POLYPHONIC INSTRUMENT RECOGNITION USING SPECTRAL CLUSTERING

Luis Gustavo Martins
Telecommunications and Multimedia Unit
INESC Porto
Porto, Portugal
lmartins@inescporto.pt

Juan José Burred
Communication Systems Group,
Technical University of Berlin,
Berlin, Germany
burred@nue.tu-berlin.de

George Tzanetakis, Mathieu Lagrange
Computer Science Department,
University of Victoria
Victoria, BC, Canada
[gtzan, lagrange]@uvic.ca

ABSTRACT

The identification of the instruments playing in a polyphonic music signal is an important and unsolved problem in Music Information Retrieval. In this paper, we propose a framework for the sound source separation and timbre classification of polyphonic, multi-instrumental music signals. The sound source separation method is inspired by ideas from Computational Auditory Scene Analysis and formulated as a graph partitioning problem. It utilizes a sinusoidal analysis front-end and makes use of the normalized cut, applied as a global criterion for segmenting graphs. Timbre models for six musical instruments are used for the classification of the resulting sound sources. The proposed framework is evaluated on a dataset consisting of mixtures of a variable number of simultaneous pitches and instruments, up to a maximum of four concurrent notes.

1 INTRODUCTION

The increasing quantity of music titles available in digital format added to the huge amount of personal music storage capacity available today has resulted in a growing demand for more efficient and automatic means of indexing, searching and retrieving music content. The computer identification of the instruments playing in a music signal can assist the automatic labeling and retrieval of music.

Several studies have been made on the recognition of musical instruments on isolated notes or in melodies played by a single instrument. A comprehensive review of those techniques can be found in [1]. However, the recognition of musical instruments in multi-instrumental, polyphonic music is much more complex and presents additional challenges. The main challenge stands from the fact that tones from performing instruments can overlap in time and frequency. Therefore, most of the isolated note recognition techniques that have been proposed in the literature are inappropriate for polyphonic music signals. Some of the proposed techniques for the instrument recognition on polyphonic signals consider the entire audio mixture, avoiding any prior source separation [2, 3]. Other approaches are based on the separation of the playing sources, requiring the prior knowledge or estimation of the pitches of the different notes [4, 5]. However, robustly extracting the fundamental frequencies in such multiple pitch scenarios is difficult.

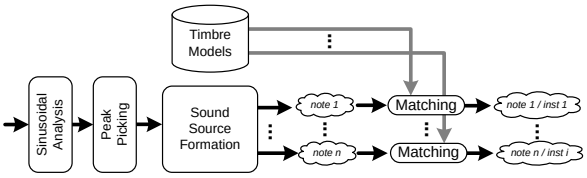


Figure 1. System diagram block.

In this paper, we propose a framework for timbre classification of polyphonic, multi-instrumental music signals using automatically separated sound sources. Figure 1 presents a block-diagram of the complete system. It starts by taking a single-channel audio signal and uses a sinusoidal analysis front-end for estimating the most prominent spectral peaks over time. The detected spectral peaks are then grouped into clusters according to cues inspired from Computational Auditory Scene Analysis (i.e. frequency, amplitude and harmonic proximity) and formulated as a graph partitioning problem. The normalized cut, a technique from the Computer Vision field, is then used as a global criterion for segmenting graphs. Contrary to other approaches [6, 7], this source separation technique does not require any prior knowledge or pitch estimation.

As demonstrated in previous works by the authors [8, 9] and later in section 4, the resulting clusters capture reasonably well the underlying sound sources and events (i.e. notes, in the case of music signals) present in the audio mixture. After the sound source separation stage, each identified cluster is matched to a collection of six timbre models namely piano, oboe, clarinet, trumpet, violin and alto sax. These models are a compact description of the spectral envelope and its evolution in time, and were previously trained using isolated note audio recordings. The

design of the models, as well as their application to isolated note classification, were described in [10].

The outline of the paper is as follows. In section 2 we describe the sound source separation technique, which starts from a sinusoidal representation of the signal followed by the application of the normalized cut for source separation. In section 3 we briefly describe the training of the timbre models and focus on the matching procedure used to classify the separated clusters. We then evaluate the system performance in section 4 and close with some final conclusions.

