POLYPHONIC MUSICAL INSTRUMENT RECOGNITION BASED ON A DYNAMIC MODEL OF THE SPECTRAL ENVELOPE

Juan José Burred, Axel Röbel

Analysis/Synthesis Team
IRCAM - CNRS STMS
75004 Paris, France
{burred,roebel}@ircam.fr

Thomas Sikora

Communication Systems Group
Technical University of Berlin
10587 Berlin, Germany
sikora@nue.tu-berlin.de

ABSTRACT

We propose a new method for detecting the musical instruments that are present in single-channel mixtures. Such a task is of interest for audio and multimedia content analysis and indexing applications. The approach is based on grouping sinusoidal trajectories according to common onsets, and comparing each group's overall amplitude evolution with a set of pre-trained probabilistic templates describing the temporal evolution of the spectral envelopes of a given set of instruments. Classification is based on either an Euclidean or a probabilistic definition of timbral similarity, both of which are compared with respect to detection accuracy.

Index Terms— Musical instrument classification, Music Information Retrieval, spectral envelope, Gaussian Processes.

1. INTRODUCTION

We present a method for detecting the musical instruments that are present in monaural (single-channel) linear mixtures. This is of interest for music content analysis applications such as indexing and retrieval, transcription and source separation.

Past research work concerning automatic classification of musical instruments has mostly concentrated on the isolated-instrument case. In comparison, the more demanding and realistic polyphonic case has only been addressed recently. Approaches aiming at that goal typically either consider the mixture as a whole [1] or attempt to separate the constituent sources with prior knowledge related to pitch [2].

The proposed method is based on the grouping and 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 can detect highly inharmonic instruments. The amplitude of common-onset sinusoidal trajectories is matched against a set of probabilistic time-frequency (t-f) models of the spectral envelope.

2. DYNAMIC SPECTRAL ENVELOPE MODELING

The used timbre models are based on the spectral envelope and its evolution in time, which are two of the most important factors contributing to the characteristic timbre of each musical instrument. Detailed validation experiments of the models were reported in [3]. The first step of the training consists in performing Principal Component Analysis (PCA) on the set of all training spectral envelopes extracted by means of sinusoidal modeling and frequency interpolation, extracted from a database of isolated notes. We used a subset of the RWC database [4]. PCA was used as spectral decomposition transform because of its optimal compression capabilities.

For each instrument, a sequence of notes belonging to a section of the chromatic scale are considered for the training of each model. To obtain the rectangular data matrix $\mathbf{X}$ needed for PCA, the amplitudes of the training envelopes are linearly interpolated to a regular frequency grid of K bins. Then, spectral decomposition via PCA factorizes the data matrix as $\mathbf{X} = \mathbf{PY}$, where the columns of the $K \times K$ basis matrix $\mathbf{P}$ are the eigenvectors of the covariance matrix of the data matrix $\mathbf{X}$, and $\mathbf{Y}$ are the projected coefficients. After whitening, the final projection $\mathbf{Y}_\rho$ 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.

In PCA space, the projected coefficients are then grouped into a set of generic models representing the classes. Here, for the sake of accuracy, the time variation of the envelope is modeled as a trajectory rather than using cluster-based approximations, such as Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM). For each class, all training trajectories are collapsed into a single *prototype curve*.

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 $\mathbf{p}_{ir} \sim N(\boldsymbol{\mu}_{ir}, \mathbf{\Sigma}_{ir})$ 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 *Gaussian Process* (GP) with time-varying means and covariances:

$$\mathcal{C}_i \sim GP\left(\boldsymbol{\mu}_i(r), \mathbf{\Sigma}_i(r)\right). \tag{2}$$

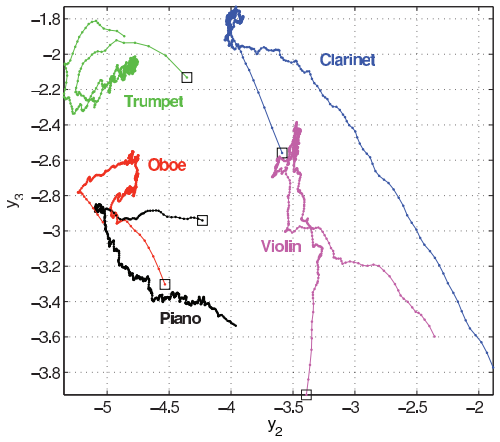


Fig. 1. Two-dimensional projection of the prototype curves corresponding to a 5-class database. Squares denote the starting points.

