A COMPARISON OF SOUND SEGREGATION TECHNIQUES FOR PREDOMINANT INSTRUMENT RECOGNITION IN MUSICAL AUDIO SIGNALS

Juan J. Bosch, Jordi Janer, Ferdinand Fuhrmann and Perfecto Herrera
Universitat Pompeu Fabra, Music Technology Group, Roc Boronat 138, Barcelona
juanjo.bosch@gmail.com, jordi.janer@upf.edu,
ferdinand.fuhrmann@gmail.com, perfecto.herrera@upf.edu

ABSTRACT

The authors address the identification of predominant music instruments in polytimbral audio by previously dividing the original signal into several streams. Several strategies are evaluated, ranging from low to high complexity with respect to the segregation algorithm and models used for classification. The dataset of interest is built from professionally produced recordings, which typically pose problems to state-of-art source separation algorithms. The recognition results are improved a 19% with a simple sound segregation pre-step using only panning information, in comparison to the original algorithm. In order to further improve the results, we evaluated the use of a complex source separation as a pre-step. The results showed that the performance was only enhanced if the recognition models are trained with the features extracted from the separated audio streams. In this way, the typical errors of state-of-art separation algorithms are acknowledged, and the performance of the original instrument recognition algorithm is improved in up to 32%.

1. INTRODUCTION

The amount of music available has dramatically increased in recent years. There is thus a clear need of effectively organizing and retrieving this content. Music Information Retrieval (MIR) is a research field dealing with the extraction of music content information, and can be used for such purposes. Instrumentation is a very useful description of musical data, since it can be exploited successfully in different forms; songs can be retrieved using the information about the presence of an instrument, and the identification of the musical genre is easier when knowledge about the instrumentation is available (e.g. a banjo makes the piece more likely to be country than classical music). Additionally, instrumentation is a key aspect for the perceived similarity in music [1].

Audio source separation deals with the recuperation of the original signals from the acoustical sources constituting an audio mixture by computational means. Even though there is still much room for improvement when applied to real world music, state-of-art separation algorithms can be used to, at least, increasing the presence of a source or a group of sources in a mixture, such as harmonic-percussive separation [10]. They can potentially be a useful pre-step to improve the results of MIR tasks, such as chord detection, melody extraction, etc.

The automatic recognition of instruments is usually based on timbre models or features such as MFCCs or MPEG-7 combined with statistical classifiers. An extensive review of approaches for isolated musical instrument classification can be found in [8], with several classification techniques, a number of instrumental categories below ten, and accuracies that reach up to 90%.

More recent works deal with instrument recognition in polytimbral musical signals, which is a more realistic and demanding problem. For instance, Tzanetakis focused on the detection of voice [12], while Essid [4] presented an approach using a taxonomy-based hierarchical classification, in which the classifiers were trained on combinations of instruments such as: piano, tenor sax, double bass and drums. Kitahara et al. [9] proposed several techniques to improve instrument recognition in duo and trio music by dealing with three issues: the feature variations caused by sound mixtures, the pitch dependency of timbres, and the use of musical context. With the proposed techniques, they achieved an 85.8% average recognition rate in trio music. Fuhrmann [5] proposed a method for automatic recognition of predominant instruments with Support Vector Machine (SVM) classifiers trained with features extracted from real musical audio signals. One of the problems identified in this system is that it often missed some of the labels in excerpts containing more than one predominant instrument.

The recognition of the instruments present in a mixture becomes more complex as the number of instruments increases. Reducing the number of instruments in the audio to be analyzed should thus help in the recognition of instruments, and the idea of using source separation as a pre-step has already been investigated in previous research. Heittola et al. [6] use Non-Negative Matrix Factorization (NMF) with a source filter model, based on

previous work by Virtanen and Klapuri [13]. Klapuri's multipitch estimation is used in the separation, with aid of an optional streaming algorithm which organizes individual notes into sound sources. The Viterbi algorithm is then employed to find the most likely sequence of notes. The classifiers use MFCC's (with a 40 channel filter bank) along with their first time derivatives. A Gaussian Mixture Model (GMM) is used to model the instrument conditional densities of the features, and the parameters are estimated using the Expectation Maximization (EM) algorithm from the training material. A Maximum Likelihood classifier is then used for classification. The dataset was artificially created from the RWC dataset, with a maximum of six note polyphony, and 19 different pitched instruments, reaching a 59.1% F1-measure. Burred [3] also presents an instrument classification approach with a stereo blind source separation pre-step, using Gaussian likelihood as a timbre similarity measure. The reported accuracy reaches 86.7%, with polyphony of 2 instruments, and 5 classes. Results are significantly better than in monaural separation, with 79.8% accuracy.

