A SEGMENTAL SPECTRO-TEMPORAL MODEL OF MUSICAL TIMBRE

Juan José Burred *, *Axel Röbel*

Analysis/Synthesis Team,
IRCAM
Paris, France
{burred,roebel}@ircam.fr

ABSTRACT

We propose a new statistical model of musical timbre that handles the different segments of the temporal envelope (attack, sustain and release) separately in order to account for their different spectral and temporal behaviors. The model is based on a reduced-dimensionality representation of the spectro-temporal envelope. Temporal coefficients corresponding to the attack and release segments are subjected to explicit trajectory modeling based on a non-stationary Gaussian Process. Coefficients corresponding to the sustain phase are modeled as a multivariate Gaussian. A compound similarity measure associated with the segmental model is proposed and successfully tested in instrument classification experiments. Apart from its use in a statistical framework, the modeling method allows intuitive and informative visualizations of the characteristics of musical timbre.

1. INTRODUCTION

Our goal is to develop a computational model of musical instrument sounds that is accurate and flexible enough for several sound processing and content analysis applications. We seek a compact representation of both temporal and spectral characteristics, distinctive for each instrument, that is able to describe or predict the essential time-frequency behaviours of a range of isolated notes of a particular instrument. We formulate the problem as a supervised learning task, based on a labeled training database, that estimates a statistical model.

We put special emphasis on the temporal aspect: since the early studies by Helmholtz, it is well-known that not only the spectral shape, but also its evolution in time plays a crucial role in the distinction between instruments, i.e. in our perception of timbre. However, when it comes to computational modeling of music for analysis or synthesis purposes, research has traditionally given more importance to the spectral aspect. This is true for the two research fields of relevance here: music content analysis (or information retrieval) and music sound transformation and synthesis.

In music information retrieval, in which pattern recognition algorithms are applied for classification or search by similarity, the predominant architecture is to extract a set of short-time features that roughly describe the spectral shape, followed by a simple temporal modeling consisting of a statistical measure of their evolution across a certain fixed-length temporal segment. Common features range from low-level measures, describing the spectral shape with a scalar (such as spectral centroid, flatness, kurtosis, etc.) to mid-level multidimensional features including a moderate level of auditory modeling, such as *Mel Frequency Cepstral Coefficients* (MFCC) or auditory filter banks. Examples of temporal modeling approaches include computing velocity and acceleration coefficients, measuring statistical moments across a mid- to long-term window or using autoregressive models [1]. Feature extraction is typically followed by a statistical classification model that either completely ignores the temporal sequence of the features, such as *Gaussian Mixture Models* (GMM) or *Support Vector Machines* (SVM), or reduces it to a discrete sequence of states, such as *Hidden Markov Models* (HMM). Only recently, more detailed temporal models have been proposed in this context. As an example, we cite the work by Joder *et al.*[2], where alignment kernels are studied as a replacement of traditional static kernels for SVM classification. It should be noted that the adequate level of spectral and temporal accuracy of the model will strongly depend on the exact application context. When analyzing full music tracks, it will be unhelpful to attempt a highly accurate extraction of both spectral and temporal envelopes, due to the huge variability they will present in the training database. However, if the goal is to analyze or classify isolated instrumental sounds (as it is in the present contribution), both spectral and temporal characteristics will be highly structured and can thus be exploited by a more accurate model. The high variability and unpredictability of full music tracks is also the reason why the music analysis community has focused less on temporal structure than the speech analysis community.

Concerning sound transformation and synthesis, much attention has been given to the accurate estimation of the spectral envelope [3], and to the study of the corresponding formant structures. When signal reconstruction is needed (sound transformation, synthesis, source separation), source models have to be far more accurate than in information retrieval applications. Thus, more sophisticated models are typical in this area, such as spectral basis decompositions [4] or models based on sinusoidal modeling [5]. Still, when it comes to statistical learning, the temporal evolution is also often ignored, or approximated by simple temporal smoothness constraints. For instance, Virtanen [5] and Kameoka *et al.* [6] both model temporal smoothness as a superposition of temporal windows, and in [7] a Markov chain prior is imposed on the temporal coefficients controlling the superposition of a set of spectral bases. Bloit *et al.* [8] use a more explicit modeling of feature trajectories by a generalization of HMM in which the static distributions of the states are replaced by a collection of *curve primitives* that represent basic trajectory segment shapes.

Our main motivation is to model temporal evolution at a still higher degree of accuracy. As will be seen, in some cases we avoid temporal discretization altogether and attempt to explicitly model the trajectories in feature space. Such a model was presented in our previous works [9, 10], and will be briefly summarized in Sect. 2.