Figure 1 shows a 2-dimensional projection of an example set of the mean prototype curves corresponding to a training set of 5 classes: piano, clarinet, oboe, violin and trumpet. The database consists of all dynamic levels (piano, mezzoforte and forte) of two or three exemplars of each instrument type, with normal playing style, covering a range of one octave (C4-B4).

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 frequency grid. The reconstructed mean vector is

$$\hat{\boldsymbol{\mu}}_{ir} = \mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\mu}_{ir} + E\{\mathbf{X}\} \tag{3}$$

and, assuming diagonal covariance for simplicity, the corresponding variance vector is

$$\hat{\boldsymbol{\sigma}}_{ir}^2 = \mathrm{diag}\left(\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2} \boldsymbol{\Sigma}_{ir} (\mathbf{P}_\rho \boldsymbol{\Lambda}_\rho^{1/2})^T\right), \tag{4}$$

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 Gaussian Process, but this time it is unidimensional and parametrized with means and variances varying in the t-f plane:

$$\mathcal{E}_i \sim GP\left(\mu_i(t,f), \sigma_i^2(t,f)\right). \tag{5}$$

3. ONSET DETECTION AND TRACK GROUPING

For classification, the mixture is first subjected to inharmonic sinusoidal extraction, yielding a set of sinusoidal tracks with frame-wise evolution in amplitude and frequency (phase is discarded). This is followed by a simple onset detection stage, based on the detection function $o(r) = \sum_{p\in\mathcal{N}_r} \frac{1}{\hat{f}_{pr}}$, where $\hat{f}_{pr}$ is the estimated frequency of partial p at frame r and $\mathcal{N}_r$ is the set of indices of the partials born at frame r. The peaks of this function are declared as the onset positions L_o^{on} for $o = 1,\ldots,O$ (given in frames).

After onset detection, all tracks $\mathbf{t}_t$ having its first frame within the interval $[L_o^{on} - Q, L_o^{on} + Q]$ for a given onset location L_o^{on} are grouped into the set $\mathcal{T}_o$, where o is the onset index. A value of $Q = 2$ was chosen. 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 it 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.

4. TIMBRE DETECTION

The timbre detection stage 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 aim, the core problem is to design an appropriate distance measure between the track groups and the models. A similar situation was described in our previous work [5], where the aim was to match partial clusters already separated by an external and independent separation method. In that case, an averaged Euclidean distance between the clusters and the t-f prototypes was used. Here, that basic idea is further developed and enhanced.

The first measure tested was the total Euclidean distance between the amplitude of each t-f bin belonging to $\mathcal{T}_o'$ and the surface $\mathbf{M}_i$ evaluated at the frequencies of $\mathcal{T}_o'$:

$$d(\mathcal{T}_o', \tilde{\mathbf{M}}_{io}) = \sum_{t\in\mathcal{T}_o'} \sum_{r=1}^{R_t} |A_{tr} - \mathbf{M}_i(f_{tr})|, \tag{6}$$

where R_t is the number of frames in track $\mathbf{t}_t \in \mathcal{T}_o'$ and A_{tr} and f_{tr} are the logarithmic amplitude and frequency, respectively, of the r-th frame of that track. In order to obtain the evaluation at the frequency support $\tilde{\mathbf{M}}_{io} = \mathbf{M}_i(\mathbf{F}_o)$, 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.

A probabilistic reformulation of such a measure allows to take into account not only the metric distance to the mean surfaces $\mathbf{M}_i$, but also the spread of their distribution as modeled by $\mathbf{V}_i$. To this end, the distance-minimization problem was redefined as a likelihood maximization. In particular, as measure of timbre similarity between $\mathcal{T}_o'$ and the instrument model formed by parameters $\boldsymbol{\theta}_i = (\mathbf{M}_i, \mathbf{V}_i)$, the following

likelihood function is used:

$$L(\mathcal{T}'_o|\boldsymbol{\theta}_i) = \prod_{t\in\mathcal{T}'_o}\prod_{r=1}^{R_t} p\left(A_{tr}|\mathbf{M}_i(f_{tr}), \mathbf{V}_i(f_{tr})\right), \quad (7)$$

where $p(x)$ denotes a unidimensional Gaussian distribution. The evaluation of the variance surface at the frequency support $\tilde{\mathbf{V}}_{io} = \mathbf{V}_i(\mathbf{F}_o)$ is performed in the same way as before.