2 SOUND SOURCE SEPARATION

Computational Auditory Scene Analysis (CASA) systems aim at identifying perceived sound sources (e.g. notes in the case of music recordings) and grouping them into auditory streams using psycho-acoustical cues [11]. However, as remarked in [6] the precedence rules and the relevance of each of those cues with respect to a given practical task is hard to assess. Our goal is to use a flexible framework where these perceptual cues can be expressed in terms of similarity between time-frequency components. The separation task is then carried out by clustering components which are close in the similarity space (see Figure 2). Once identified, those clusters will be matched to timbre models in order to perform the instrument identification task.

2.1 Sinusoidal Modeling

Most CASA approaches consider auditory filterbanks and/or correlograms as their front-end [12]. In these approaches the number of time-frequency components is relatively small. However closely-spaced components within the same critical band are hard to separate. Other approaches [6, 13] consider the Fourier Spectrum as their front-end. In these approaches, in order to obtain sufficient frequency resolution a large number of components is required. Components within the same frequency region can be pre–clustered together according to a stability criterion computed using statistics over the considered region. However, this approach has the drawback of introducing another clustering step, and opens the issue of choosing the right descriptors for those pre-clusters. Alternatively, a sinusoidal front-end is helpful to provide meaningful and precise information about the auditory scene while considering only a limited number of components, and is the representation we consider in this work.

Sinusoidal modeling aims to represent a sound signal as a sum of sinusoids characterized by amplitudes, frequencies, and phases. A common approach is to segment the signal into successive frames of small duration so that the stationarity assumption is met. For each frame, the local maxima of the power spectrum are identified and a bounded set of sinusoidal components is estimated selecting the peaks with the highest amplitudes.

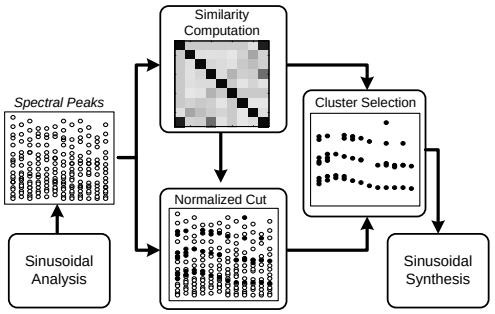


Figure 2. Block-Diagram of the Sound Source Separation algorithm.

The discrete signal $x_k(n)$ at frame index k is then modeled as follows:

$$x_k(n) = \sum_{l=1}^{L_k} a_{lk} \cos\left(\frac{2\pi}{F_s} f_{lk} \cdot n + \phi_{lk}\right) \quad (1)$$

where F_s is the sampling frequency and ϕ_{lk} is the phase at the beginning of the frame of the l-th component of L_k sine waves. The f_l and a_l are the frequency and the amplitude of the l-th sine wave, respectively, both of which are considered as constant within the frame. For each frame k, a set of sinusoidal parameters $\mathcal{S}_k = \{p_{1k}, \cdots, p_{L_k k}\}$ is estimated. The system parameters of this Short-Term Sinusoidal (STS) model $\mathcal{S}_k$ are the L_k triplets $p_{lk} = \{f_{lk}, a_{lk}, \phi_{lk}\}$, often called *peaks*.

2.2 Spectral Clustering

In order to simultaneously optimize partial tracking and source formation, we construct a graph over the entire duration of the sound mixture. Unlike approaches based on local information [14], we utilize the global normalized cut criterion to partition the graph (spectral clustering). This criterion has been successfully used for image and video segmentation [15]. In our perspective, each partition is a set of peaks that are grouped together such that the similarity within the partition is minimized and the dissimilarity between different partitions is maximized. By appropriately defining the similarity between peaks a variety of perceptual grouping cues can be used.

The edge weight connecting two peaks p_{lk} and $p_{l'k'}$ (k is the frame index and l is the peak index) depends on the proximity of frequency, amplitude and harmonicity:

$$W(p_{lk}, p_{l'k'}) = W_f(p_{lk}, p_{l'k'}) \cdot W_a(p_{lk}, p_{l'k'}) \cdot W_h(p_{lk}, p_{l'k'}) \quad (2)$$

where W_x are typically radial basis functions of distance among the two peaks in the x axis. For more details see [8, 9].