* J.J. Burred is now with Audionamix, Paris, France.

In short, our previous model extracts a set of dimension-reduced coefficients describing the spectral envelope, while keeping their temporal ordering. Then, all coefficient trajectories for each instrument class are collapsed into a prototype trajectory that corresponds to a *Gaussian Process* (GP) with varying mean and covariance.

The fact that our previous model used a single GP prototype trajectory per instrument gave rise to important limitations, as will be described. This contribution builds on those works by replacing the single-GP model with a compound model in which the attack, sustain and release segments of the temporal envelope are modeled separately. This solves two important drawbacks of the GP model. First, it allows using different statistical models for different segments, thus accounting for their possibly very different behaviours at the feature level. As will be seen, the shapes of feature trajectories are very descriptive in the transient phases (attack and release), but, as can be expected, they will vary less in sustained regions. In the latter case, a cluster model will be more appropriate than an explicit trajectory. And second, it avoids the implicit time-stretching of the attack and release phases that was needed when learning the GP model. This issue will be better understood when we will address it in more detail in the next section.

We will begin our presentation with a brief summary of our previous GP-based modeling approach (Sect. 2). Sect. 3 will introduce the assumptions and methods we use for the segmentation of the temporal envelope. The new *spectro-temporal segmental model* will be presented in detail in Sect. 4. Finally, we will present two applications of the segmental model: to classification of isolated samples (Sect. 6), where an increase of performance compared to the GP model is reported, and to timbre visualization (Sect. 5).

2. DYNAMIC SPECTRAL ENVELOPE MODELING

We aim at modeling the spectral envelope and its evolution in time, to which we will jointly refer as *spectro-temporal envelope*. Since our previous approach to that end has been described and evaluated in detail in our previous works [9, 10], we will only present it here very briefly.

The first step is to extract the spectro-temporal envelopes from a large set of files belonging to a training database. To that end, we perform sinusoidal modeling (i.e., peak picking and partial tracking) on the individual notes, followed by an inter-peak interpolation in frequency to obtain a smooth spectral shape. Then, dimensionality reduction is performed via *Principal Component Analysis* (PCA). All the spectro-temporal envelopes need thus to be organized into a rectangular data matrix $\mathbf{X}$ that will be subjected to a factorization of the form

$$\mathbf{X} = \mathbf{P}\mathbf{Y}, \tag{1}$$

where $\mathbf{P}$ is a $K \times K$ matrix of spectral bases and $\mathbf{Y}$ is a $K \times T$ matrix of temporal coefficients (K is the frequency bin index and T is the time frame index). To accommodate the envelopes into $\mathbf{X}$ while keeping formants aligned in frequency, the envelopes are sampled at a regular frequency grid defined by $k = 1, \ldots, K$. The reduced-dimensional PCA projection of size $D \times T$ with $D < K$ is then given by

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

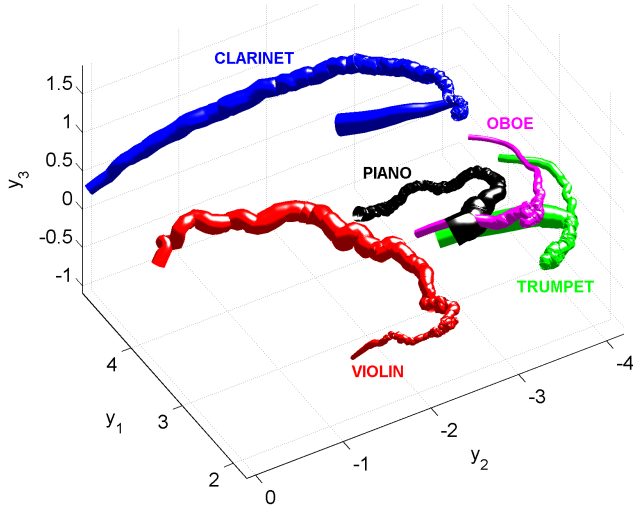


Figure 1: First three dimensions of the prototype tubes corresponding to a set of 5 Gaussian Process (GP) timbre models.

where $\mathbf{\Lambda}_\rho = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ and λ_d are the D largest eigenvalues of the covariance matrix

$$\mathbf{\Sigma}_{\boldsymbol{X}} = E\{(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^T\}. \tag{3}$$

Each point in the PCA space defined by the above equations will correspond to a spectral envelope shape, and a trajectory will correspond to a variation in time of the spectral envelope, i.e., to a spectro-temporal envelope in the time-frequency domain.

