MUSICAL GENRE CLASSIFICATION ENHANCED BY IMPROVED SOURCE SEPARATION TECHNIQUES

Aristomenis S. Lampropoulos
University of Piraeus
Department of Informatics
Piraeus 185 34, Greece.
arislamp@unipi.gr

Paraskevi S. Lampropoulou
University of Piraeus
Department of Informatics
Piraeus 185 34, Greece.
vlamp@unipi.gr

George A. Tsihrintzis
University of Piraeus
Department of Informatics
Piraeus 185 34, Greece.
geoatsi@unipi.gr

ABSTRACT

We present a system for musical genre classification based on audio features extracted from signals which correspond to distinct musical instrument sources. For the separation of the musical sources, we propose an innovative technique in which the convolutive sparse coding algorithm is applied to several portions of the audio signal. The system is evaluated and its performance is assessed.

Keywords: Musical Genre Classification, Source Separation, Convolutive Sparse Coding.

1 INTRODUCTION AND WORK OVERVIEW

Recent advances in digital storage technology and the tremendous increase in the availability of digital music files have led to the creation of large music collections for use by broad classes of computer users. In turn, this fact gives rise to a need for systems that have the ability to manage and organize efficiently large collections of stored music files. Many currently available music search engines and peer-to-peer systems (e.g. Kazaa, emule, Torrent) rely on textual meta-information such as file names and ID3 tags as the retrieval mechanism. This textual description of audio information is subjective and does not make use of the musical content and the relevant metadata have to be entered and updated manually, which implies significant effort in both creating and maintaining the music database. Therefore, it is expected that extracting the information from the actual music data through an automated process could overcome some of these problems.

There have been many works on audio content analysis which use various features and methods (Lampropoulos et al., 2004c; Dowling and Harwood, 1996; Aucoutier and Pachet, 2003; Tzanetakis and Cook, 2002, 2000), most of which focus on automatic musical genre classification. These methods provide techniques to organize digital music into categorical labels created by human experts using objective features of the audio signal that relate to instrumentation, timbral texture, rhythmic and pitch content (Aucoutier and Pachet, 2003; Tzanetakis and Cook, 2000). These techniques rely on pattern recognition algorithms and offer possibilities for content-based indexing and retrieval. However, all these works use the complex sound structure of the audio signal in a music file to extract the feature vector.

In this paper, we propose a new approach for musical genre classification based on the features extracted from signals that correspond to musical instrument sources. Contrary to previous works, our approach uses first a sound source separation method to decompose the audio signal into a number of component signals, each of which corresponds to a different musical instrument source, (see Figure 1). In this way timbral, rhythmic and pitch features are extracted from separated instrument sources and used to classify a music clip, detect its various musical instruments sources and classify them into a musical dictionary of instrument sources or instrument teams. This procedure attempts to mimic a human listener who is able to determine the genre of a music signal and, at the same time, identify a number of different musical instruments in a complex sound structure.

The problem of separating the component signals that correspond to the musical instruments that generated an audio signal is ill-defined as there is no prior knowledge about the instrumental sources. Many techniques have been successfully used to solve the general blind source separation problem in several application areas; among these, the Independent Component Analysis (ICA) method (Plumbley et al., 2002; Martin, 1999) appears to be one of the most promising. ICA assumes that the individual source components in an unknown mixture have the property of mutual statistical independence. This property is exploited in order to algorithmically identify the latent sources. Moreover, ICA-based methods require certain limiting assumptions, such as the assumption that the number of observed mixture signals be at least as high as the number of source signals and that the mixing matrix be full rank. However, a method has been proposed which is based on ICA but relaxes the constraint on the number of observed mixture signals. This is called the Independent Subspace Analysis (ISA) method and can separate