Most existing approaches that apply the Ncut algorithm to audio [16] consider the clustering of components over one analysis frame only. However, the time integration (i.e. partial tracking) is as important as the frequency one

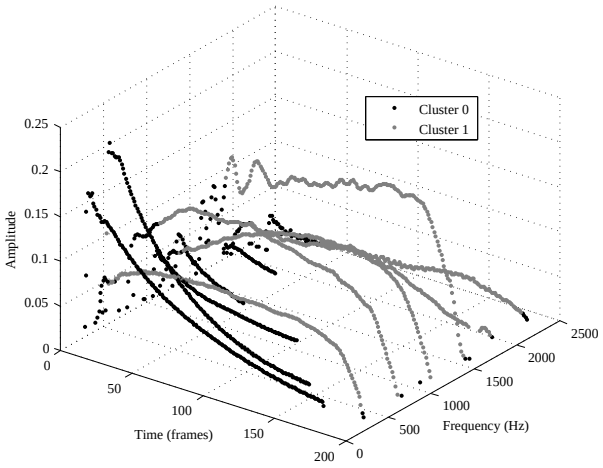


Figure 3. Resulting sound source formation clusters for two notes played by a piano and an oboe (E4 and B4, respectively).

(i.e. source formation) and should be carried out at the same time. We therefore consider the sinusoidal components extracted within the entire mixture as proposed in [8]. We considered a maximum of 20 sinusoids per frame which are 46 ms long, using a hop size of 11 ms.

Figure 3 depicts the result of the sound source separation using the normalized cut for a single-channel audio signal with mixture of two notes (E4 and B4 [1], same onset, played by a piano and an oboe, respectively). Each dot corresponds to a peak in the time-frequency space and the different coloring reflects the cluster to which it belongs (i.e. its source).

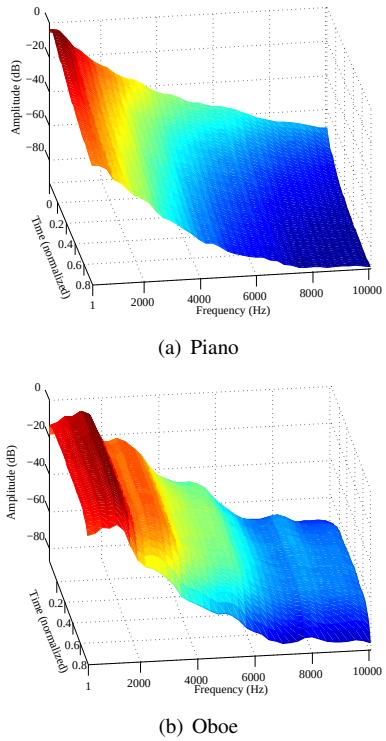
3 TIMBRE IDENTIFICATION

3.1 Timbre Models

Once each single-note cluster of sinusoidal parameters has been extracted, it is classified into an instrument from a predefined set of six: piano (**p**), oboe (**o**), clarinet (**c**), trumpet (**t**), violin (**v**) and alto sax (**s**). The method models each instrument as a set of time-frequency templates, one for each instrument. The template describes the typical evolution in time of the spectral envelope of a note. The spectral envelope is an appropriate representation to generate features to analyze sounds described by sinusoidal modeling, since it matches the salient peaks of the spectrum, i.e., the amplitudes a_{lk} of the partials.

The training process consists of arranging the training dataset as a time-frequency matrix $\mathbf{X}(g,k)$ of size $G \times K$, where g is the frequency bin index and k is the frame index, and performing spectral basis decomposition upon it using Principal Component Analysis (PCA). This yields a factorization of the form $\mathbf{X} = \mathbf{BC}$, where the columns of the $G \times G$ matrix $\mathbf{B}$ are a set of spectral basis sorted in decreasing order of contribution to the total variance, and $\mathbf{C}$ is the $G \times K$ matrix of projected coefficients. By keeping a reduced set of $R < G$ basis, we obtain both a reduction of the data needed for a reasonable approximation and, more importantly for our purpose, a representation based only on the most essential spectral shapes.

[1] Throughout this paper we use the convention A4 = 440Hz.



(a) Piano

(b) Oboe

Figure 4. Examples of prototype envelopes for a range of one octave.

Having as goal a pitch-independent classification, the time-frequency templates should be representative for a wide range of notes. In the training process, notes from several pitches must be considered to obtain a single model. The training samples are subjected to sinusoidal modeling, and arranged in the data matrix $\mathbf{X}$ by linearly interpolating the amplitude values to a regular frequency grid defined at the locations of the G bins. This is important for appropriately describing formants, which are mostly independent of the fundamental frequency.

