria that guided the design of the modeling approach (Section II). The main part of this paper (Sections III and IV) is a detailed description of the proposed sound modeling method, which is divided into two main blocks: the *representation stage* and the *prototyping stage*. The representation stage (Section III) corresponds to what, in the pattern recognition community, is called the feature extraction stage. It describes how the spectro–temporal envelopes are estimated from the training samples by means of sinusoidal modeling and subsequent frequency interpolation and dimensionality reduction via PCA, and places special emphasis on discussing the formant alignment issues that arise when using notes of different pitches for the training. This section includes the description of a set of experiments (Section III-D) aimed at evaluating the appropriateness of the chosen spectral front-end. The prototyping stage (Section IV) aims at learning statistical models (one model per instrument) out of the dimension-reduced coefficients generated in the representation stage. In order to reflect the temporal evolution in detail, the projected coefficient trajectories are modeled as a set of Gaussian processes (GP) with changing means and variances. This offers possibilities for visualization and objective timbre characterization, as will be discussed in detail. Finally, the application of the trained models in two MIR tasks will be presented: Section V addresses the classification of isolated musical instrument samples and Section VI the more demanding task of detecting which instruments are present on a single-channel mixture of up to four instruments. Conclusions are summarized in Section VII, together with several possible directions for future research.

The modeling method presented here was first introduced in [11]. That work addressed the evaluation of the representation stage, but it lacked detail about the sinusoidal modeling and basis decomposition procedures and, most importantly, it only provided a very brief mention of the prototyping stage (i.e., the temporal modeling as Gaussian processes), without any formalized presentation. The present contribution provides all missing details and contains a full presentation and discussion of the prototyping stage, together with new experiments and observations concerning the interpretation of the obtained prototypical spectral shapes. More specifically, it addresses the influence of the extracted timbre axes (introduced later) on the spectral shape, the observation of formants (Section IV), and the influence of the frequency alignment procedure on the inter-instrument classification confusion (Section V). The application of the models for polyphonic instrument recognition has been presented more extensively in [12]. Since the main focus here was the design of the modeling approach, we only provide a brief presentation thereof in Section VI, and we refer the reader to that work for further details concerning that particular application. Finally, another related article is [13], where the models were used for source separation purposes. In particular, source separation is based on extending the polyphonic recognition procedure of Section VI to recover missing or overlapping partials by interpolating the prototypical time–frequency templates. However, since the emphasis here was on sound analysis, such a topic is not covered here.

II. Design Criteria

In benefit of the desired multipurpose nature of the models, the following three design criteria were followed and evaluated during the development process: *representativeness*, *compactness*, and *accuracy*. The above mentioned methods fulfill some of the criteria, but do not meet the three conditions at the same time. The present work was motivated by the goal of combining all three advantages into a single algorithm. Each criterion has an associated objective measure that will be defined later (Section III-D). It should be noted that these measures were selected according to their appropriateness within the context of the signal processing methods used here, and they should be considered only an approximation to the sometimes fairly abstract criteria (e.g., representativeness) they are intended to quantify. Another simplification of this approach worth mentioning is that the criteria are considered independent from each other, while dependencies do certainly exist. What follows is a detailed discussion of how the approaches from the literature reviewed above meet or fail to meet the criteria, and how those limitations are proposed to be overcome.

A. Representativeness

An instrument model should be able to reflect the essential timbral characteristics of any exemplar of that instrument (e.g., the piano model should approximate the timbre of any model and type of piano), and be valid for notes of different pitches, lengths, dynamics and playing styles. We will refer to this

requirement as the representativeness criterion. This requires using a training database containing samples with a variety of those factors, and a consequent extraction of prototypes.

Many of the above-mentioned methods focus on the automatic generation of timbre spaces for the subsequent timbral characterization of individual notes, rather than on training representative instrument models valid for a certain range of pitches, dynamics, etc. For instance in [3], [4], and [6], several notes are concatenated to obtain common bases for generating the timbre spaces; there is however no statistical learning of the notes' projections from each instrument into a parametric model. In [5], a static Gaussian modeling approach is proposed for the clusters formed by the projected coefficients. MFCCs and the MPEG-7 approach are indeed intended for large-scale training with common pattern recognition methods, but as mentioned they do not meet the requirement of accuracy of the envelope description. In this paper, we propose a training procedure consisting of extracting common spectral bases from a set of notes of different pitches and dynamics, followed by the description of each instrument's training set as a Gaussian process. Only one playing style per instrument has been considered (i.e., no pizzicati, stacatti, or other articulations). It can be strongly assumed that such special playing styles would require additional specific models, since they heavily change the spectro–temporal behavior.

It should be noted that, while there have been several works dealing with an explicit modeling of the dependency of timbre on the fundamental frequency (f_0) or on the dynamics (see e.g., the work by Kitahara *et al.* [14] and Jensen's Instrument Definition Attributes model in [15]), that was not our goal here. Specifically, we address f_0-dependency from a different perspective: instead of seeking an f_0-dependent model, we accommodate the representation stage such that the modeling error produced by considering notes of different pitches is minimized. In other words, we seek prototypical spectro–temporal shapes that remain reasonably valid for a range of pitches. This allows avoiding a preliminary multipitch extraction stage in applications involving polyphonic mixtures, such as polyphonic instrument detection (Section VI) or source separation [13]. This important characteristic of the model will be discussed in detail in the next section.

In our experiments, we measure representativeness by the averaged distance in feature space between all samples belonging to the training database and all samples belonging to the test database. A high similarity between both data clouds (both in distance and in shape) indicates that the model has managed to capture essential and representative features of the instrument. The significance of such a measure, like in many other pattern recognition tasks, will benefit from a good-quality and well-populated database.

B. Compactness

Compactness refers to the ability to include as much information (variance, entropy) in models as simple as possible. It does not only result in more efficient computation, storage and retrieval but, together with representativeness, implies that the model has captured the essential characteristics of the source. In [4], compactness was considered one of the goals, but no training was performed. MFCCs are highly compact but, again, inaccurate. This work will use PCA spectral basis decomposition to attain compactness. In such a context, the natural measure of compactness is the variance explained by the retained PCA eigenvalues.

C. Accuracy

Some applications require a high representation accuracy. As an example, in a polyphonic detection task, the purpose of the models is to serve as a template guiding the separate detection of the individual overlapping partials. The same is valid if the templates are used to generate a set of partial tracks for synthesis. Model accuracy is a demanding requirement that is not always necessary in classification or retrieval by similarity, where the goal is to extract global, discriminative features. Many approaches relying on sinusoidal modeling [3]–[6] are based on highly accurate spectral descriptions, but fail to fulfill either compactness or representativeness. The model used here relies on an accurate description of the spectral envelope by means of sinusoidal-modeling-based interpolation. In the present context, accuracy is measured by the averaged amplitude error between the original spectro–temporal envelope and the spectro–temporal envelope retrieved and reconstructed from the models.

III. Representation Stage

The aim of the representation stage is to produce a set of coefficients describing the individual training samples. The process of summarizing all the coefficients belonging to an instrument into a prototype subset representative of that particular instrument will be the goal of the prototyping stage.