A requirement on both formulations in order to be generally applicable is that they should not be affected by the overall gain and by the note length. To that end, a two-dimensional parameter search is performed, with one parameter controlling the amplitude scaling and one controlling the time extent. Amplitude scaling is introduced by the additive parameter α and time scaling is performed by jointly, linearly stretching the partial tracks towards the end of the note. Then, the Euclidean measure becomes the optimization problem

$$d(\mathcal{T}'_o, \tilde{\mathbf{M}}_{io}) = \min_{\alpha,N}\left\{\sum_{t\in\mathcal{T}'_o}\sum_{r=1}^{R_t}|A^N_{tr} + \alpha - \mathbf{M}_i(f^N_{tr})|\right\}, \quad (8)$$

and the likelihood-based problem is

$$L(\mathcal{T}'_o|\boldsymbol{\theta}_i) = \max_{\alpha,N}\left\{\prod_{t\in\mathcal{T}'_o} w_t \prod_{r=1}^{R_t} p\left(A^N_{tr} + \alpha|\mathbf{M}_i(f^N_{tr}), \mathbf{V}_i(f^N_{tr})\right)\right\}, \quad (9)$$

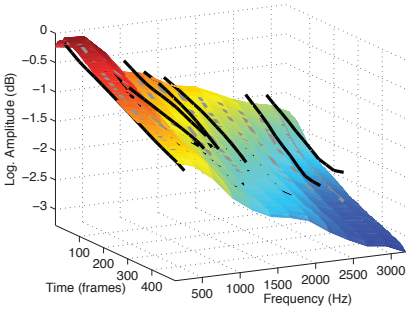
where A^N_{tr} and f^N_{tr} denote the amplitude and frequency values for a track belonging to a group that has been stretched so that its last frame is N. The factor w_t denotes an optional track-wise weighting defined by $w_t = e^{R_t/\bar{f}_t}$, where $\bar{f}_t$ is the mean frequency of the track, such that lower-frequency and longer tracks have a greater impact on the matching measure than higher-frequency and shorter tracks. Two different versions of the timbre likelihood were tested: weighted and unweighted (for the latter, $w_t = 1$).

Figure 2(a) shows an example of a good match between a track group belonging to a piano note (solid black lines) and a segment of the piano prototype envelope. The tracks have an overall strong similarity in both their frequency-dependent amplitude distribution and dynamic variation. In contrast, Fig. 2(b) is an example of a weak match between the same piano track group and the oboe model. Both spectral shape and dynamic behaviors differ significantly.

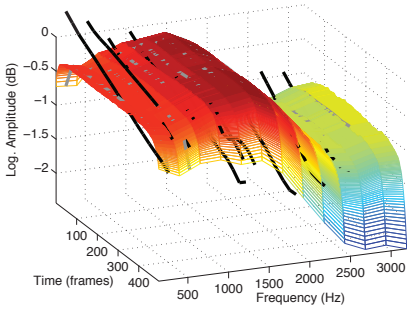
Figure 3(a) shows the optimization surfaces produced by an example parameter search (α, N) for a piano note, using the previous 5-instrument database. Figures 3(b) and 3(c) show representative projection profiles of the surfaces with fixed stretching and scaling parameters, respectively.

5. EXPERIMENTAL RESULTS

The single-channel mixtures used for the experiments were generated by linearly mixing samples of isolated notes from the RWC database [4]. Three different types of mixtures



(a) Good match: piano tracks versus piano model.

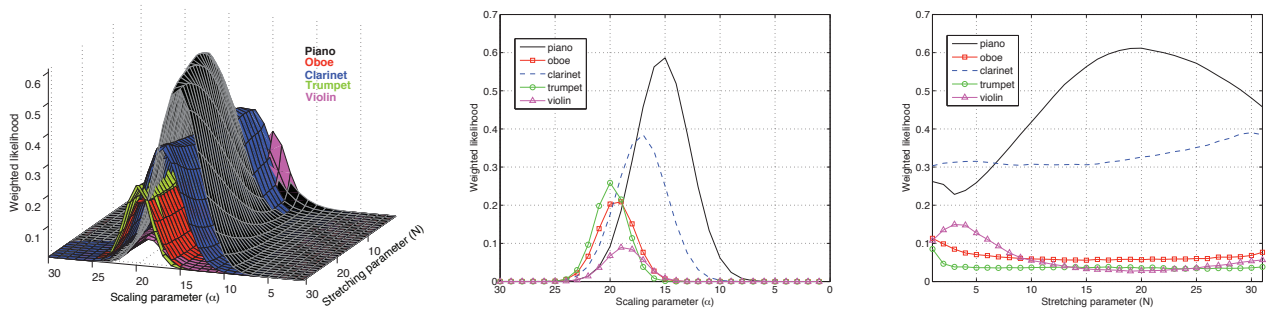


(b) Bad match: piano tracks versus oboe model.