The projected coefficients of each instrument in the R-dimensional PCA space are summarized as a prototype curve by interpolating the trajectories corresponding to the individual training samples at common time points and point-wise averaging them. When projecting back into the time-frequency domain by a truncated inverse PCA, each P^i-point prototype curve will correspond to a $G \times P^i$ prototype envelope $\mathbf{M}^i(g,k)$ for instrument i. We consider the same number of time frames $P = P_i$ for all instrument models. Figure 4 shows the obtained prototype envelopes for the fourth octave of a piano and of an oboe.

Depending on the application, it can be more convenient to perform further processing on the reduced-dimensional PCA space or back in the time-frequency domain. When classifying individual notes, a distance

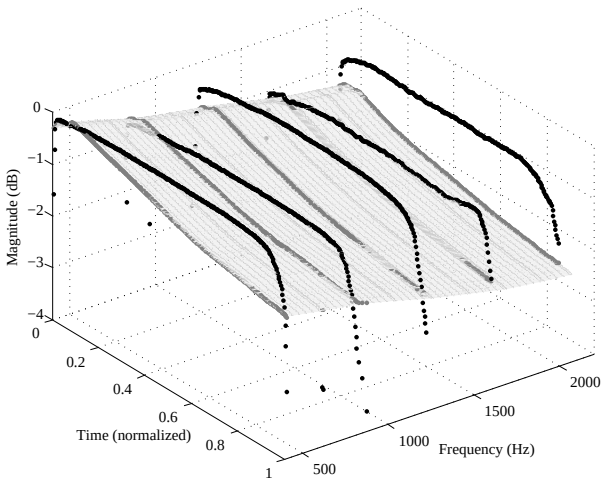


Figure 5. Weak matching of an alto sax cluster and a portion of the piano prototype envelope.

measure between unknown trajectories and the prototype curves in PCA space has proven successful [10]. In the current source separation application, the clusters to be matched to the models can contain regions of unresolved overlapping partials or outliers, which can introduce important interpolation errors when adapted to the G-bin frequency grid needed for projection onto the bases. This makes working in the time-frequency domain more convenient in the present case.

3.2 Timbre Matching

Each one of the clusters obtained by the sound source separation step is matched against each one of the prototype envelopes. Let us denote a particular cluster of K frames represented as an ordered set of amplitude and frequency vectors $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_K)$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_K)$ of possibly differing lengths $L_1, \ldots, L_K$.

We need to evaluate the prototype envelope of model i at the frequency support of the input cluster j. This operation is denoted by $\tilde{\mathbf{M}}^{ij} = \mathbf{M}^i(\mathbf{F}^j)$. To that end, the time scales of both input and model are first normalized. Then, the model frames closest to each one of the input frames in the normalized time scale are selected. Finally, each new amplitude value $\tilde{m}_{lk}^{ij}$ is linearly interpolated from the neighboring amplitude values of the selected model frame.

We then define the distance between a cluster j and an interpolated prototype envelope i as

$$d(\mathbf{A}^j, \tilde{\mathbf{M}}^{ij}) = \frac{1}{K^j} \sum_{k=1}^{K^j} \sqrt{\sum_{l=1}^{L_k^j} (a_{lk}^j - \tilde{m}_{lk}^{ij})^2} \qquad (3)$$

i.e., the average of the Euclidean distances between frames of the input clusters and interpolated prototype envelope at the normalized time scale. The model $\tilde{\mathbf{M}}^{ij}$ minimizing this distance is chosen as the predicted instrument for classification. Figure 5 shows an attempt to match a cluster extracted from an alto sax note and the corresponding section of the piano prototype envelope. As it is clearly visible, this weak match results in a high distance value.

True instruments						*classified*
p	**o**	**c**	**t**	**v**	**s**	*as*
100	0	0	0	0	0	**p**
0	**100**	8	8	0	0	**o**
0	0	**67**	0	33	0	**c**
0	0	0	**92**	0	8	**t**
0	0	0	0	**58**	8	**v**
0	0	25	0	8	**83**	**s**

Table 1. Confusion matrix for single–note instrument identification. We considered 6 different instruments from the RWC database: piano (**p**), oboe (**o**), clarinet (**c**), trumpet (**t**), violin (**v**), alto sax (**s**).

4 EXPERIMENTS

The current framework implementation does still not fully take into consideration timing information and continuity issues, such as note onsets and durations. Given so, we will limit the evaluation procedure to the separation and classification of concurrent notes sharing the same onset and played from different instruments.