A. Envelope Estimation Through Sinusoidal Modeling

The first step of the training consists in extracting the spectro–temporal envelope of each individual sound sample of the training database. For its effectiveness, simplicity, and flexibility, we chose the interpolation approach to envelope estimation. It consists in frame-wise selecting the prominent sinusoidal peaks extracted with sinusoidal modeling and defining a function between them by interpolation. Linear interpolation results in a piecewise linear envelope containing edges. In spite of its simplicity, it has proven adequate for several applications [16]. Cubic interpolation results in smoother curves, but is more computationally expensive.

Sinusoidal modeling [16], also called additive analysis, performs a frame-wise approximation of amplitude, frequency, and phase parameter triplets $\hat{s}_{pr} = (\hat{A}_{pr}, \hat{f}_{pr}, \hat{\theta}_{pr})$. Here, p is the partial index and r is the frame (time) index. Throughout this paper, logarithmic amplitudes will be used. The set of frequency points $\hat{f}_{pr}$ for all partials during a given number of frames is called *frequency support*. In this paper, the phases $\hat{\theta}_{pr}$ will be ignored.

To perform the frame-wise approximations $\hat{s}_{pr}$, sinusoidal modeling implements the consecutive stages of peak picking and partial tracking. A sinusoidal track is the trajectory described by the amplitudes and frequencies of a sinusoidal peak across consecutive frames. To denote a track $\mathbf{t}_t$, the following notation will be used: $\mathbf{t}_t = \{\hat{s}_{p_t r} | R_t^0 \leq r \leq R_t^L\}$, where p_t is

the partial index associated with the track and R_t^0 and R_t^L are, respectively, its first and last frames. These stages have two possible modes of operation: harmonic and inharmonic. The harmonic mode is used whenever the f_0 is known beforehand. It is more robust since the algorithm can guess that the partials will be positioned close to integer multiples of f_0, and also because the analysis parameters can be adapted accordingly. In this paper, harmonic sinusoidal modeling is used for the representation stage experiments (Section III-D) and for training the models for the classification and polyphonic detection applications (Sections V and VI). Inharmonic mode will be used when analyzing the mixtures for polyphonic instrument detection (Section VI). In harmonic mode, a Blackmann window of size $W = 5f_0$ and a hop size of $W/4$ were used, with a sampling rate of $f_s = 44.1$ kHz. In inharmonic mode, a Blackmann window of fixed size $W = 8192$ samples was used, with a hop size of 2048 samples and the same f_s.

Given a set of additive analysis parameters, the spectral envelope can finally be estimated by frame-wise interpolating the amplitudes $\hat{A}_{pr}$ at frequencies $\hat{f}_{pr}$ for $p = 1, \ldots, P_r$.

B. Spectral Basis Decomposition

Spectral basis decomposition [10] consists of performing a factorization of the form $\mathbf{X} = \mathbf{PY}$, where $\mathbf{X}$ is the data matrix containing a time–frequency (t-f) representation with K spectral bands and R time frames (usually $R \gg K$), $\mathbf{P}$ is the transformation basis whose columns $\mathbf{p}_i$ are the basis vectors, and $\mathbf{Y}$ is the projected coefficient matrix. If the data matrix is in *temporal orientation* (i.e., it is a $R \times K$ matrix $\mathbf{X}(r,k)$), a temporal $R \times R$ basis matrix $\mathbf{P}$ is obtained. If it is in *spectral orientation* ($K \times R$ matrix $\mathbf{X}(k,r)$), the result is a spectral basis of size $K \times K$. Having as goal the extraction of spectral features, the latter case is of interest here.

PCA realizes such a factorization under the constraint that the variance is concentrated as compactly as possible in a few of the transformed dimensions. It meets our need for compactness and was thus chosen for the basis decomposition stage. After centering (i.e., removing the mean) and whitening (i.e., normalizing the dimensions by their respective variances), the final projection of reduced dimensionality $D < K$ is given by

$$\mathbf{Y}_\rho = \mathbf{\Lambda}_\rho^{-1/2} \mathbf{P}_\rho^T (\mathbf{X} - E\{\mathbf{X}\}) \tag{1}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ and λ_d are the D largest eigenvalues of the covariance matrix $\mathbf{\Sigma}_\mathbf{X}$, whose corresponding eigenvectors are the columns of $\mathbf{P}_\rho$. The ρ subscript denotes dimensionality reduction and indicates the mentioned eigenvalue and eigenvector selection. The truncated model reconstruction would then yield the approximation

$$\hat{\mathbf{X}} = \mathbf{P}_\rho \mathbf{\Lambda}_\rho^{1/2} \mathbf{Y}_\rho + E\{\mathbf{X}\}. \tag{2}$$

C. Frequency Alignment

To approach the design criterion of representativeness we need to consider notes of different instrument exemplars, dynamics and pitches into the training set. More specifically, we concatenate in time the spectro–temporal envelopes of different exemplars, dynamics and pitches into a single input data matrix, and extract the common PCA bases. However, since the spectro–temporal envelope can greatly vary between pitches, concatenating the whole pitch range of a given instrument can produce excessively flat common bases, thus resulting in a poor timbral characterization. On the other hand, it can be expected that the changes in envelope shape will be minor for notes that are consecutive in the chromatic scale. It was thus necessary to find an appropriate trade-off and choose a moderate range of consecutive semitones for the training. After preliminary tests, a range between one and two octaves was deemed appropriate for our purposes.

In Casey's original proposal [10] and related works, basis decomposition is performed upon the *short-time Fourier transform* (STFT) spectrogram, with fixed frequency positions given by the regular frequency-domain sampling of the DFT. In contrast, here the decomposition is performed on the spectro–temporal envelope, which we defined as a set of partials with varying frequencies plus an interpolation function. Thus, when concatenating notes of different pitches, the arrangement into the data matrix is less straightforward.

The simplest solution is to ignore interpolation and use directly the sinusoidal amplitude parameters as the elements of the data matrix. In this case, the number of partials to be extracted for each note is fixed and the partial index p is used as frequency index, obtaining $\mathbf{X}(p,r)$ with elements $x_{pr} = \hat{A}_{pr}$. We will refer to this as *Partial Indexing* (PI).

The PI approach is simple and appropriate in some contexts ([3], [4]), but when concatenating notes of different pitches, several additional considerations have to be taken into account. These concern the formant- or resonance-like spectral features, that can either lie at the same frequency, irrespective of the pitch, or be correlated with the fundamental frequency. In this paper, the former will be referred to as f_0-invariant features, and the latter as f_0-correlated features. When concatenating notes of different pitches for the training, their frequency support will change logarithmically. If the PI arrangement is used, this has the effect of misaligning the f_0-invariant features in the data matrix. On the contrary, possible features that follow the logarithmic evolution of f_0 will become aligned.

An alternative to PI is to interpolate between partial amplitudes to approximate the spectral envelope, and to sample the resulting function at a regular grid of G points uniformly spaced within a given frequency range $f_g = (f_{\max}/G)g$. The spectral matrix is now defined by $\mathbf{X}(g,r)$, where $g = 1, \ldots, G$ is the grid index and r the frame index. Its elements will be denoted by $x_{gr} = A_{gr}$. This approach shall be referred to as Envelope Interpolation (EI). This strategy does not change formant alignments, but introduces an interpolation error.

In general, frequency alignment is desirable for the present modeling approach because, if subsequent training samples share more common characteristics, prototype spectral shapes will be learned more effectively. In other words, the data matrix will be more correlated and thus PCA will be able to obtain a better compression. In this context, the question arises of which one of the alternative preprocessing methods—PI (aligning f_0-correlated features) or EI (aligning f_0-invariant

features)—is more appropriate. In order to answer that question, the experiments outlined in the next section were performed.