Fig. 2. Examples of matches between track groups (solid black lines) and prototype envelopes.

were generated: simple, consonant mixtures consisting of one single note per instrument separated by consonant intervals (fifths, fourths, thirds, major and minor sixths), simple, dissonant mixtures with single notes separated by dissonant intervals (major and minor seconds, augmented fourths, major and minor sevenths), and sequences of more than one note per instrument, containing both consonant and dissonant interval relationships. Predominantly dissonant mixtures are expected to be easier to classify than predominantly consonant ones, because of the higher degree of partial overlaps of the latter. For each type of mixture and polyphony level, 10 mixtures were generated for the simple-mixture experiments and 20 for the sequence experiments, making a total of 100 mixtures. The sample onsets were separated at least by one analysis frame. The training database consists of the 5 instruments mentioned before, covering 2 octaves (C4-B5) and contains 1098 samples in total. For the evaluation, the database was partitioned into separate training (66% of the database) and testing sets (33% of the database).

The classification measure chosen for the experiments was the note-by-note accuracy, i.e. the percentage of detected individual notes correctly assigned to their instrument. Table 1 shows the results for all three timbre similarity measures and all three mixture types. The likelihood approach worked better than the Euclidean distance in all cases, showing the advantage of taking into account the model variances. Using the track-wise length and frequency weighting in the likelihood clearly improves performance in the dissonant case. That is not the case, however, for high, consonant poly-



(a) Weighted likelihood optimization surfaces. (b) Amplitude scaling profile. (c) Time stretching profile.

Fig. 3. Examples of likelihood optimization results for a piano note.

Polyphony	Consonant, simple 2	3	4	**Av.**	Dissonant, simple 2	3	4	**Av.**	Sequences 2	3	**Av.**
Euclidean distance	63.14	34.71	40.23	46.03	73.81	69.79	42.33	61.98	64.66	50.64	57.65
Likelihood	66.48	**53.57**	**51.95**	**57.33**	**79.81**	57.55	56.40	64.59	63.68	**56.40**	**60.04**
Weighted likelihood	**76.95**	43.21	40.50	53.55	**79.81**	**77.79**	**61.40**	**73.00**	**65.16**	54.35	59.76

Table 1. Experimental results: instrument detection accuracy (%).

phonies. This can be explained by the fact that, in consonant intervals, it is very likely that the lowest-frequency partials of one of the notes are overlapping, and thus ignored for the matching, cancelling their proportionally more important contribution to the weighted likelihood as compared to the unweighted likelihood. In contrast, lowest partials in dissonant intervals are in fact very unlikely to overlap, and the overlapping will more commonly occur in higher frequency areas. As expected, performance decreases with increasing polyphony and is better with dissonant than with consonant mixtures. The best obtained performances were of 79.71% with 2 voices, 77.79% with 3 voices, and 61.40% for 4 voices. For the sequences, the likelihood approach again outperforms the Euclidean distance. The improvement is however less important, and the difference in accuracy between the weighted and non-weighted likelihoods is not statistically significant.

6. CONCLUSIONS AND FUTURE WORK

The proposed method for detection of instruments in monaural polyphonic mixtures focuses on the analysis of the amplitude evolution of the partials, matching it with a set of pretrained time-frequency templates. The obtained results shows the viability of such a task without requiring multipitch estimation, and the importance of a detailed assessment of the temporal evolution of the spectral envelope.

Future improvements can include the refinement of the models by a decomposition of the envelope into attack, sustain and release phases, the evaluation of other measures of timbre similarity, and the consideration of delayed or reverberant mixtures. Another improvement would be to avoid the onset separability constraint by either timbre matching of individual sinusoidal tracks or using models of mixed timbres.

7. ACKNOWLEDGEMENTS

This work was supported by the French National Agency of Research (ANR) within the RIAM project Sample Orchestrator and by the European Commission under the IST research network of excellence VISNET II of the 6th Framework Programme.

8. REFERENCES

[1] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music," in *Proc. ICASSP*, Philadelphia, USA, 2005.

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

[3] 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 Semantics of Audio Signals (LSAS)*, Athens, Greece, December 2006.

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

[5] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, September 2007.