MONAURAL SOURCE SEPARATION FROM MUSICAL MIXTURES BASED ON TIME-FREQUENCY TIMBRE MODELS

Juan José Burred and Thomas Sikora
Communication Systems Group
Technical University of Berlin, Germany
{burred,sikora}@nue.tu-berlin.de

ABSTRACT

We present a system for source separation from monaural musical mixtures based on sinusoidal modeling and on a library of timbre models trained a priori. The models, which rely on Principal Component Analysis, serve as time-frequency probabilistic templates of the spectral envelope. They are used to match groups of sinusoidal tracks and assign them to a source, as well as to reconstruct overlapping partials. The proposed method does not make any assumptions on the harmonicity of the sources, and does not require a previous multipitch estimation stage. Since the timbre matching stage detects the instruments present on the mixture, the system can also be used for classification and segmentation.

1 INTRODUCTION

Separation of a musical mixture into its sources can greatly facilitate content analysis for Music Information Retrieval purposes, and allows other applications like remixing or upmixing to a larger number of channels than the original if multitrack recordings are not available. We address separation from a single channel, which is a highly underdetermined problem that requires either strong assumptions about the nature of the sources, a fair amount of a priori information, or a combination of both.

The main assumption taken in underdetermined separation is the sparsity of the sources, which leads to the use of elaborate signal models. An example thereof is the use of Nonnegative Sparse Coding [1]. Other approaches are based on sinusoidal modeling, which allows a detailed handling of overlapping partials and is also a highly sparse model. They are based on grouping the extracted partials according to Auditory Scene Analysis cues. In [2], amplitude smoothness is modeled by performing basis decomposition on the harmonic structures and their evolution in time. In [3], spectral filtering techniques are used to resolve overlapping sinusoids. Most approaches based on sinusoidal modeling rely either on a previous multipitch estimation stage or on the knowledge of the pitch score of the mixture [2, 3].

The above methods are unsupervised (i.e., there is no training phase) and are based on generic source models. To further improve separation, statistical models of the sources can be trained beforehand on a database of isolated source samples. Examples of this supervised approach include the use of learnt spectral priors with bayesian harmonic models [4] and the derivation of templates for timbral features [5].

In the present contribution, we propose a system for the separation of sources from single-channel mixtures of musical instruments based on sinusoidal modeling and on a library of pre-trained timbre models. Since it also outputs onset/offset information and the instrument each note belongs to, it can also be used for segmentation or polyphonic instrument recognition. The timbre models are time-frequency templates that describe in detail spectral shapes and their evolution in time. In contrast to most previously existing approaches, no assumptions on harmonicity are made, which allows to separate highly inharmonic sounds or to separate chords played by a single instrument. Furthermore, no previous multipitch estimation or any kind of a priori pitch-related score is needed. Instead, separation is solely based on common onset properties of the partials, and on the analysis of the evolution in time of the spectral envelope they define. The knowledge of the number and names of the instruments is not mandatory, but will obviously increase the performance.

Figure 1 shows an overview of the proposed separation system. First, the mixture signal is subjected to sinusoidal modeling, obtaining a set of sinusoidal tracks. A simple onset detector based on identifying synchronously starting tracks then allows to select the partial tracks that are going to be matched with the timbre models in the next stage. After each common-onset group of partial tracks has been assigned to an instrument, the overlapping part of the tracks is retrieved from the models. Finally, the separated tracks are synthesized using additive synthesis.

2 TRAINING OF THE TIMBRE MODELS

The used timbre models are based on the spectral envelope and its evolution in time. Their design and training process has been described in detail in [6]. It is based on performing Principal Component Analysis (PCA) on the set of training spectral envelopes extracted from a database

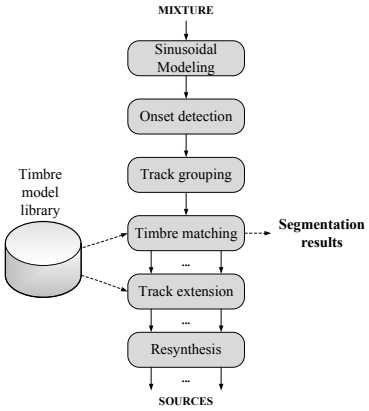


Figure 1. System overview.

of isolated notes. The final result is a set of prototype curves in a reduced-dimensional timbre space. When projected back to the t-f domain, each prototype trajectory corresponds to a prototype envelope consisting of a mean surface and a variance surface, which we will denote by $\mathbf{M}_i(k,r)$ and $\mathbf{\Sigma}_i(k,r)$, respectively, where $i = 1,\ldots,I$ is the instrument index, $k = 1,\ldots,K$ is the frequency bin index, and $r = 1,\ldots,R$ is the frame index (we will consider the same number of frames R for all models). Analogously, this can be interpreted as a Gaussian Process with parameters varying in the time-frequency plane.