D. Evaluation of the Representation Stage

A cross-validated experimental framework was implemented to test the validity of the representation stage and to evaluate the influence of the PI, linear EI, and cubic EI methods. Here, some experimental results will be presented. Further results and evaluation details can be found in [11].

The used samples are part of the RWC database [17]. One octave (C4 to B4) of two exemplars from each instrument type was trained. As test set, the same octave from a third exemplar from the database was used. All sound samples belonging to each set were subjected to sinusoidal modeling, concatenated in time and arranged into a data matrix using either the PI or the EI method. For the PI method, $P = 20$ partials were extracted. For the EI method, $f_{\max}$ was set as the frequency of the 20th partial of the highest note present in the database, so that both methods span the same maximum frequency range, and a frequency grid of $G = 40$ points was defined.

As mentioned earlier, representativeness was measured in terms of the global distance between the training and testing coefficients. We avoid probabilistic distances that rely on the assumption of a certain probability distribution, which would yield inaccurate results for data not matching that distribution. Instead, average point-to-point distances were used. In particular, the averaged minimum distance between point clouds, normalized by the number of dimensions, was computed:

$$\Delta_D(\omega_1,\omega_2) = \frac{1}{D}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}\min_{\mathbf{y}_j\in\omega_2}\{d(\mathbf{y}_i,\mathbf{y}_j)\} + \frac{1}{n_2}\sum_{j=1}^{n_2}\min_{\mathbf{y}_i\in\omega_1}\{d(\mathbf{y}_i,\mathbf{y}_j)\}\right\} \quad (3)$$

where ω_1 and ω_2 denote the two clusters, n_1 and n_2 are the number of points in each cluster, $\mathbf{y}_i$ are the PCA coefficients, and $d(\cdot)$ denotes the Mahalanobis distance

$$d(\mathbf{y}_0,\mathbf{y}_1) = \sqrt{(\mathbf{y}_0-\mathbf{y}_1)^T\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}(\mathbf{y}_0-\mathbf{y}_1)} \quad (4)$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}}$ is the global covariance matrix.

Compactness was measured by the explained variance (EV) of the PCA eigenvalues λ_i

$$EV(D) = 100\frac{\sum_i^D \lambda_i}{\sum_i^K \lambda_i}. \quad (5)$$

Accuracy was defined in terms of the reconstruction error between the truncated t-f reconstruction of (2) and the original data matrix. To that end, the *relative spectral error* (RSE) [18] was measured

$$\text{RSE} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{\sum_{p=1}^{P_r}(A_{pr}-\tilde{A}_{pr})^2}{\sum_{p=1}^{P_r}A_{pr}^2}} \quad (6)$$

where $\tilde{A}_{pr}$ is the reconstructed amplitude at support point (p,r) and R is the total number of frames. In order to measure the RSE, the envelopes must be compared at the points of the original frequency support. This means that, in the case of the EI method, the back-projected envelopes must be reinterpolated using the original frequency information. As a consequence, the RSE accounts not only for the errors introduced by the dimension reduction, but also for the interpolation error itself, inherent to EI.

Fig. 1 shows the results for the particular cases of the piano (as an example of a non-sustained instrument) and of the violin (as an example of a sustained instrument). Fig. 1(a) and (d) demonstrates that EI has managed to reduce the distance between training and test sets in comparison to PI. Fig. 1(b) and (e) shows that EI achieves a higher compression than PI for low dimensionalities. A 95% of variance is achieved already for $D = 7$ in the case of the piano and of $D = 8$ in the case of the violin. Finally, Fig. 1(c) and (f) demonstrates that EI also reduces the reconstruction error in the low-dimensionality range. The RSE curves for PI and EI must always cross because of the zero reconstruction error of PI with $D = K$ and of the reinterpolation error of EI. In general, cubic and linear interpolation performed very similarly.

IV. Prototyping Stage

In model space, the projected coefficients must be reduced into a set of generic models representing the classes. Common MIR methods include *Gaussian mixture models* (GMMs) and *hidden Markov models* (HMMs). Both are based on clustering the transformed coefficients into a set of densities, either static (GMM) or linked by transition probabilities (HMM). The evolution of the envelope in time is either completely ignored in the former case, or approximated as a sequence of states in the latter. For a higher degree of accuracy, however, the time variation of the envelope should be modeled in a more faithful manner, since it plays an important role when characterizing timbre. Therefore, the choice here was to always keep the sequence ordering of the coefficients, and to represent each class as a trajectory rather than as a cluster. For each class, all training trajectories are to be collapsed into a single *prototype curve* representing that instrument.

To that end, the following steps are taken. Let $\mathcal{Y}_{si}$ denote the coefficient trajectory in model space corresponding to training sample s (with $s = 1,\ldots,S_i$) belonging to instrument i (with $i = 1,\ldots,I$), of length R_{si} frames: $\mathcal{Y}_{si} = (\mathbf{y}_{si1},\mathbf{y}_{si2},\ldots,\mathbf{y}_{siR_{si}})$. First, all trajectories are interpolated in time using the underlying time scales in order to obtain the same number of points. In particular, the longest trajectory, of length $R_{\max}$ is selected and all the other ones are interpolated so that they have that length. In the following, the sign $\breve{}$ will denote interpolation

$$\breve{\mathcal{Y}}_{si} = \text{interp}_{R_{\max}}\{\mathcal{Y}_{si}\} = (\breve{\mathbf{y}}_{si1},\breve{\mathbf{y}}_{si1},\ldots,\breve{\mathbf{y}}_{siR_{\max}}). \quad (7)$$

Then, each point in the resulting prototype curve for instrument i, of length $R_{\max}$, denoted by $\mathcal{C}_i = (\mathbf{p}_{i1},\mathbf{p}_{i2},\ldots,\mathbf{p}_{iR_{\max}})$, is considered to be a D-dimensional