2.1. Gaussian Process Model

The projected coefficients $\mathbf{Y}_\rho$ are considered the features that will be subjected to statistical learning. Each training sample will result in a feature trajectory in PCA space. The aim of the learning stage of the GP model is to collapse all individual training trajectories into a prototype curve, one for each instrument class. To that end, the following steps are taken. First, all trajectories are interpolated in time using the underlying time scales in order to obtain the same number of points. Then, each point of index r in the resulting prototype curve for instrument i is considered to be a D-dimensional Gaussian random variable $\mathbf{p}_{ir} \sim \mathcal{N}(\boldsymbol{\mu}_{ir}, \mathbf{\Sigma}_{ir})$ with empirical mean $\boldsymbol{\mu}_{ir}$ and empirical covariance matrix $\mathbf{\Sigma}_{ir}$. A prototype curve can be thus interpreted as a D-dimensional, nonstationary GP with time-varying means and covariances parametrized by the frame index r:

$$\mathcal{M}_i \sim \mathcal{GP}\left(\boldsymbol{\mu}_i(r), \mathbf{\Sigma}_i(r)\right). \tag{4}$$

Rather than prototype curves (corresponding to the means $\boldsymbol{\mu}_i(r)$), the resulting models in PCA space have the shape of *prototype tubes* with varying widths proportional to the covariance $\mathbf{\Sigma}_i(r)$. Figure 1 shows the representation in the first 3 dimensions of PCA space of a set of 5 GP models learnt from a database of 174 audio samples. The used samples are a subset of the RWC database [11]. As measured in [10] in terms of explained variance, the first 3 principal components already contain around 90% of information.

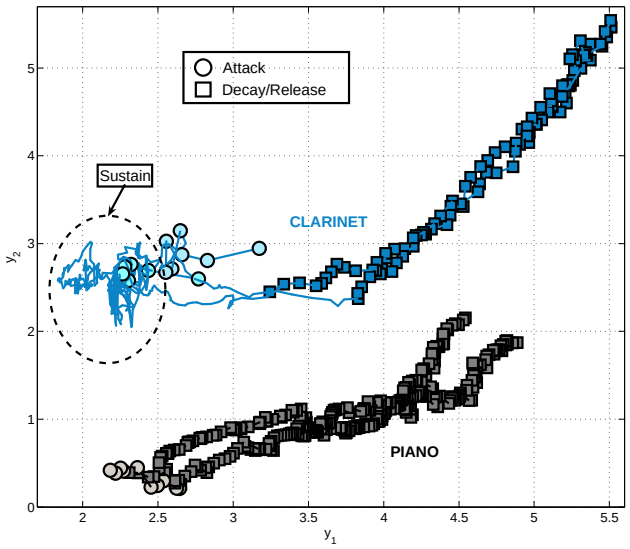


Figure 2: Example of attack, sustain and decay/release segments in PCA space: 2 clarinet and 2 piano notes from the training database.

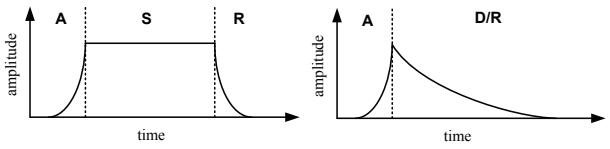
2.2. Limitations of the GP Model

GP models of the spectro-temporal envelope, and their corresponding visualization as prototype tubes, are adequate for trajectories with a slowly evolving gradient (i.e., not changing direction too often). As was observed with individual training samples, this is the case for the attack, release and decay sections of the notes. In sustained segments, the spectral envelope stays relatively constant and thus the corresponding feature trajectory will oscillate inside a small region of space, with little or no net displacement, suggesting a cluster rather than a trajectory. Interpolating and keeping the time alignment to learn a GP in such segments will mostly lead to complicated and highly random trajectories that can hinder both classification performance and generalization.

A graphical example of this observation is shown in Fig. 2. Four coefficient trajectories corresponding to four individual training samples (two clarinet notes, in blue, and two piano notes, in gray) are shown in their projection onto the first two dimensions of PCA space. The trajectory curves are superimposed by circles in the attack segments and by squares in the release/decay segments. The piano notes are non-sustained: their trajectories show a net displacement across their whole duration. The clarinet notes, being sustained, show a clearly different graphical behavior. The sustain part corresponds to the indicated cluster-like area, where there is little net displacement. The "tails" corresponding to attack and relase/decay and coming out (or into) the cluster are clearly recognizable. Although not represented here, the cluster-like behavior of the sustain phase is also observable under other space projections and other dimensions.