In this paper, we address the combination of source separation and instrument recognition, in order to improve the identification of predominant instruments in professionally produced western music recordings. As opposed to audio data created by artificially mixing several instrument with no musical relation between them, this real world scenario adds more complexity to source separation algorithms, due to several reasons. First, in real world western music, instruments are harmonically related, and thus their spectral components usually share some of the frequencies. Furthermore, effects such as reverbs, delays, etc. make the separation much more difficult. On the other hand, such scenario allows the algorithms to take advantage of the spatial information present in stereophonic recordings.

2. METHOD

This section introduces the methodology proposed to investigate if the performance of an instrument recognition algorithm can be improved with a previous audio segregation step, as introduced in subsection 2.1. The dataset is described in subsection 2.2, and the evaluation methodology is introduced in subsection 2.3.

2.1 Audio segregation for instrument recognition

The algorithm used by Fuhrmann in [5] is considered as the baseline instrument recognition. It is conceived to output a set of labels corresponding to the predominant instruments in an excerpt of polytimbral music. Ten pitched instruments are used in this study: cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and additionally human singing voice. The original system uses SVM, which outputs probabilistic estimates for each of the modeled categories. As previously introduced, the main problem is that it sometimes misses some labels in excerpts with multiple instruments.

The hypothesis is that in order to enhance its performance, a previous step could be performed, separating input audio data into several streams. These streams are then separately processed by the instrument recognition algorithm, resulting in several sets of labels. The sets of labels are then combined and given as output labels. Several segregation methods are considered, as well as different strategies for the label combination, and also several models used for instrument recognition. Figure 1 illustrates the combination of a segregation process followed by the instrument recognition in each of the streams.

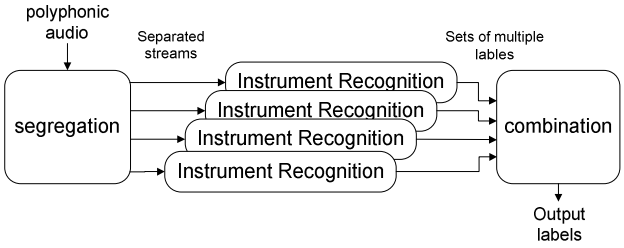

Figure 1. Generic flow diagram for the application of audio segregation as a previous step to the instrument recognition

We consider two different segregation methods. The first is FASST (A Flexible Audio Source Separation Framework), presented by Ozerov et al. [10]. It is based on structured source models, which allow the introduction of constrains according to the available prior knowledge about the separation problem. It aims at generalizing several existing source separation methods, and allows creating new ones. The second segregation method is a simple Left/Right-Mid/Side (LRMS) separation based on panning information, where M = L+R and S = L-R.

In this research, the FASST algorithm is used in a configuration which separates the polytimbral audio input into four streams: "drums", "bass", "melody", and "other" (dbmo). This is a default configuration provided with the FASST framework, and it fits our interest in the recognition of the predominant pitched instruments, as the classifier neither considers bass nor drums. After the separation, the "melody" stream would ideally contain the main instrument to be recognized, and the "other" stream would contain the rest of the instruments, with no presence of bass and percussive instruments. Recognition of the predominant instruments in these streams of audio should be easier than in the case of the original polytimbral mixture. However, there are limitations in most separation algorithms, especially when applied to real world music. They commonly create artifacts and errors in the separation, producing some leakage of instruments in streams where they should not be present. This

could affect the recognition of instruments due to the changes the artifacts produce in timbre. In order to deal with these errors, we investigate if a classifier could learn how a source separation algorithm behaves, and acknowledge the errors by training models on the separated audio estimations. In simple words, the models would learn, with the features of the estimated "drums", "bass", "melody", "other" stream, when the predominant instrument of the audio is a cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, or voice. We consider the use of different models for each of the separated streams, in order to allow the usage of a different set of (automatically selected) audio features, as well as different parameters for training the classifiers.