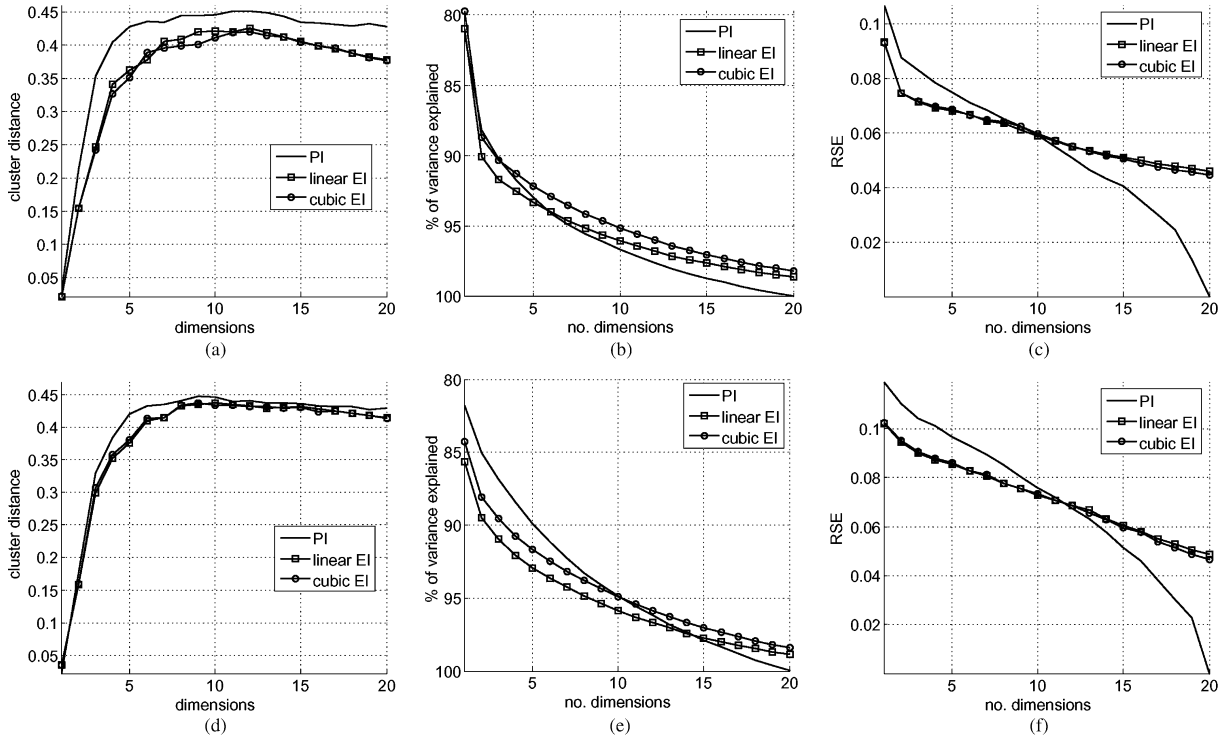

Fig. 1. Evaluation of the representation stage: results for the piano [Fig. 1(a)–(c)] and the violin [Fig. 1(d)–(f)]. Note that the y-axes of the explained variance graphs have been inverted so that for all measures, "better" means downwards. (a) Piano: train/test cluster distance (representativeness criterion). (b) Piano: explained variance (compactness criterion). (c) Piano: RSE (accuracy criterion). (d) Violin: train/test cluster distance (representativeness criterion). (e) Violin: explained variance (compactness criterion). (f) Violin: RSE (accuracy criterion).

Gaussian random variable $\mathbf{p}_{ir} \sim N(\boldsymbol{\mu}_{ir}, \boldsymbol{\Sigma}_{ir})$ with empirical mean

$$\boldsymbol{\mu}_{ir} = \frac{1}{S_i} \sum_{s=1}^{S_i} \breve{\mathbf{y}}_{sir} \tag{8}$$

and empirical covariance matrix $\boldsymbol{\Sigma}_{ir}$, which for simplicity will be assumed diagonal, where $\boldsymbol{\sigma}_{ir}^2 = \mathrm{diag}(\boldsymbol{\Sigma}_{ir})$ is given by

$$\boldsymbol{\sigma}_{ir}^2 = \frac{1}{S_i - 1} \sum_{s=1}^{S_i} (\breve{\mathbf{y}}_{sir} - \boldsymbol{\mu}_{ir})^2. \tag{9}$$

The obtained prototype curve is thus a discrete-temporal sequence of Gaussian distributions in which means and covariances change over time. This can be interpreted as a D-dimensional, nonstationary GP parametrized by r (in other words, a collection of Gaussian distributions indexed by r)

$$\mathcal{C}_i \sim GP\left(\boldsymbol{\mu}_i(r), \boldsymbol{\Sigma}_i(r)\right). \tag{10}$$

Fig. 2 shows an example set of mean prototype curves corresponding to a training set of five classes: piano, clarinet, oboe, violin, and trumpet, in the first three dimensions of the PCA space. The database consists of three dynamic levels (piano, mezzoforte and forte) of two to three exemplars of each instrument type, covering a range of one octave between C4 and B4. This makes a total of 423 sound files. Here, only the mean curves formed by the values $\boldsymbol{\mu}_{ir}$ are plotted. It must be noted, however,

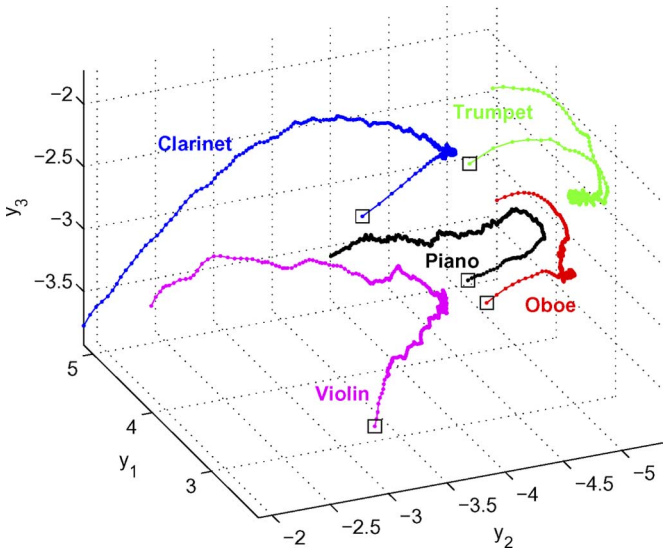

Fig. 2. Prototype curves in the first three dimensions of model space corresponding to a five-class training database of 423 sound samples, preprocessed using linear envelope interpolation. The starting points are denoted by squares.

that each curve has an "influence area" around it as determined by their time-varying covariances.

Note that the time normalization defined by (7) implies that all sections of the ADSR temporal envelope are interpolated with the same density. This might be disadvantageous for sustained sounds, in which the length of the sustained part is arbitrary. For example, comparing a short violin note with a long

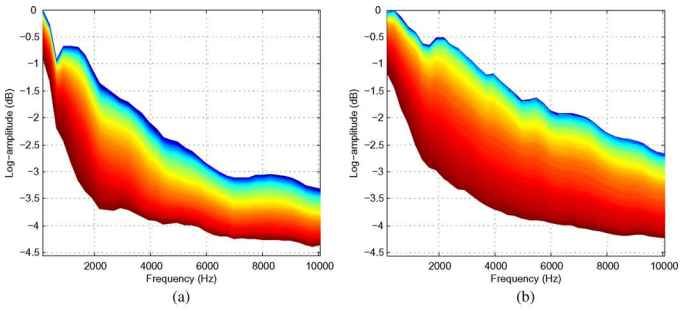

Fig. 3. Frequency profile of the prototype envelopes corresponding to two of the curves in Fig. 2. (a) Clarinet. (b) Violin.

violin note will result in the attack part of the first being excessively stretched and matched with the beginning of the sustained part of the second. The experiments in the next section will help to assess the influence of this simplification.

When projected back to the t-f domain, each prototype trajectory will correspond to a *prototype envelope* consisting of a mean surface and a variance surface, which will be denoted by $\mathbf{M}_i(g,r)$ and $\mathbf{V}_i(g,r)$, respectively, where $g = 1,\ldots,G$ denotes the sample points of the frequency grid and $r = 1,\ldots R_{\max}$ for all the models. Each D-dimensional mean point $\boldsymbol{\mu}_{ir}$ in model space will correspond to a G-dimensional vector of mean amplitudes constituting a time frame of the reconstructed spectro–temporal envelope. Undoing the effects of whitening and centering, the reconstructed means are

$$\hat{\boldsymbol{\mu}}_{ir} = \mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\mu}_{ir} + E\{\mathbf{X}\} \tag{11}$$

and the corresponding variance vector

$$\hat{\boldsymbol{\sigma}}_{ir}^2 = \operatorname{diag}\left(\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\Sigma}_{ir} \left(\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2}\right)^T\right) \tag{12}$$

both of G dimensions, which form the columns of $\mathbf{M}_i(g,r)$ and $\mathbf{V}_i(g,r)$, respectively.