3 SEGMENTATION AND SEPARATION

3.1 Sinusoidal modeling

The sinusoidal model approximates a signal as a sum of sinusoids with time-varying amplitudes, frequencies and phases. The successive stages of spectral peak picking and partial tracking are performed to obtain a frame-wise approximation to that model, yielding a triplet of amplitude, frequency and phase information $x_{pr} = (A_{pr}, f_{pr}, \theta_{pr})$, for each partial p and each time frame r. We use a standard procedure, as described in [8].

3.2 Onset detection

Sinusoidal extraction is followed by a basic onset detection stage consisting of counting the number of new tracks within a certain frame range. If $b(r)$ is a function giving the number of tracks born at frame r, we define an onset detection function $o(r)$ as a moving average of order C: $o(r) = 1/C \sum_{c=0}^{C-1} b(r-c)$. Its highest peaks are declared as the onset positions L_o^{on} for $o = 1,\ldots,O$.

3.3 Track grouping and labeling

All tracks $\mathbf{t}_t$ having its first frame within the interval $[L_o^{on} - C, L_o^{on} + C]$ for a given onset location L_o^{on} are grouped into the set $\mathbf{T}_o$. A track belonging to this set can be of one of the following types:

1. *Nonoverlapping:* if it corresponds to a new partial not present in the previous note or group of notes.
2. *Overlapping with previous track:* if its initial frequency is close, within a narrow margin, to the final frequency of a partial from the previous note or group of notes.
3. *Overlapping with synchronous track:* if it coincides in frequency, within a narrow margin, with a track belonging to the same track group $\mathbf{T}_o$.

Tracks of type 2 are easily detected, and correspondingly labeled, by searching the set $\mathbf{T}_{o-1}$ for a track fulfilling the narrow frequency margin condition. Whether the rest of tracks are of type 1 or type 3 cannot be detected at this point of the system. Tracks of type 2 and 3 can be furthermore classified as resulting from overlaps between partials belonging to the same or different instruments. Whether a track of type 2 corresponds to the same or to different instruments is irrelevant for our purposes since the corresponding notes will be segmented and separated anyway. On the contrary, tracks of type 3 belonging to the same instrument will be left intact without separation in order to detect same-instrument chords as belonging to a single source (note that our goal is different than that of transcription, which would require to detect the constituent notes of the chord). Currently, type 3 tracks from different instruments cannot be currently reliably separated, and thus the system will not support separation of notes from different instruments and exactly the same onsets. Thus, all tracks of types 1 and 3 are considered as belonging to the same source. Finally, the offset L_o^{off} corresponding to a given onset L_o^{on} is declared as the last frame of the longest partial of group $\mathbf{T}_o$.

3.4 Timbre detection

The timbre detection stage matches each one of the onset-related track groups $\mathbf{T}_o$ with each one of the prototype envelopes derived from the timbre models, and selects the instrument corresponding to the highest match. As measure of timbre similarity between the track group $\mathbf{T}_o$ and the model formed by parameters $\boldsymbol{\theta}_i = (\mathbf{M}_i, \mathbf{\Sigma}_i)$, we use the following likelihood function:

$$L(\mathbf{T}_o|\boldsymbol{\theta}_i) = \prod_{t,r} p\left(A_{tr}|\mathbf{M}_i(f_{tr}), \mathbf{\Sigma}_i(f_{tr})\right) \qquad (1)$$

where A_{tr} and f_{tr} are the amplitude and frequency, respectively, on the r-th frame of the t-th track $\mathbf{t}_t \in \mathbf{T}_o$, $p(x)$ denotes a unidimensional Gaussian distribution, and $\mathbf{M}_i(f_{tr})$ and $\mathbf{\Sigma}_i(f_{tr})$ denote the evaluation of each parameter matrix at the frequency point f_{tr}. In order to obtain the quantities $\mathbf{M}_i(f_{tr})$ and $\mathbf{\Sigma}_i(f_{tr})$ for each data point, the model frames closest in time to the input frames are chosen, and the corresponding values for the mean and the variance are linearly interpolated from neighboring data.