Finally, the strategy for the combination of the labels given as output by the individual instrument recognition models is also important. Two strategies are explored in our experiments: 1) selecting some of the classifiers' output only (e.g. only the sets of labels from the "melody" and "other" streams), and 2) requiring a degree of agreement (overlap) between all sets of labels. In the second strategy, output labels correspond to the ones present in more than N sets of labels predicted by the models.

2.2 Data

Two different datasets have been created for training and testing, based on the database originally compiled by Fuhrmann [6]. Firstly, the training dataset contains 6700 annotated excerpts of 3 seconds in which only one instrument is predominant. These data are unevenly distributed among the modeled categories, ranging from minimum 388 to a maximum of 778. A second training dataset is derived by separating the original one into dbmo streams with FASST. Secondly, the testing set consists of around 3000 excerpts annotated with one to five instruments. This set was created by dividing the original music pieces of the original database [6] into segments with the following properties: 1) the predominant instruments are the same in the whole excerpt, 2) the length is between 5 and 20 seconds, and 3) the excerpts are stereo. The first property allows us to disregard segmentation according to the instrumentation into the recognition evaluation, since the predominance of instruments typically changes amongst or even within sections of a music piece. The second property ensures that the instrument labeling process has enough information to output the labels with a certain confidence. The third property corresponds to the use case of interest: professionally produced music recordings, in stereo format.

2.3 Evaluation methodology

The evaluation method is based on comparing the output labels against the manually annotated ground-truth labels. Following the traditional information retrieval evaluation measures, we calculate: true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) for each of the instruments (labels). We consider L the closed set of labels $L = \{l_i\}$, with $i = 1...N$, N the number of instruments, and the dataset $X = \{x_i\}$, with $i = 1...M$, and M the number of excerpts. We define $\hat{Y} = \{\hat{y}_i\}$, with $i = 1...M$ as the set of ground-truth labels, and $Y = \{y_i\}$, with $i = 1...M$, and $y_i \subseteq L$, the set of predicted labels assigned to each instance i. Precision and recall are defined for each of the labels l in L as:

$$P_l = \frac{tp_l}{tp_l + fp_l} = \frac{\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{i=1}^{M} y_{l,i}} \quad (1)$$

$$R_l = \frac{tp_l}{tp_l + fn_l} = \frac{\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{i=1}^{M} \hat{y}_{l,i}} \quad (2)$$

where $y_{l,i}$ and $\hat{y}_{l,i}$ are boolean variables referring to instance i, which indicate the presence of the label l in the set of predicted labels, or in the set of ground-truth labels respectively. Additionally, we define the F1 as the harmonic mean between precision and recall:

$$F_l = \frac{2P_lR_l}{P_l + R_l} \quad (3)$$

We also define macro and micro averages of the previous metrics, in order to obtain more general performance metrics, which consider all labels. The macro is here understood as an unweighted average of the precision or recall taken separately for each label (average over labels).

$$P_{macro} = \frac{1}{|L|}\sum_{l=1}^{L} P_l\,,\; R_{macro} = \frac{1}{|L|}\sum_{l=1}^{L} R_l \quad (4)$$

On the other hand, the micro average is an average over instances, and thus, giving more weight to the labels with a higher number of instances:

$$P_{micro} = \frac{\sum_{l=1}^{L} tp_l}{\sum_{l=1}^{L}\left(tp_l + fp_l\right)} = \frac{\sum_{l=1}^{L}\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{l=1}^{L}\sum_{i=1}^{M} y_{l,i}} \quad (5)$$

$$R_{micro} = \frac{\sum_{l=1}^{L} tp_l}{\sum_{l=1}^{L}\left(tp_l + fn_l\right)} = \frac{\sum_{l=1}^{L}\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{l=1}^{L}\sum_{i=1}^{M} \hat{y}_{l,i}} \quad (6)$$

The macro and micro F1 are defined as the harmonic mean of respectively, the macro and micro averages.

$$F_{macro} = \frac{2P_{macro}R_{macro}}{P_{macro} + R_{macro}},\; F_{micro} = \frac{2P_{micro}R_{micro}}{P_{micro} + R_{micro}} \quad (7)$$

The following section details the experiments performed according to the presented methodology.