Analogously as in model space, a prototype envelope can be interpreted as a GP, but in a slightly different sense. Instead of being multidimensional, the GP is unidimensional (in amplitude), but parametrized with means and variances varying in the two-dimensional t-f plane. Such prototype envelopes are intended to be used as t-f templates that can be interpolated at any desired t-f point. Thus, the probabilistic parametrization can be considered continuous, and therefore the indices t and f will be used, instead of their discrete counterparts r and k. The prototype envelopes can then be denoted by

$$\mathcal{E}_i \sim GP\left(\mu_i(t,f), \sigma_i^2(t,f)\right). \tag{13}$$

Fig. 3 shows the frequency-amplitude projection of the mean prototype envelopes corresponding to the clarinet and violin prototype curves of Fig. 2. The shades or colors denote the different time frames. Note the different formant-like features in the mid-low frequency areas. On the figures, several prominent formants are visible, constituting the characteristic averaged spectral shapes of the respective instruments. Again, only the mean surfaces are represented, but variance influence areas are also contained in the model.

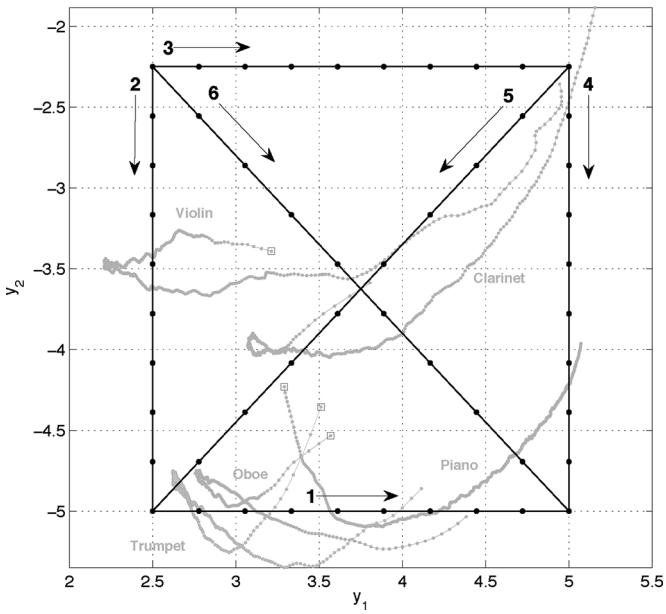

Fig. 4. Envelope evaluation points and traces for Fig. 5.

The average resonances found with the modeling procedure presented here are consistent with previous acoustical studies. As an example, the frequency profile of the clarinet [Fig. 3(a)] shows a spectral hill that corresponds to the first measured formant, which has its maximum between 1500 and 1700 Hz [19]. Also, the bump around 2000 Hz on the violin profile [Fig. 3(b)] can be identified as the "bridge hill" observed by several authors [20], produced by renonances of the bridge.

Depending on the application, it can be more convenient to perform further processing on the reduced-dimensional PCA space or back in the t-f domain. When classifying individual notes, such as introduced in the next section, a distance measure between unknown trajectories and the prototype curves in PCA space has proven a successful approach. However, in applications where the signals to be analyzed are mixtures of notes, such as polyphonic instrument recognition (Section VI), the envelopes to be compared to the models can contain regions of unresolved overlapping partials or outliers, which can introduce important interpolation errors when adapted to the frequency grid needed for projection onto the bases. In those cases, working in the t-f domain will be more convenient.

To gain further insight into the meaning of the timbre axes, the spectral envelope was evaluated and plotted at different points of the space. In benefit of clarity, a two-dimensional projection of the space onto the first two dimensions was performed, and several evaluation "traces" were chosen as indicated by the numbered straight lines on Fig. 4. Fig. 5 represents the evolution of the spectral envelope alongside the traces defined in Fig. 4, sampled uniformly at ten different points. The thicker envelopes correspond to the starting points on the traces, which are then followed in the direction marked by the arrows. Each envelope representation in Fig. 5 corresponds to a sample point as indicated by the dots on the traces of Fig. 4. Traces 1 to 4 are parallel to the axes, thus illustrating the latter's individual influence.

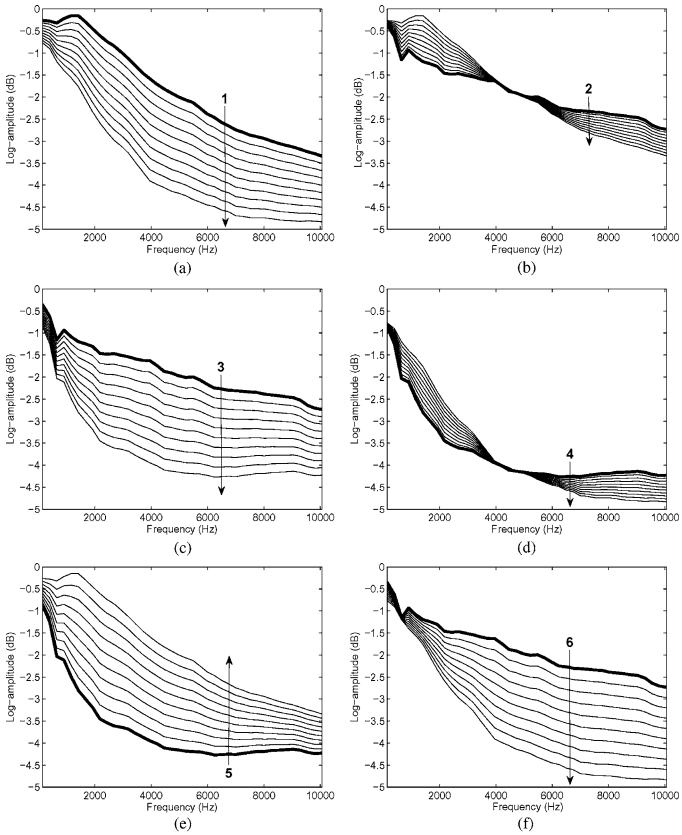

Fig. 5. Evolution of the spectral envelope alongside the traces in Fig. 4. (a) Trace 1. (b) Trace 2. (c) Trace 3. (d) Trace 4. (e) Trace 5. (f) Trace 6.

From traces 1 and 3 it can be asserted that the first dimension (axis $\mathbf{y}_1$) mostly affects the overall energy and slope of the spectral envelope. Such slope can be approximated as the slope of the straight line one would obtain performing linear regression on the spectral envelope. Along traces 2 and 4 (axis $\mathbf{y}_2$), the envelope has the clear behavior of changing the ratio between low-frequency and high-frequency spectral content. For decreasing values of $\mathbf{y}_2$, high-frequency contents decreases and low-frequency contents increases, producing a rotation of the spectral shape around a pivoting point at approximately 4000 Hz. Traces 5 and 6 travel alongside the diagonals and represent thus a combination of both behaviors.

V. Application to Sample Classification

In the previous sections, it has been shown that the proposed modeling approach is successful in capturing timbral features of individual instruments. For many applications, however, dissimilarity between different models is also desired. Therefore, we evaluate the performance of the model in a classification context involving solo instrumental samples. Such a classification task is a popular application [21], aimed at the efficient managing and searching of sample databases.