Such observations suggest the segmentation of the training samples into sustained and non-sustained sections before the learning stage, so that sustained sections can be learnt by cluster-like models and non-sustained ones by trajectory-like models.

Another limitation of the single-GP approach arises from the interpolation performed previous to the learning of the time-



(a) Attack - Sustain - Release. (b) Attack - Decay/Relase.

Figure 3: Simplified temporal segmentation models for sustained (a) and non-sustained (b) notes.

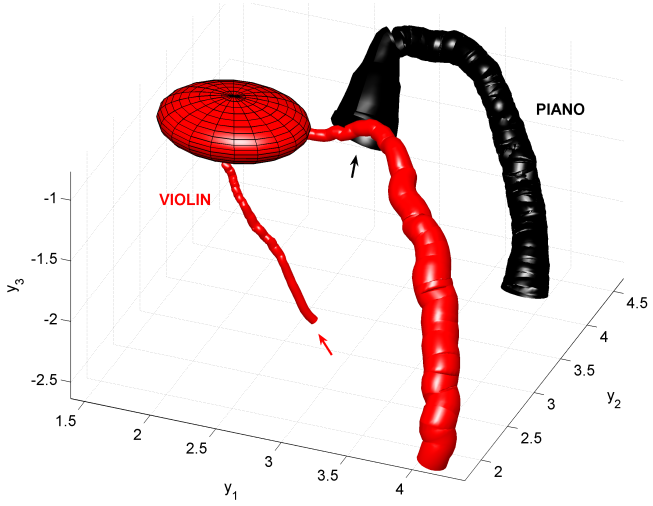
varying means $\boldsymbol{\mu}_i(r)$ and covariances $\boldsymbol{\Sigma}_i(r)$. Interpolating all curves with the same number of points corresponds to time normalization. Thus, for sustained sounds, this will have the implicit effect of misaligning the attack and release phases. When aligning a short sustained note with a long sustained note of the same instrument, the attack and release portions of the short note will be excessively stretched. This results in portions of the attack and release of some notes being modeled together with sustained portions of other notes, hindering model performance and unnaturally increasing its variance. Instead, attack and release segments vary relatively little in duration across notes in similar pitch ranges for a particular instrument, whereas the sustain segment can have an arbitrary duration. This further motivates the temporal segmentation of the input signals.

3. TEMPORAL SEGMENTATION

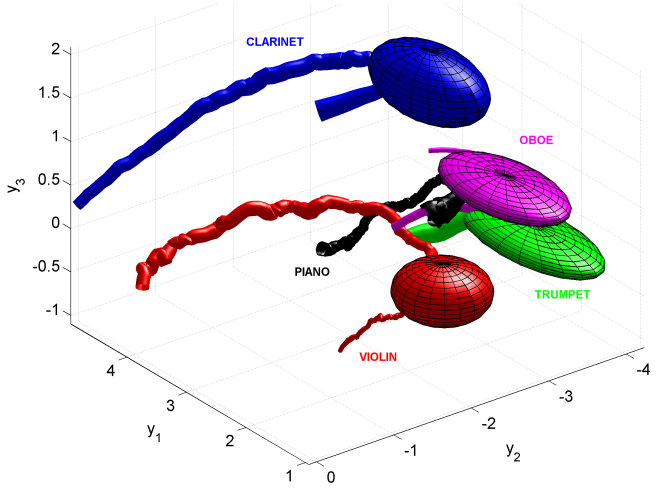
The segmentation of a musical note into its attack, sustain and release components is usually performed by applying thresholds to its amplitude or energy temporal envelope. The best known segmentation model, the attack-decay-sustain-release (ADSR) envelope, popularized by early analog synthesizers, is hardly generalizable to acoustic musical instruments. Instead, we consider two separate simple segmentation schemes (see Fig. 3), one for sustained sounds (e.g. wind instruments or bowed strings) and one for non-sustained sounds (e.g. struck or plucked strings, membranes or bars):

- **ASR model (sustained sounds).** Consisting of an attack segment, a sustain segment (of arbitrary length) and a release segment between the end of the excitation and the end of the vibrations.
- **AD/R model (non-sustained sounds).** Consisting of an attack segment and a "rest" segment that can be interpreted as either decay D or release R. This is to account for the fact that some authors call the rest segment "decay" (the energy is freely decaying), while others call the rest segment "release" (the excitation has been released).