3. EXPERIMENTS

We conducted five experiments to investigate the benefits of the segregation of the audio signal into different streams prior to the application of an instrument recognition algorithm. In the first four experiments, the SVM models used for the instrument recognition were trained with parameters that optimized the performance in Experiment 1: a polynomial kernel of degree 4 and a cost parameter = 0.1. In each experiment, we consider all combinations of sets of labels to find the best recognition performance. These are notated with the initials of the streams considered, e.g.: "Exp3:dbo" refers to the combination of the labels outputted from the recognition of the d (drums) + b (bass) + o (other) streams in Experiment 3. The combination strategy was initially the union of the labels predicted by each of the models. Then, in Experiment 5 we explored a partial overlap strategy, and we optimized recognition performance by tuning the parameters for each of the models.

3.1 Experiment 1: original algorithm

The original algorithm is employed without a previous separation step, as shown in Figure 2:

Figure 2: Original instrument recognition algorithm

The labels obtained in this experiment are named "n" for "no separation". In this configuration, the stereo audio input is transformed into mono, by adding the left and right channels. We obtain the following micro averages: precision = 0.708, recall = 0.258 and F1 = 0.378.

3.2 Experiment 2: Left/Right-Mid/Side separation + original models

In this experiment, audio was segregated into four streams with l = Left, r = Right, n = l+r (Mid), and s = l-r (Side), and the original model was used for classification, as depicted in Figure 5.

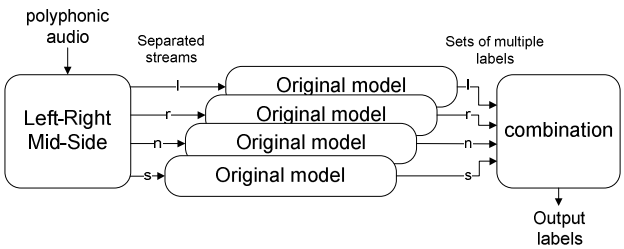

Figure 5. LRMS separation into lrns streams, used as input of the original instrument recognition models.

The label "n" is used for the "Mid" stream in order to be consistent with the notation in Experiment 1, also performed in the addition of the Left and Right channels.

Evaluation results showed that the best combination is with "Exp2:lrns", obtaining a micro F1 = 0.451. This represents an absolute improvement of 7.3 percent points in the micro F1 with respect to the original algorithm "Exp1:n", or in relative terms, a 19.3%. This is a considerable improvement, especially taking into account that this is a very simple segregation method which could even be performed in real time.

3.3 Experiment 3: FASST + original models

In this experiment, FASST separation into the bass (b), drums (d), melody (m) and other (o) streams is used, along with the original models for the instrument recognition, as shown in Figure 3.

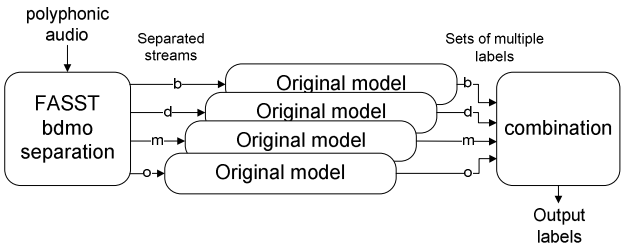

Figure 3. FASST separation into the drum, bass, melody and other streams, combined with the original instrument recognition models.

The evaluation showed that the original algorithm without source separation provides better results than any of the combinations of the labels obtained in Experiment 3. The best micro F1 (0.355) is obtained with a combination of all separated streams ("Exp3:dbmo"). In this case, recall (0.385) is better than with the original algorithm ("Exp1:n"), but precision is quite worse (0.330), so the F1-measure is lower. There is thus a decrease in performance when using source separation as a pre-step of the original instrument recognition models. This is probably due to the fact that the separation is not perfect, there is some energy of instruments in streams where there should not be present, and their timbre is modified. Additionally, the separation algorithm has the drawback of its complexity and execution time, which is above one minute per second of to-be-separated audio (Intel Core 2 Duo @ 2.4 GHz, 4 GB RAM with Windows XP – 32 bits).