We perform such a classification task extracting a common basis from the whole training set, computing one prototype curve for each class and measuring the distance between an input curve and each prototype curve. Like for prototyping, the curves must have the same number of points, and thus the input curve must be interpolated with the number of points of the densest prototype curve, of length $R_{\max}$. The distance between an interpolated unknown curve $\breve{\mathcal{U}}$ and the ith prototype curve $\mathcal{C}_i$ is defined here as the average Euclidean distance between their mean points

$$d(\breve{\mathcal{U}},\mathcal{C}_i) = \frac{1}{R_{\max}} \sum_{r=1}^{R_{\max}} \sqrt{\sum_{k=1}^{D} (\breve{u}_{rk} - \mu_{irk})^2}. \qquad (14)$$

For the experiments, another subset of the same five classes (piano, clarinet, oboe, violin, and trumpet) was defined, again from the RWC database [17], each containing all notes present in the database for a range of two octaves (C4 to B5), in all different dynamics (forte, mezzoforte, and piano) and normal playing style, played by two to three instrument exemplars of each instrument type. This makes a total of 1098 individual note files, all sampled at 44.1 kHz. For each method and each number of dimensions, the experiments were iterated using ten-fold random cross-validation. The same parameters as in the representation stage evaluations were used: $P = 20$ partials for PI, and a frequency grid of $G = 40$ points for EI.

The obtained classification accuracy curves are shown in Fig. 6. Note that each data point is the result of averaging the ten folds of cross-validation. The experiments were iterated up to a dimensionality of $D = 20$, which is the full dimensionality in the PI case. The best classification results are given in Table I. With PI, a maximal accuracy of 74.9% was obtained. This was outperformed by around 20 percent units when using the EI approach, obtaining 94.9% for linear interpolation and 94.6% for cubic interpolation.

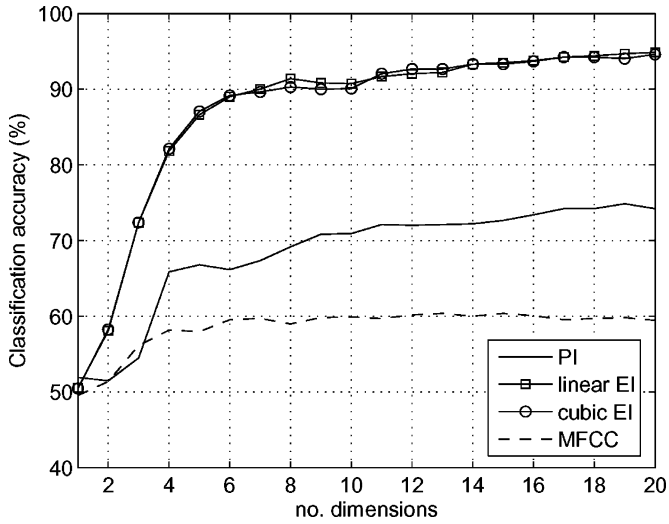

Fig. 6. Isolated sample classification results: Averaged classification accuracy.

TABLE I
Isolated Sample Classification: Maximum Averaged Accuracy and Standard Deviation (STD)

Representation	Accuracy	STD
PI	74.9%	$\pm$ 2.8%
Linear EI	94.9%	$\pm$ 2.1%
Cubic EI	94.6%	$\pm$ 2.7%
MFCC	60.4%	$\pm$ 4.1%

TABLE II
CONFUSION MATRIX FOR THE MAXIMUM ACCURACY OBTAINED WITH PI ($D = 19$)

	found				
real	**p**	**c**	**o**	**v**	**t**
p	**81.40**	0.47	2.79	4.19	11.16
c	6.07	**86.45**	5.14	1.40	0.93
o	1.40	4.20	**55.94**	12.59	25.87
v	1.87	0.80	11.50	**77.54**	8.29
t	4.86	0.69	18.75	15.97	**59.72**

TABLE III
CONFUSION MATRIX FOR THE MAXIMUM ACCURACY OBTAINED WITH LINEAR EI ($D = 20$)

	found				
real	**p**	**c**	**o**	**v**	**t**
p	**95.81**	1.40	0.47	0	2.33
c	1.40	**92.52**	5.14	0.93	0
o	0	2.10	**95.10**	2.10	0.70
v	1.07	0.53	0	**95.45**	2.94
t	0	0	0	3.47	**96.53**

To assess instrument-wise performances, two confusion matrices are shown in Table II (for the best performance achieved with PI) and in Table III (for the best performance achieved with linear EI). The initials on the matrices denote: piano $(\mathbf{p})$, clarinet $(\mathbf{c})$, oboe $(\mathbf{o})$, violin $(\mathbf{v})$, and trumpet $(\mathbf{t})$. All individual performances are better with EI than with PI, but the difference in performances between instruments show a completely different behavior. In particular, note that the clarinet obtained both the best performance of all instruments with PI (86.45%) and the worst performance with EI (92.52%). Recall that PI aligns f_0-correlated features and EI aligns f_0-invariant features. The spectrum of the clarinet has the particularity that the odd partials are predominant. When estimating the spectral envelope, this produces important inter-peak valleys that are, in effect, f_0-correlated features, which are thus kept aligned by PI. It follows that for the clarinet, f_0-correlated features predominate over static formants, and the contrary is valid for the other four considered instruments.

Another conclusion that can be drawn from the confusion matrices is that the piano, the only non-sustained instrument considered, did not perform significantly better than the sustained instruments. This suggests that the simplicity of the time normalization process (which, as mentioned above, is uniform in all phases of the ADSR envelope) has a relatively small effect on the performance, at least for this application scenario.

For comparison, the representation stage was replaced with a standard implementation of MFCCs. Note that MFCCs follow a similar succession of stages than our approach (envelope estimation followed by compression), but they are expected to perform worse because the estimation stage delivers a rougher envelope (based on fixed frequency bands), and the DCT produces only a suboptimal decorrelation. The MFCC coefficients were subjected to GP prototyping, and a set of MFCC prototype curves was thus created. The results are again shown in Fig. 6 and Table I. The highest achieved classification rate was only of 60.4% (with $D = 13$).

The obtained accuracies are comparable to those of other systems from the literature. A review of approaches can be found in [21]. As examples of methods with a similar number of classes, we can cite the work by Brown *et al.* [22], based on a Naïve Bayes Classifier and attaining a classification accuracy of 84% for four instrument classes, the work by Kaminskyj and Materka [23], based on a feedforward Neural Network and reaching an accuracy of 97% with four classes, and the work by Livshin and Rodet [24], where a k-Nearest Neighbors algorithm attains a performance of 90.53% for ten classes, interestingly using only the sinusoidal part of the signals.

VI. APPLICATION TO POLYPHONIC INSTRUMENT RECOGNITION

Isolated sample classification, as presented in the previous section, is useful for applications involving sound databases intended for professional musicians or sound engineers. A broader group of users will potentially be more interested in analysis methods that can handle more realistic and representative musical data, such as full musical tracks containing mixtures of different instruments. While far from being applicable to a wide range of instrumentations and production styles, current methods aiming at the detection of instruments in a polyphonic mixture aim towards that ideal goal of generalized *auditory scene analysis*.