The evaluation dataset was artificially created mixing audio samples of isolated notes of piano, oboe, clarinet, trumpet, violin and alto sax, all from the RWC Music Database [17]. The training dataset used to derive the timbre models for each instrument (see Section 3) is composed of audio samples of isolated notes, also from the RWC Music Database. However, in order to get meaningful timbre recognition results, we used independent instances of each instrument for the evaluation dataset and for the training dataset. Ground-truth data was also created for each mixture and includes information about the notes played and the corresponding instrument. Given that the timbre models used in this work showed good results for a range of about two octaves [10], we constrained the notes used for evaluation to the range C4 to B4. Furthermore, for simplicity's sake, we have only considered notes with a fixed intensity in this evaluation.

4.1 Timbre identification for single note signals

We started by evaluating the performance of the timbre matching block (as discussed in Section 3.2) for the case of isolated notes coming from each of the six instruments modeled. This provides a base-ground with which will be possible to compare the ability of the framework to classify notes separated from mixtures. For the case of isolated notes, the sound source separation block reduces its action to just performing sinusoidal analysis, since there are no other sources to be separated. This basically only results in the loss of the non-harmonic residual, which although not irrelevant to timbre identification, has been demonstrated to have a small impact in the classification

	2-note			3-note			4-note			total		
	RCL	**PRC**	**F1**	**RCL**	**PRC**	**F1**	**RCL**	**PRC**	**F1**	**RCL**	**PRC**	**F1**
p	83	100	91	22	100	36	0	0	0	23	100	38
o	100	75	86	100	46	63	67	40	50	86	50	63
c	33	100	50	33	100	50	40	86	55	36	93	52
t	89	100	94	58	100	74	58	64	61	67	85	75
v	67	67	67	83	45	59	83	36	50	80	43	56
s	100	43	60	67	60	63	60	75	67	67	62	64
total	**75**	**79**	**77**	**56**	**64**	**59**	**46**	**56**	**50**	**56**	**64**	**60**

Table 2. Recall and precision values for instrument presence detection in multiple-note mixtures.

performance [18]. Table 1 presents the confusion matrix for the instrument classification for a dataset of 72 isolated notes, ranging from C4 to B4, from each one of the six considered instruments. The system presents an overall classification accuracy of 83.3%, being violin and clarinet the instruments posing the biggest difficulties.

4.2 Instrument presence detection in mixtures of notes

We then evaluated the ability of the system to separate and classify the notes from audio files with up to 4 simultaneously sounding instruments. A combination of 54 different instruments and mixtures of 2-, 3- and 4-notes was created (i.e. 18 audio files for each case).

The first and simplest evaluation we performed was to test the system ability to detect the presence of an instrument in a mixture of up to 4 notes. In this case it was just a matter of matching each one of the six timbre models with all the separated clusters and counting the *true* and *false positives* for each instrument. A *true positive (TP)* is here defined as the number of separated clusters correctly matched to an instrument playing in the original mixture (such information is available in the dataset ground-truth). A *false positive (FP)* can be defined as the number of clusters classified as an instrument not present in the original audio mixture. Given these two values, it is then possible to define three performance measures for each instrument - *Recall (RCL)*, *Precision (PRC)* and *F-Measure (F1)*:

$$RCL = \frac{TP}{COUNT} \quad PRC = \frac{TP}{TP + FP} \tag{4}$$

$$F1 = \frac{2 \times RCL \times PRC}{RCL + PRC} \tag{5}$$

where $COUNT$ is the total number of instances of an instrument over the entire dataset (i.e. the total number of notes it plays). As shown in Table 2, the system was able to correctly detect 56% of the occurrences of instruments in mixtures of up to 4 notes, with a precision of 64%. Piano appears as the most difficult timbre to identify, specifically for the case of 4-note mixtures, where from the existing 15 notes playing in the dataset, none was correctly detected as coming from that instrument. As anticipated, the system performance degrades with the increase of the number of concurrent notes. Nevertheless, it was still possible to retrieve 46% of the present instruments in 4-note mixtures, with a precision of 56%.

4.3 Note separation and timbre identification in mixtures of notes

Although informative, the previous evaluation has a caveat – it does not allow to precisely verify if a separated and classified cluster does in fact correspond to a note played with the same instrument in the original audio mixture. In order to fully assess the separation and classification performance of the framework, we tried to make a correspondence between each separated cluster and the notes played in the mix (available in the ground-truth).