In order to deal with amplitude and temporal scaling uncertainty, a 2-D parameter search must be performed to

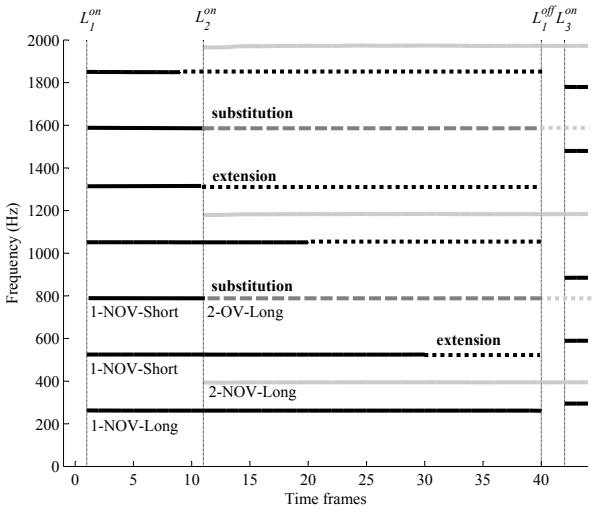


Figure 2. Track extension and substitution.

find the best matches. Amplitude scaling is denoted by term α and time scaling is performed by stretching the partial tracks towards the offset. Then, the finally used likelihood function becomes:

$$L(\mathbf{T}_o|\boldsymbol{\theta}_i) = \max_{\alpha,N} \prod_{t,r} p\left(A_{tr}^N + \alpha|\mathbf{M}_i(f_{tr}^N), \boldsymbol{\Sigma}_i(f_{tr}^N)\right) \tag{2}$$

where 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.

3.5 Track extension and substitution

Once a track group $\mathbf{T}_o$ has been declared as produced by instrument i, the corresponding prototype envelope means $\mathbf{M}_i$ are used for the two following purposes:

1. Tracks of type 1 or 3 that are shorter than the current note (which can either result from a partial amplitude approximating the noise threshold and thus remaining undetected or by the imminent appearance of a partial from the next onset group overlapping with it) are extended towards the offset by selecting the appropriate frames from $\mathbf{M}_i$ and linearly interpolating the amplitudes at the mean frequency of the remainder of the track. The amplitudes retrieved from the model are scaled so that the amplitude transition between original and extended sections of the partial is smooth.

2. Overlapping tracks of type 2 are retrieved from the model in their entirety by interpolating the model at the frequency support of the track. If the track is shorter than the note, it is again extended using the same procedure as above.

Figure 2 shows an example of the results of the track extension block.The OV labels indicate overlapping tracks of type 2 and NOV means nonoverlapping tracks (type 1). Short, nonoverlapping partials are extended to the offset (marked by *extension*) and overlapping tracks of the second offset are marked by *substitution*. Note that any region marked as *substitution* additionally implies an extension of the nonverlapping tracks from the previous onset.

Polyphony	2	3
1. Intervals / arpeggios	8.95	5.38
2. Sequences	3.17	2.26

Table 1. Results from experiments 1 and 2: Spectral Source-to-Residual Ratios (SSRR) in dB.

4 EXPERIMENTS AND RESULTS

Since the phases of the sinusoids are not preserved by the separation algorithm, the evaluation is done in the t-f domain. We use a spectral signal-to-residual ratio (SSRR):

$$SSRR = 10 \log \frac{\sum_{k,r} |S(k,r)|^2}{\sum_{k,r} (|S(k,r)| - |\hat{S}(k,r)|)^2} \tag{3}$$

where $S(k,r)$ and $\hat{S}(k,r)$ are respectively the spectrograms of the original and separated sources. Although the SSRR is a separation quality measure, it should be noted that it will also reflect errors of any other part of the system, like timbre detection or onset/offset definition. A note being classified with the wrong instrument will have its overlapping partials extracted from the wrong model, and thus will decrease separation quality. All samples used for the experiments were extracted from the RWC Musical Instrument sound database [7]. A selection of separation audio examples is available on line [1] .

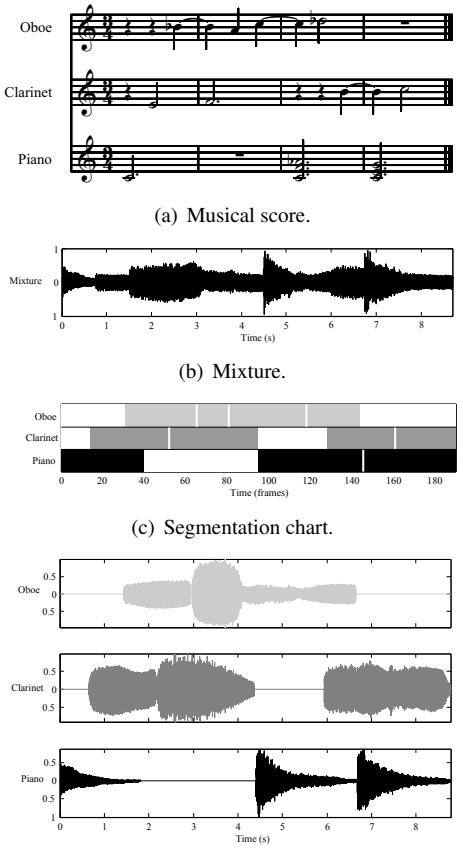
4.1 Experiment 1: Intervals and Arpeggios