3.4 Experiment 4: FASST + models trained with separated audio

In this experiment, the instrument recognition models have been trained with the dbmo audio streams obtained from separating the training dataset with FASST. Four different models have been created; one for each of the output streams of the FASST bdmo separation algorithm, as shown in Figure 4. A different set of features has been automatically selected for each of the SVM models during the training process.

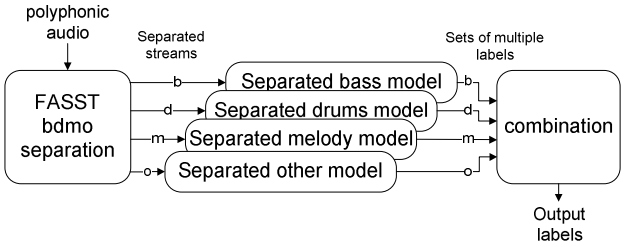

Figure 4. FASST separation into the drum, bass, melody and other streams, combined with the instrument recognition using models trained on the separated audio.

The evaluation showed that using the models trained on each of the streams of separated audio provides better results than using the original models, and better than the original algorithm without any sound segregation. The combination of the "m" and "o" labels already improves the results obtained in "Exp1:n", obtaining a micro F1 = 0.411. The best micro F1-measure (0.446) is obtained with the bdmo combination. If the "n" labels are additionally combined, the micro F1 increases to 0.480.

If we analyze the recognition results per instrument, the best are obtained with the voice, achieving 0.902 precision, 0.574 recall and a 0.701 F1-measure. Clarinet seems to be the most challenging instrument to be recognized, with a F1-measure = 0.113. A further observation is that there is a relation between the stream and the instruments which are better recognized. For instance, the recognition in the bass stream is better for instruments with low frequency content, such as the cello, which is not so well recognized in the rest of the streams.

3.5 Experiment 5: Optimizing the performance of FASST + models trained with separated audio

In this experiment, we aimed at improving the results obtained in Experiment 4 — FASST dbmo separation + models trained with separated audio. Different models are used for the recognition of each of the four audio streams, and thus it is possible to optimize the parameters of each of them. Additionally, we also investigate the requirement of a certain degree N of overlap in the combination of labels. The evaluation showed that if the value of N is increased, the precision increased as well, at the expense of a lower recall. With N = 0 (no overlap required), the obtained micro F1 is equal to 0.446. If N = 1, which is equivalent to outputting only the labels which had been predicted by at least two of the classifiers, we obtain the best precision found in all experiments (0.733), but the recall is considerably reduced (0.354), and the F1-measure is thus smaller. Therefore, the overall performance is considered to be worse when N increases.

As in all previous experiments, the minimum degree of overlap between labels was set to N=0 in Experiment 5, which provided the best results in terms of the F1-measure. The output labels were thus the union of all labels predicted by each of the models. On the other hand, the use of a different configuration for the training of each of the four models led to some improvements in the results, achieving a micro F1= 0.497. In order to further improve the results we tried combining the labels derived from both, source separation and panning-based segregation streams. The combination Exp5:dbmonslr achieved the best F1-measure from all experiments, equal to 0.503.

The most relevant results of all experiments are presented in Table 1.

	Mac Prec	Mac Rec	Mic Prec	Mic Rec	Mac F1	Mic F1
Exp1:n	**57.8**	24.9	**70.8**	25.8	34.9	37.8
Exp2:lrns	48.5	33.8	58.2	36.7	39.8	45.1
Exp3:dbmo	31.0	37.0	33.0	38.5	33.7	35.5
Exp4:dbmo	49.0	30.6	62.5	34.7	37.7	44.6
Exp4:dbmon	47.5	37.3	59.3	40.3	41.8	48.0
Exp5:dbmon	44.0	41.5	54.9	45.5	42.7	49.7
Exp5:dbmolrns	41.0	**45.5**	50.4	**50.1**	**43.2**	**50.3**

Table 1. Instrument recognition measures (in %). See text for details regarding the studied experimental methods and their acronyms

In the following section we analyze the obtained results, and compare the evaluated approaches.