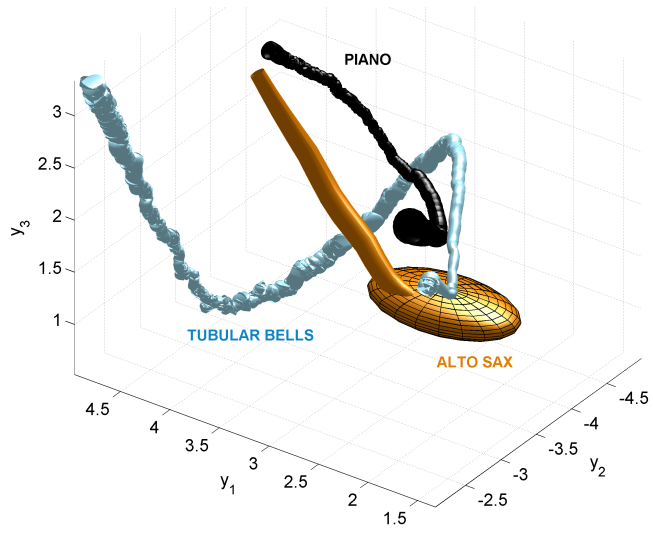
We use the automatic segmentation method proposed in [12], based on measuring the change rate of the slopes of the energy envelope and using adaptive thresholds. In spite of the simplicity of the segmentation scheme used, it has proven adequate enough for our purposes. Of course, the modeling process will benefit from other, more sophisticated, temporal segmentation methods. For example, automatic segmentation should also take spectral cues into account, as suggested in [13].



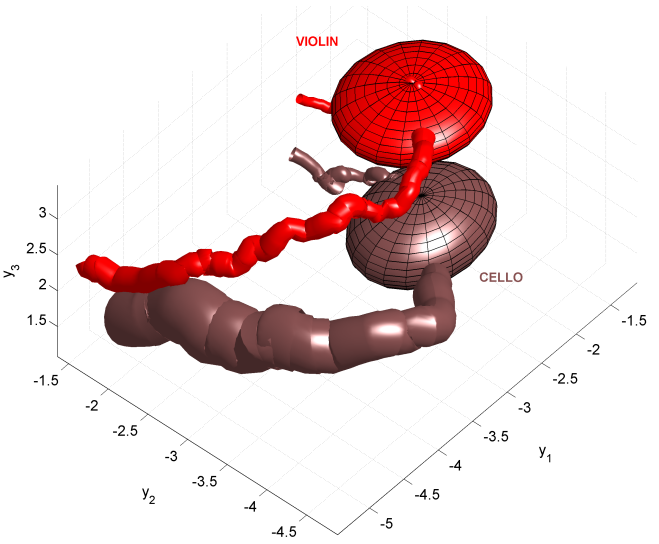
(a) Comparison of a sustained instrument with a non-sustained instrument. The arrows indicate the starting points of the models.



(b) Segmental version of Fig. 1.



(c) Non-sustained struck strings (piano) vs. non-sustained struck bars (tubular bells) vs. sustained woodwind (alto sax).



(d) Comparison of instruments from the same family (bowed strings).

Figure 4: Examples of timbre visualizations with segmental spectro-temporal models.

4. SEGMENTAL SPECTRO-TEMPORAL MODEL

Following the previous observations, we propose to replace the GP model with a compound model with heterogeneous models for each segment of the temporal envelope, which we call the *segmental spectro-temporal model* (SST). Attack and release/decay segments will be modeled by trajectory-like models, for which we use the interpolated GP approach that was applied in Sect. 2.1 to the trajectory as a whole, giving rise to the, respectively, attack and release/decay tubes with the following probability distributions:

$$p_i^A(\boldsymbol{x}) = \mathcal{GP}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i^A(r), \boldsymbol{\Sigma}_i^A(r), r \in \mathcal{R}_i^A\right) \quad (5)$$

$$p_i^{D/R}(\boldsymbol{x}) = \mathcal{GP}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i^{D/R}(r), \boldsymbol{\Sigma}_i^{D/R}(r), r \in \mathcal{R}_i^{D/R}\right) \quad (6)$$

where $\mathcal{R}_i^A$ and $\mathcal{R}_i^{D/R}$ are, respectively, the index sets for the A and D/R segments after interpolation. Note that interpolation (with implicit time normalization) is now only performed on the corre-

sponding subset of indices, avoiding excessive time stretching due to the influence of the sustain segment.