Thus, a second, more demanding, analysis application was selected to test the appropriateness of the models. In particular, we address the detection of the occurrence of instruments in single-channel mixtures. The main difficulty of such a task, compared to the single-voice case, arises from the fact that the observed partials correspond to overlapping notes of different timbres, thus not purely following the predicted t-f template approximations. In such a case, it will be more convenient to work in the t-f domain. Also, since the notes have to be compared one-by-one to the templates, they must first be located in the audio stream by means of an onset detection stage.

Past approaches towards polyphonic timbre detection typically either consider the mixture as a whole [25] or attempt to separate the constituent sources with prior knowledge related to pitch [26]. The method proposed here is based on the grouping and partial separation of sinusoidal components, but has the particularity that no harmonicity is assumed, since classification is solely based on the amplitude of the partials and their evolution in time. As a result, no pitch-related *a priori* information or preliminary multipitch detection step are needed. Also, it has the potential to detect highly inharmonic instruments, as well as single-instrument chords.

The mixture is first subjected to inharmonic sinusoidal extraction, followed by a simple onset detection, based on counting the tracks born at a particular frame. Then, all tracks $\mathbf{t}_t$ having its first frame close to a given onset location L_o^{on} are grouped into the set $\mathcal{T}_o$. A track belonging to this set can be either non-overlapping (if it corresponds to a new partial not present in the previous track group $\mathcal{T}_{o-1}$) or overlapping with a partial of the previous track (if its mean frequency is close, within a narrow margin, to the mean frequency of a partial from $\mathcal{T}_{o-1}$). Due to the fact that no harmonicity is assumed, it cannot be decided from the temporal information alone if a partial overlaps with a partial belonging to a note or chord having the onset within the same analysis window or not. This is the origin of the current onset separability constraint on the mixture, which hinders

two notes of being individually detected if their onsets are synchronous. For each track set $\mathcal{T}_o$, a reduced set $\mathcal{T}'_o$ was created by eliminating all the overlapping tracks in order to facilitate the matching with the t-f templates.

Then, the classification module matches each one of the track groups $\mathcal{T}'_o$ with each one of the prototype envelopes, and selects the instrument corresponding to the highest match. To that end, envelope similarity was first defined as the following optimization problem, based on the total Euclidean distance between amplitudes:

$$d\left(\mathcal{T}'_o, \tilde{\mathbf{M}}_{io}\right) = \min_{\alpha,N}\left\{\sum_{t\in\mathcal{T}'_o}\sum_{r=1}^{R_t}\left|A_{tr}^N + \alpha - \mathbf{M}_i\left(f_{tr}^N\right)\right|\right\} \tag{15}$$

where R_t is the number of frames in track $\mathbf{t}_t \in \mathcal{T}'_o$, α is an amplitude scaling parameter, and A_{tr}^N and f_{tr}^N denote the amplitude and frequency values for a track belonging to a group that has been stretched so that its last frame is N. The optimization based on amplitude scaling and track stretching is necessary to avoid the overall gain and note length having an effect on the measure. In order to perform the evaluation $\tilde{\mathbf{M}}_{io} = \mathbf{M}_i(\mathbf{F}_o)$ at the frequency support $\mathbf{F}_o$, for each data point the model frames closest in time to the input frames are chosen, and the corresponding values for the mean surface are linearly interpolated from neighboring data points.

To also take into account the variance of the models, a corresponding likelihood-based problem was defined as

$$L\left(\mathcal{T}'_o|\boldsymbol{\theta}_i\right) = \max_{\alpha,N}\left\{\prod_{t\in\mathcal{T}'_o}\prod_{r=1}^{R_t} p\left(A_{tr}^N + \alpha|\mathbf{M}_i\left(f_{tr}^N\right), \mathbf{V}_i\left(f_{tr}^N\right)\right)\right\}$$

where $p(x)$ denotes a unidimensional Gaussian distribution.

The single-channel mixtures used for the experiments were generated by linearly mixing samples of isolated notes from the RWC database [17] with separated onsets. Two different types of mixtures were generated: simple mixtures consisting of one single note per instrument and sequences of more than one note per instrument. A total of 100 mixtures were generated. The training database consists of the five instruments mentioned before, covering two octaves (C4-B5), and contains 1098 samples in total. For the evaluation, the database was partitioned into separate training (66% of the database) and test sets (33% of the database). The training set contains samples from one or two exemplars, and the test set contains samples from a further instrument exemplar. More precisely, this means that 66% of the samples were used to train the models, and the remaining 33% were used to generate the 100 mixtures.

The classification measure chosen was the note-by-note accuracy, i.e., the percentage of individual notes with correctly detected onsets that were correctly classified. Table IV shows the results. The likelihood approach worked better than the Euclidean distance in all cases, showing the advantage of taking into account the model variances. Note that these experiments had the goal of testing the performance of the spectral matching module alone, and do not take into account the performance of the onset detection stage.

TABLE IV
POLYPHONIC INSTRUMENT RECOGNITION ACCURACY (%)

	Simple mixtures			Sequences	
Polyphony	2	3	4	2	3
Euclidean distance	68.48	52.25	41.28	64.66	50.64
Likelihood	73.15	55.56	54.18	63.68	56.40

While a fully significant performance comparison with other systems is difficult due to the lack of a common database and evaluation procedure, we can cite the previous work [27], which used the same timbre modeling procedure and a similar database (linear mixtures from the RWC samples, albeit six instruments are considered, instead of five). The onset detection stage and subsequent track grouping heuristics, used here, are replaced in that work by a graph partitioning algorithm. The note-by-note classification accuracy was of 65% with two voices, 50% with three voices, and 33% with four voices.

VII. CONCLUSION AND FUTURE WORK

The task of developing a computational model representing the dynamic spectral characteristics of musical instruments has been addressed. The development criteria were chosen and combined so that such models can be used in a wide range of MIR applications. To that end, techniques aiming at compactness (PCA), accuracy of the envelope description (sinusoidal modeling and spectral interpolation) and statistical learning (training and prototyping via Gaussian Processes) were combined into a single framework. The obtained features were modeled as prototype curves in a reduced-dimensional space, which can be projected back into the t-f domain to yield a set of t-f templates called prototype envelopes.

We placed emphasis on the evaluation of the frequency misalignment effects that occur when notes of different pitches are used in the same training database. To that end, data preprocessing methods based on PI and EI were compared in terms of explained variance, reconstruction error and training/test cluster similarity, with EI being better in most cases for low and moderate dimensionalities of up to around 1/4 of the full dimensionality. It follows that the interpolation error introduced by EI was compensated by the gain in correlation in the training data.

The developed timbre modeling approach was first evaluated for the task of classification of isolated instrument samples, consisting in projecting the spectro–temporal envelope of unknown samples into the PCA space and comparing an average distance between the resulting trajectory and each one of the prototype curves. This approach reached a classification accuracy of 94.9% with a database of five classes, and outperformed using MFCCs for the representation stage by 34 percent units.

As a second, more demanding application, detection of instruments in monaural polyphonic mixtures was tested. Such a task focused on the analysis of the amplitude evolution of the partials, matching it with the pre-trained t-f templates. The obtained results show the viability of such a method without requiring multipitch estimation. Accuracies of 73.15% for two voices, 55.56% for three voices, and 54.18% for four voices were obtained. To overcome the current constraint on the separability of the onsets, the design of more robust spectro–temporal similarity measures will be needed.