4. DISCUSSION

The highest precision is obtained with the instrument recognition algorithm [5] by itself ("Exp1:n"), at the expense of having a low recall, which provides a medium F1-measure. In "Exp2:nslr" we considerably improve the results with a simple panning-based segregation, achieving a 19.2% relative increase in the micro-F1 with respect to the original algorithm. Experiment 3 makes use of the FASST dbmo separation as a pre-step to the instrument recognition. In this experiment, the precision drops, and the recognition performance is worse. After training the recognition models with source separated data, we obtain considerably better results in "Exp4:dbmo" compared to "Exp3:dbmo" and also "Exp1: n" in terms of F1-measure. With the aggregation of the sets of labels obtained with the original algorithm, we obtain a further increase in the performance in "Exp4:dbmon". The results from "Exp5:dbmon" show that it is possible to further improve the instrument recognition by tuning the parameters of each of the dbmo models. Finally "Exp5:dbmonslr" corresponds to the best results obtained in any of the automatic instrument recognition experiments, by combining "dbmo" sets of labels from the tuned models trained with separated streams, and the "Exp2:nslr" sets of labels obtained with the LRMS separation. The detailed results for all possible combination of labels and experiments can be found in [2]. The best micro F1-measure = 0.503, thanks to the recall gained by the combination of all labels. The

micro F1-measure obtained with the original algorithm without segregation was 0.378, so we were able to improve 12.2 percent points, which represents a 32.3% relative to the initial value. It is interesting to note that the micro averages are better than the macro averages, since the majority of categories with the most frequent instances (e.g. voice) are more easily recognized than the rest.

5. CONCLUSIONS

We presented novel methods to improve the automatic recognition of predominant musical instruments, by its combination with audio segregation algorithms. A comparison with previous similar approaches is not straightforward, since the number of classes and datasets are different. However, if we compare the performance of the original algorithm with the best of our presented approaches combining source separation and instrument recognition, there is around 32% improvement of the micro F1-measure. The way in which the combination is made is very important to be able to improve the results of the algorithms: we found that the application of a source separation pre-step may not provide a better recognition of the instruments if the models do not consider the limitations and errors of the separation algorithms. Training the classification models with the different streams of separated audio has been found to be an effective strategy for acknowledging the typical source separation errors. This leads to a better performance, which can be further enhanced by tuning the parameters of each of the different models used in the instrument recognition. A drawback of the use of the proposed separation algorithm is its computational complexity. As a simple, fast and efficient alternative, we propose the decomposition of the stereophonic polytimbral audio into the left, right, mid and side streams, and the combination of the labels identified by the instrument recognition algorithms in each of the streams. This increased a 19.2% the performance of the predominant instrument recognition.

Acknowledgements: This research was partially supported by "La Caixa" Fellowship Program, and the following projects: Classical Planet: TSI-070100-2009-407 (MITYC), DRIMS: TIN2009-14247-C02-01 (MICINN) and MIRES: EC-FP7 ICT-2011.1.5 Networked Media and Search Systems, grant agreement No. 287711.

6. REFERENCES

[1] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception*, vol. 27, no. 3, pp. 223–242, 2010.

[2] J.J. Bosch, "Synergies between Musical Source Separation and Instrument Recognition", Master's Thesis, Universitat Pompeu Fabra, 2011.

[3] J. J. Burred, "From sparse models to timbre learning: new methods for musical source separation," PhD Thesis, Technical University of Berlin, Berlin, 2008.

[4] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Transactions On Audio Speech and Language Processing*, vol. 14, no. 1, pp. 68-80, 2006.

[5] F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music," in *Proc. of ISMIR*, 2009.

[6] F. Fuhrmann and P. Herrera, "Polyphonic Instrument Recognition for exploring semantic Similarities in Music," in *Proc. of 13th Int. Conference on Digital Audio Effects DAFx10*, Graz Austria, 2010, pp. 1-8.

[7] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. of ISMIR*, 2009.

[8] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic Classification of Musical Instrument Sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3-21, Mar. 2003.

[9] T. Kitahara , M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps", *EURASIP Journal on Advances in Signal Processing*, 2007.

[10] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishomoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related tasks" in *Advances in Music Information Retrieval*. Vol. 274 Springer, 2010, pp. 213–236.

[11] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133.

[12] G. Tzanetakis, "Song-specific bootstrapping of singing voice structure," in *Proc IEEE International Conference on Multimedia and Expo ICME*, 2004.

[13] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing*, Neural Information Processing Systems Workshop, 2006.