Sustain is modeled by a multivariate Gaussian cluster with full covariance matrix:

$$p_i^S(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_i^S, \boldsymbol{\Sigma}_i^S\right). \quad (7)$$

Note that, for the A and D/R segments, we have used the notation (r) to denote explicit temporal dependence, whereas for the S segment, the notation denotes a static model in which the individual samples are statistically independent from each other.

We thus obtain the following compound mixture models for, respectively, sustained and non-sustained sounds:

$$p_i^{\text{sust}}(\boldsymbol{x}) = p_i^A(\boldsymbol{x}) + p_i^S(\boldsymbol{x}) + p_i^{D/R}(\boldsymbol{x}) \quad (8)$$

$$p_i^{\text{n.sust}}(\boldsymbol{x}) = p_i^A(\boldsymbol{x}) + p_i^{D/R}(\boldsymbol{x}). \quad (9)$$

5. APPLICATION TO TIMBRE VISUALIZATION

The segmental modeling method is highly appropriate for the graphical representation of timbre characteristics. The use of dimension reduction via PCA implies that most information (in terms of variance) will be concentrated in the first few dimensions, and thus 2-D or 3-D representations of the feature space will be highly illustrative of the essential timbral features. Also, since a common set of bases is used for the entire training set, it is possible to visually assess the timbre similarities and dissimilarities between different instruments through the distance of their models in space. Finally, the use of compound models allows the use of different geometrical objects for a visually appealing presentation and fast assessment of spectro-temporal behavior. Sustain segments correspond to ellipsoids, from which variable-diameter tubes arise that correspond to attack and decay/release phases. The length of the ellipsoid axes and the variable widths of the tubes are proportional to the model covariances, with the proportionality factor selected for an adequate visual characterization.

Several graphical examples of timbre visualizations based on SST models are presented in Fig. 4. Fig. 4(a) shows the visual comparison between a sustained (violin) and a non-sustained instrument (piano). This figure corresponds to a training database of 171 samples. The sustain segment of the violin is represented as an ellipsoid described by the covariance of its Gaussian distribution. The attack segment of the piano shows a greater variance than the decay segment. Fig. 4(b) is the segmental counterpart of Fig. 1, showing the resulting SST models from the exact same database of 5 instruments.

Figure 4(c) shows the comparison between a struck bar percussion instrument (tubular bells), a struck string instrument (piano) and a sustained woodwind instrument (alto saxophone). Notable in this figure is the great spectral variability of the bells: their prototype curve traverses more regions in space than the other models. It should be recalled at this point that longer curves in PCA space do not correspond to longer notes, since time has been normalized by interpolation. Longer curves in space correspond to a greater variability of spectral envelope shape.

Finally, Fig. 4(d) shows the timbre comparison between two instruments (violin and cello) from the same family (bowed strings), and playing the same range of notes. It can be observed that the general shape of the model is similar, suggesting a similarity in timbre. From the third dimension on, however, the models are indeed shifted from each other. Also notable in this case is the much higher variance of the cello in the release phase.

Since it is difficult to find one particular projection that highlights the important features for all instruments at the same time, a better visualization can be achieved by letting the user rotate the figures on a computer.

6. APPLICATION TO CLASSIFICATION

An example of application of the models to the field of information retrieval is the classification of isolated musical samples. An evaluation of the models in such a task also helps assessing their discriminative power. Classification can be performed by projecting an unknown sound into feature space and defining a global distance or likelihood between the projected interpolated unknown trajectory $\breve{\mathcal{U}}$ and the stored compound models. In our previous work based on instrument-wise GP modeling [10], such distance was simply the average Euclidean distance between the input trajectory and each one of the stored prototype curves:

$$d(\breve{\mathcal{U}}, \mathcal{M}_i) = \frac{1}{R_{max}} \sum_{r=1}^{R_{max}} \sqrt{\sum_{k=1}^{D} (\breve{u}_{rk} - \mu_{irk})^2}, \quad (10)$$

where R_{max} denotes the maximum number of frames among the stored models and the $\breve{}$ symbol denotes interpolation. In order to also take into account the variance of the prototypes, classification based on GP models can be instead reformulated as a maximum likelihood problem based on the following point-to-point likelihood:

$$\mathcal{L}(\breve{\mathcal{U}}|\boldsymbol{\mu}_i(r), \boldsymbol{\Sigma}_i(r)) = \prod_{r=1}^{R_{max}} \mathcal{N}\left(\breve{\boldsymbol{u}}(r)|\boldsymbol{\mu}_i(r), \boldsymbol{\Sigma}_i(r)\right). \quad (11)$$

For the SST model, the different model types call for the use of hybrid distance measures. The first step is to segment the incoming signal following the method of Sect. 3. Afterwards, the sound is identified as either sustained or non-sustained. This will be necessary for the later choice of appropriate distance measure. This detection is performed here with the following simple but efficient rule: a sound is classified as non-sustained if the beginning of the release/decay segment is detected before half the duration of the sound. Once the input sound has been segmented, for comparison of the A and D/R segments, the GP likelihood definition of Eq. 11 will be used, after replacing the parameters with the ones corresponding to either segment.