For the first, simplest evaluation test, we consider mixtures of single notes from different instruments playing 2-note intervals or 3-note arpeggios. For each case, 10 different mixtures were generated. Here, the instruments contained in the mixture were considered unknown and a library of 5 instrument models (piano, oboe, clarinet, trumpet and violin), trained over the 4th octave, was used. Table 1 shows the results of averaging all SSRR values for each detected source and for all mixtures.

4.2 Experiment 2: Note sequences

For the second test we used sequences of up to 4 consecutive notes played by each instrument. This situation is more demanding, since all notes from each instrument need to be correctly classified in order to be grouped into the same track. 5 mixtures with different melodies and with up to 3 simultaneous instruments were generated. For this test, the instruments were known a priori.

[1] www.nue.tu-berlin.de/research/projects/sourcesep/sepmodels/



(a) Musical score.

(b) Mixture.

(c) Segmentation chart.

(d) Separated sources.

Figure 3. Example of separation including chords.

4.3 Experiment 3: Sequences including chords

Our final test involves more complex sequences that can contain common-onset chords played by a single instrument. Several test sequences with up to 3 simultaneous instruments were generated. Figure 3 shows an example of segmentation and separation of a sequence by three instruments (piano, clarinet and oboe), in which piano chords appear. In this case, no systematic evaluation has been performed to the current time. This and other audio examples corresponding to this experiment can be found on the web page mentioned above.

5 CONCLUSIONS

A method has been proposed that allows separation of musical instrument sounds from a single-channel mixture without making any assumptions on harmonicity, and without a previous multipitch estimation stage. This makes the extraction of the notes and chords played by a single instrument possible without pitch-related a priori information. The system is based on a stored library of probabilistic timbre models describing the characteristic behavior of each instrument's spectral envelope in time and in frequency. Experiments using mixtures of up to 3 notes from up to 5 instruments, including mixtures with single-instrument chords, have been shown to demonstrate the viability of the method.

The main limitation of the system is that notes with common onsets played by different instruments cannot be currently separated. A first direction towards solving this was to match the tracks to the timbre models individually, rather than in common-onset groups, and declaring an onset group as a mixture of two instruments if the individual track classification result was spread across the corresponding classes. Although some success has been obtained using this approach, it still lacks robustness. In our future research, the present system will be extended to stereo mixtures, so that the additionally available spatial information will allow to detect common onset notes from different instruments.

6 ACKNOWLEDGEMENTS

Part of this research was performed at the Analysis/Synthesis team, IRCAM, Paris. The research work leading to this paper has been supported by the European Commission under the IST research network of excellence K-SPACE of the 6th Framework Programme.

7 REFERENCES

[1] Virtanen, T. "Separation of Sound Sources by Convolutive Sparse Coding", *Proc. ISCA SAPA Workshop*, Jeju, Korea, 2004.

[2] Virtanen, T. "Algorithm for the Separation of Harmonic Sounds with Time-Frequency Smoothness Constraint", *Proc. DAFX*, London, UK, 2003.

[3] Every, M. R. and Szymanski, J. E. "Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics", *IEEE Trans. on Audio, Speech, and Language Processing*, 2006.

[4] Vincent, E. and Plumbley, M. D. "Single-Channel Mixture Decomposition Using Bayesian Harmonic Models", *Proc. Int. Conference on Independent Component Analysis and Blind Source Separation*, Charleston, USA, 2006.

[5] Kinoshita, T., Sakai, S. and Tanaka, H. "Musical Sound Source Identification Based on Frequency Component Adaptation", *Proc. IJCAI CASA Workshop*, Stockholm, Sweden, 1999.

[6] Burred, J. J., Röbel, A. and Rodet, X. "An Accurate Timbre Model for Musical Instruments and its Application to Classification", *Proc. Workshop on Learning the Semantics of Audio Signals*, Athens, Greece, 2006.

[7] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R. "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", *Proc. ISMIR*, Baltimore, USA, 2003.

[8] Serra, X. "Musical Sound Modeling with Sinusoids plus Noise", in Roads, C., Pope, S., Picialli, A. and De Poli, G. (Eds.), *Musical Signal Processing*, Swets & Zeitlinger, 1997.