A possibility for further research is to separate prototype curves into segments of the ADSR envelope. This can allow three enhancements: first, different statistical models can be more appropriate to describe different segments of the temporal envelope. Second, such a multi-model description can allow a more abstract parametrization at a morphological level, turning timbre description into the description of geometrical relationships between objects, and finally, it would allow treating the segments differently when performing time interpolation for the curve averaging, and time stretching for maximum-likelihood timbre matching, thus avoiding stretching the attack time in the same degree than the sustained part.

It is also possible to envision sound-transformation or synthesis applications involving the generation of dynamic spectral envelope shapes by navigating through the timbre space, either by a given set of deterministic functions or by user interaction. If combined with multi-model extensions of the prototyping stage, like the ones mentioned above, this could allow approaches to morphological or object-based sound synthesis. It can be strongly assumed that for such possible future applications involving sound resynthesis, perceptual aspects (such as auditory frequency warpings or masking effects) will have to be explicitly considered as part of the models in order to obtain a satisfactory sound quality.

The presented modeling approach is valid for sounds with predominant partials, both harmonic or inharmonic, and in polyphonic scenarios it can handle linear mixtures. Thus, a final evident research goal would be to extend the applicability of the models to perform with more realistic signals of higher polyphonies, different mixing model assumptions (e.g., delayed or convolutive models due to reverberation) and real recordings that can contain, e.g., different levels of between-note articulations (transients), playing modes, or noisy or percussive sounds.

REFERENCES

[1] J. F. Schouten, "The perception of timbre," in *Proc. 6th Int. Congr. Acoust.*, Tokyo, Japan, 1968, vol. GP-6-2, pp. 35–44.

[2] J. M. Grey, "Multidimensional perceptual scaling of musical timbre," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, 1977.

[3] C. Hourdin, G. Charbonneau, and T. Moussa, "A multidimensional scaling analysis of musical instruments' time-varying spectra," *Comput. Music J.*, vol. 21, no. 2, pp. 40–55, 1997.

[4] G. Sandell and W. Martens, "Perceptual evaluation of principal-component-based synthesis of musical timbres," *J. Audio Eng. Soc.*, vol. 43, no. 12, pp. 1013–1028, Dec. 1995.

[5] G. D. Poli and P. Prandoni, "Sonological models for timbre characterization," *J. New Music Research*, vol. 26, no. , pp. 170–197, 1997.

[6] M. Loureiro, H. de Paula, and H. Yehia, "Timbre classification of a single musical instrument," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[7] K. Jensen, "The timbre model," in *Proc. Workshop Current Research Directions Comput. Music*, Barcelona, Spain, 2001.

[8] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representations," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[10] M. Casey, "Sound classification and similarity tools," in *Introduction to MPEG-7*, B. S. Manjunath and T. Sikora, Eds. New York: Wiley, 2002.

[11] J. J. Burred, A. Röbel, and X. Rodet, "An accurate timbre model for musical instruments and its application to classification," in *Proc. Workshop Learn. Semantics Audio Signals (LSAS)*, Athens, Greece, Dec. 2006.

[12] J. J. Burred, A. Röbel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2010, pp. 173–176.

[13] J. J. Burred and T. Sikora, "Monaural source separation from musical mixtures based on time–frequency timbre models," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Vienna, Austria, Sep. 2007.

[14] T. Kitahara, M. Goto, and H. G. Okuno, "Musical instrument identification based on f0-dependent multivariate normal distribution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, China, 2003, pp. 421–424.

[15] K. Jensen, "Timbre models of musical sounds," Ph.D. dissertation, Dept. Comput. Sci., Univ. of Copenhagen, Copenhagen, Denmark, 1999.

[16] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral processing," in *DAFX—Digital Audio Effects*, U. Zölzer, Ed. New York: Wiley, 2002, pp. 373–438.

[17] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, 2003.

[18] A. Horner, "A simplified wavetable matching method using combinatorial basis spectra selection," *J. Audio Eng. Soc.*, vol. 49, no. 11, pp. 1060–1066, 2001.

[19] J. Backus, *The Acoustical Foundations of Music*. New York: Norton, 1977.

[20] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. New York: Springer, 1998.

[21] P. Herrera, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *J. New Music Res.*, vol. 32, no. 1, pp. 3–21, 2003.

[22] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1064–1072, 2001.

[23] I. Kaminskij and A. Materka, "Automatic source identification of monophonic musical instrument sounds," in *Proc. IEEE Int. Conf. Neural Netw.*, Perth, WA, Australia, 1995, pp. 189–194.

[24] A. Livshin and X. Rodet, "The significance of the non-harmonic "noise" versus the harmonic series for musical instrument recognition," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, 2006.

[25] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Philadelphia, PA, 2005, pp. 245–248.

[26] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, no. 4, pp. 712–729, Apr. 2004.

[27] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, Sep. 2007.

Juan José Burred (M'09) received the Telecommunication Engineering degree from the Polytechnic University of Madrid, Madrid, Spain, in 2004, and the Ph.D. degree from the Communication Systems Group, Technical University of Berlin, Berlin, Germany, in 2008.

He was a Research Assistant at the Berlin-based company zplane.development and at the Communication Systems Group of the Technical University of Berlin. In 2007, he joined IRCAM, Paris, France, where he currently works as a Researcher. His research interests include audio content analysis, music information retrieval, source separation, and sound synthesis. He also holds a degree in piano and music theory from the Madrid Conservatory of Music.

Axel Röbel received the Diploma in electrical engineering from Hannover University, Hannover, Germany, in 1990 and the Ph.D. degree (*summa cum laude*) in computer science from the Technical University of Berlin, Berlin, Germany, in 1993.

In 1994, he joined the German National Research Center for Information Technology (GMD-First), Berlin, where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996, he became Assistant Professor for digital signal processing in the Communication Science Department, Technical University of Berlin. In 2000, he obtained a research scholarship to pursue his work on adaptive sinusoidal modeling at CCRMA Standford University, Stanford, CA, and in the same year he joined IRCAM for working in the analysis–synthesis team doing research on frequency-domain signal processing. In summer 2006, he was Edgar-Varèse Guest Professor for computer music at the electronic studio of the Technical University of Berlin. Since 2008, he has been Deputy Head of the Analysis Synthesis Team, IRCAM. His current research interests are related to music and speech signal modeling and transformation.

Thomas Sikora (M'93–SM'96) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from Bremen University, Bremen, Germany, in 1985 and 1989, respectively.

In 1990, he joined Siemens, Ltd., and Monash University, Melbourne, Australia, as a Project Leader responsible for video compression research activities in the Australian Universal Broadband Video Codec consortium. From 1994 to 2001, he was the Director of the Interactive Media Department, Heinrich Hertz Institute (HHI), Berlin GmbH, Germany. He is cofounder of 2SK Media Technologies and Vis-a-Pix GmbH, two Berlin-based startup companies involved in research and development of audio and video signal processing and compression technology. He is currently a Professor and director of the Communication Systems Group at Technische Universität Berlin, Germany.

Dr. Sikora has been involved in international ITU and ISO standardization activities as well as in several European research activities for a number of years. As the Chairman of the ISO-MPEG (Moving Picture Experts Group) video group, he was responsible for the development and standardization of the MPEG-4 and MPEG-7 video algorithms. He also served as the chairman of the European COST 211ter video compression research group. He was appointed as Research Chair for the VISNET and 3DTV European Networks of Excellence. He is an Appointed Member of the Advisory and Supervisory board of a number of German companies and international research organizations. He frequently works as an industry consultant on issues related to interactive digital audio and video.