For the S segment, a different type of similarity measure is needed, without the explicit temporal ordering of Eq. 11. We wish to compare the Gaussian clusters of the sustain models (p_i^S) with a Gaussian cluster of the data points belonging to the sustain part of the unknown input sound, denoted here as $p_{\breve{\boldsymbol{u}}}^S$. The Kullback-Leibler (KL) divergence is thus an appropriate choice:

$$D_{KL}(p_{\breve{\boldsymbol{u}}}^S \| p_i^S) = \sum_{\boldsymbol{x}} p_{\breve{\boldsymbol{u}}}^S(\boldsymbol{x}) \log \frac{p_{\breve{\boldsymbol{u}}}^S(\boldsymbol{x})}{p_i^S(\boldsymbol{x})} \quad (12)$$

which in the case of multivariate Gaussian distributions has the following analytic expression:

$$\begin{aligned} D_{KL}(p_{\breve{\boldsymbol{u}}}^S \| p_i^S) &= \frac{1}{2}\left(\log\left(\frac{\det \boldsymbol{\Sigma}_i^S}{\det \boldsymbol{\Sigma}_{\breve{\boldsymbol{u}}}^S}\right) + \operatorname{tr}((\boldsymbol{\Sigma}_i^S)^{-1}\boldsymbol{\Sigma}_{\breve{\boldsymbol{u}}}^S)\right. \\ &+ \left.(\boldsymbol{\mu}_i^S - \boldsymbol{\mu}_{\breve{\boldsymbol{u}}}^S)^T(\boldsymbol{\Sigma}_i^S)^{-1}(\boldsymbol{\mu}_i^S - \boldsymbol{\mu}_{\breve{\boldsymbol{u}}}^S) - D\right), \end{aligned}$$

Model	Measure	5 dimensions	10 dimensions
GP	Euclidean	88.26 ±3.02	92.94 ± 2.12
GP	Likelihood	90.64 ±3.51	94.59 ± 2.46
SST	Likelihood	93.67 ±1.70	95.41 ± 2.16
SST	Likel. + KL	94.40 ± 2.00	96.61 ± 1.94

Table 1: Classification results (mean classification accuracy % ± standard deviation across cross-valiation folds).

where $(\boldsymbol{\mu}_{\breve{\boldsymbol{u}}}^S, \boldsymbol{\Sigma}_{\breve{\boldsymbol{u}}}^S)$ are the parameters of the sustain part of the input trajectory and D is the number of dimensions.

The global similarity measure between the unknown input trajectory and a segmental model is finally defined as the following compound log-likelihood function:

$$\begin{aligned} \log \mathcal{L}(\breve{\mathcal{U}}|\boldsymbol{\theta}_i) &= \log \mathcal{L}(\breve{\mathcal{U}}|\boldsymbol{\mu}_i^A(r), \boldsymbol{\Sigma}_i^A(r)) \\ &+ \log \mathcal{L}(\breve{\mathcal{U}}|\boldsymbol{\mu}_i^{D/R}(r), \boldsymbol{\Sigma}_i^{D/R}(r)) \\ &- \alpha D_{KL}(p_{\breve{\boldsymbol{u}}}^S \| p_i^S), \end{aligned} \quad (13)$$

where $\alpha = 1$ if the sound is classified as sustained and $\alpha = 0$ if the sound is classified as non-sustained. $\boldsymbol{\theta}_i$ denotes the ensemble of model parameters. Of course, the models not relevant to the sound class detected (sustained/non-sustained) need not to be included in the maximum likelihood evaluation.

For the classification experiments, a database of 5 instrument classes was used. The database consists of a selection of isolated samples from the RWC music database [11]. The classes include 4 sustained instruments (clarinet, oboe, violin and trumpet) and 1 non-sustained instrument (piano). Each class contains all notes for a range of two octaves (C4 to B5), in three different dynamics (forte, mezzoforte and piano) and normal playing style. This makes a total of 1098 individual note files, all sampled at 44.1 kHz. The experiments were iterated using a random partition into 10 cross-validation training/test sets. The frequency grid was of $K = 40$ points, linear interpolation was used for the frequency interpolation and cubic interpolation was used for the temporal interpolation of the GP curves in PCA space. All experiments were repeated for two different dimensionalities: $D = 5$ and $D = 10$.