A possible way to obtain such a correspondence is by estimating the pitch of each one of the detected clusters, using a simple technique. For each cluster we calculated the histogram of peak frequencies. Since the audio recordings of the instruments used in this evaluation are from notes with steady pitch over time (i.e. no vibrato, glissandos or other articulations), the peaks on the histogram provide a good indication of the frequencies of the strongest partials. Having the set of the strongest partial frequencies, we then performed another histogram of the differences among all partials and selected the highest mode as the best F0 candidate for that cluster.

Given these pitch correspondences, it is now possible to check the significance of each separated cluster as a good note candidate, as hypothesized in Section 1. For the entire dataset, which includes a total of 162 notes from all the 2-, 3- and 4-note audio mixtures, the system was able to correctly establish a pitch correspondence for 55% of the cases (67%, 57% and 49% for the 2-, 3- and 4-note mixtures, respectively). These results can not however be taken as an accurate evaluation of the sound source separation performance, as they are influenced by the accuracy of the pitch estimation technique.

The results in Table 3 show the correct classification rate for all modeled instruments and multiple-note scenarios, excluding the clusters whose correspondence was not possible to establish. This allows decoupling the source separation/pitch estimation performance from the timbre identification accuracy. Table 3 shows a correct identification rate of 47% of the separated notes overall, diminishing sharply its accuracy with the increase of concurrent notes in the signal. This shows the difficulties posed by the overlap of spectral components from different notes/instruments into a single detected cluster.

	Instrument Detection Rate			
	2-note	**3-note**	**4-note**	**overall**
p	67	67	0	55
o	100	86	60	81
c	33	29	19	26
t	75	33	22	43
v	67	100	50	75
s	75	36	42	44
total	65	50	33	**47**

Table 3. Instrument classification performance for 2-, 3- and 4-note mixtures.

5 DISCUSSION

We proposed a framework for the sound source separation and timbre classification of single-channel polyphonic music played by a mixture of instruments. Although using a constrained scenario, the experiments show the potential of the system to achieve sound source separation and identification of music instruments using timbre models. We plan on extending this framework for the analysis of continuous music by taking into consideration prior time segmentation of the music notes, based on their onsets and durations. This will allow us to deal with more realistic scenarios and to compare the proposed approach with other state-of-the-art systems.

Furthermore, the proposed framework is versatile and flexible enough to include new features at a later stage that may allow overcoming some of its current limitations. The use of timbre models as a-priori information at the sound source separation stage will be an interesting topic of future research. The extraction of new and more descriptors directly from the estimated cluster parameters (e.g. pitch, timbre features, timing information, etc.) will allow the development of innovative applications for the automatic analysis and sophisticated processing of real-world polyphonic music signals.

6 ACKNOWLEDGMENTS

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

7 REFERENCES

[1] P. Herrera, G. Peeters, and S. G. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–22, 2003.

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

[3] A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Int. Conf. on Digital Audio Effects (DAFx)*, Naples, Italy, 2004.

[4] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, , no. 27, 1999.

[5] 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.

[6] E. Vincent, "Musical source separation using time-frequency priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), 2006.

[7] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Proc. ICASSP*, 2003.

[8] M. Lagrange and G. Tzanetakis, "Sound source tracking and formation using normalized cuts," in *Proc. ICASSP*, Honolulu, USA, 2007.

[9] M. Lagrange, L.G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for singing voice separation and melody extraction," *submitted to the IEEE Trans. on Acoustics, Speech, and Signal Processing (Special Issue on MIR)*, 2007.

[10] J. J. Burred, A. Robel, and X. Rodet, "An accurate timbre model for musical instruments and its application to classification," in *Workshop on Learning the Semantics of Audio Signals*, Athens, Greece, 2006.

[11] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[12] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley, 2006.

[13] S.H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics based audio separation," in *Proc. ICASSP*, 2003, vol. 5, pp. v–640 – v–643.

[14] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34(4), pp. 744–754, 1986.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 888–905, 2000.

[16] S.H. Srinivasan, "Auditory blobs," in *Proc. ICASSP*, 2004, vol. 4, pp. iv–313 – iv–316.

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

[18] A. Livshin and X. Rodet, "The importance of the non-harmonic residual," in *AES 120th Convention*, Paris, France, 2006.