The results are shown in Table 1. The first row corresponds to the GP model evaluated with average Euclidean distances (Eq. 10), as in the previous system presented in [10]. Using the variance information by means of the likelihood of Eq. 11 improves the performance, as shown in the second row of the table. The best results, however, are obtained with the proposed segmental (SST) model. The full segmental model with the compound likelihood/divergence measure of Eq. 13 offers the best performance at 94.40% mean accuracy for $D = 5$ dimensions and at 96.61% mean accuracy for $D = 10$ dimensions.

We performed an additional experiment for testing the influence of the sustain segment in the classification. This was done by always forcing $\alpha = 0$ in Eq. 13, both for sustained and non-sustained input sounds. The results are shown in the third row of the table. Even if, as expected, the performance is lower than with the complete model, it is a remarkable result that its influence on the classification performance is rather low. This suggests that Eq. 13 might need the inclusion of different weights for its different terms, so that the influence of the individual segments are better balanced. Such a weighting scheme will be explored in the future.

7. CONCLUSIONS AND OUTLOOK

We have presented the segmental spectro-temporal (SST) model for the statistical characterization and visualization of the timbre of musical sounds. The model considers the temporal amplitude segments of each note (attack, sustain, release) separately in order to address their different behaviors in both time and frequency domains. Feature extraction is based on the estimation of the spectro-temporal envelope, followed by a dimensionality reduction step. The portions of the resulting feature trajectories corresponding to attack, release and decay segments are modeled as non-stationary Gaussian Processes with varying mean and covariances. The sustain part is modeled as a multivariate Gaussian. We proposed a compound similarity measure associated with the SST model, so that the method can readily be used for classification purposes. In particular, classification experiments with isolated samples showed an improved performance (in terms of classification accuracy) compared to our previously proposed single-Gaussian-Process model.

Apart from their use in a statistical framework, the modeling method allows intuitive and informative visualizations of the characteristics of musical timbre, including an explicit depiction of timbre similarity (or dissimilarity) between instruments.

The segmental approach is a flexible strategy that opens interesting research directions. More refined models could be envisioned for the individual segments, or for modeling variations on the playing styles. For instance, we could analyze how vibrato affects the shape of the sustain cluster, or how articulations such as stacatto, martelatto, etc., affect the behaviour of the attack trajectory.

There is also a shortcoming that needs to be addressed. Our feature extraction strategy favours the alignment of formants before performing dimensionality reduction (this issue was only briefly mentioned on this contribution, but addressed in detail in [9]). Unlike formants, other spectral features depend on pitch and will be lost in the alignment. A notable example is the predominance of odd partials in the spectra of wind instruments with both closed tubes and cylindrical bores, such as the clarinet. For such instruments, an alternative, pitch-dependent representation is desirable. In this context, a related research direction has been started in which pitch-dependent and pitch-independent features are decoupled by means of a source-filter model. This principle could be combined with the explicit trajectory modeling methods presented here.

8. REFERENCES

[1] A. Meng and J. Shawe-Taylor, "An investigation of feature models for music genre classification using the support vector classifier," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[2] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17 (1), pp. 174–186, January 2009.

[3] Axel Röbel, Fernando Villavicencio, and Xavier Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28-11, pp. 1343–1350, 2007.

[4] M. Casey and A. Westner, "Separation of mixed audio sources by Independent Subspace Analysis," in *Proc. International Computer Music Conference (ICMC)*, Berlin, Germany, 2000.

[5] T. Virtanen, "Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint," in *Proc. International Conference on Digital Audio Effects (DAFX)*, London, UK, 2003.

[6] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15 (3), pp. 982–994, March 2007.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[8] J. Bloit, N. Rasamimanana, and F. Bevilacqua, "Towards morphological sound description using segmental models," in *Proc. International Conference on Digital Audio Effects (DAFX)*, Como, Italy, September 2009.

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

[10] J.J. Burred, A. Röbel, and T. Sikora, "Dynamic spectral envelope modeling for the analysis of musical instrument sounds," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18 (3), pp. 663–674, March 2010.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003.

[12] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," in *CUIDADO I.S.T. Project Report*, 2004.

[13] J. Hajda, "A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited," *Journal of the Audio Engineering Society*, November 1996.