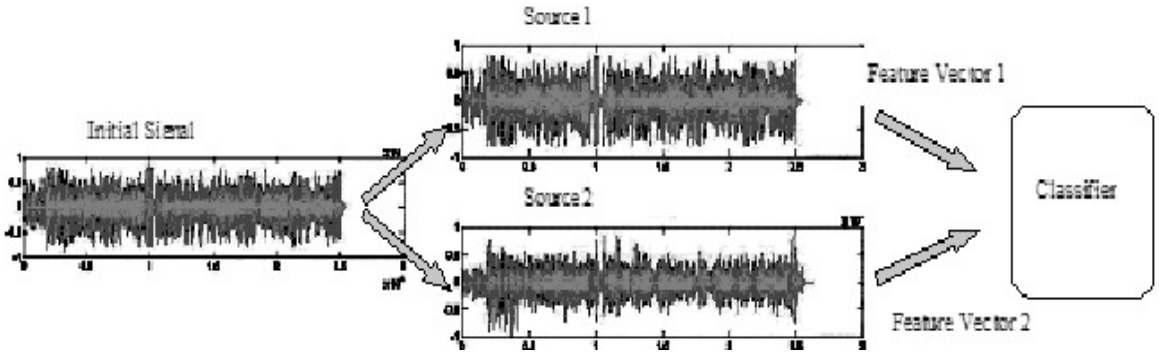

Figure 1: Source Separation: 2 component signals

individual sources from a single-channel mixture by using sound spectra (Casey and Westner, 2000). Signal independence is the main assumption of both the ICA and ISA methods. In musical signals, however, there exist dependencies in both the time and frequency domains.To overcome these limitations, we use in our system a recently proposed data-adaptive algorithm that is similar to ICA and called Convolutive Sparse Coding (CSC) (Virtanen, 2004).

In our first approach (Lampropoulos et al., 2005a), we applied the CSC algorithm in the entire input signal that corresponds to a music piece assuming that the same musical instruments were active throughout the entire music piece duration. This assumption, however, is not realistic, as it is common for different instruments to be used to generate different parts of a music piece. For example, only two instruments may be active in the introduction of a music piece, a third instrument may be added in the middle of the piece and so on. Thus, in Section 2.1, we propose a new approach for music source separation, in which we apply the CSC algorithm to three parts of a music piece.

More specifically, the paper is organized as follows: An overall architecture of our system is presented in Section 2, with Section 2.1 describing the source separation method in detail and Section 2.2 describing the extraction of audio content-based features of music pieces. Classification methods and results are given in Section 3, while conclusions and suggestions for future work are given in Section 4.

2 SYSTEM OVERVIEW

The architecture of our system consists of three main modules, as in Figure 1. The first module realizes the separation of the component signals in the input signal, while the second module extracts features from each signal produced during the source separation stage. Finally, the last module is a supervised classifier of genre and musical instrument. Each music piece can be stored in any audio file format, such as .mp3, .au, or .wav., which requires the application of a format normalization process before feature extraction. For this, we decode each music file into raw Pulse Code Modulation (PCM), using a LAME decoder (Lame) and convert it to the .wav format with resolution of 16 bit samples at a sampling rate of 22.050 Hz.

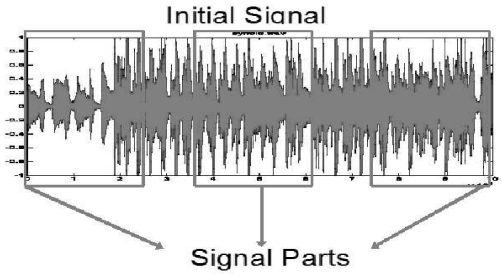

Figure 2: Parts of the Signal

2.1 Improved Source Separation Technique

For source separation, we make the assumption that portions of the signal are sufficient for reliable application of the CSC algorithm. The first step of the source separation technique is to identify the music piece portions. Specifically, we take three parts, one at the beginning, one at the middle and one at the end of the music signal, as shown in Figure 2. The length of each part is 25% of the length of the entire signal. We apply the CSC algorithm to the three signal parts in parallel. We choose the method of convolutive sparse coding because it solves, at least partially, the assumptions of spectra that remain fixed over time and the model fitting criterion of the reconstruction error, assumptions which are not valid for audio signals. The basic signal model in general sparse coding is that each observation vector x_i is a linear mixture of source vectors s_j :

$$x_i = \sum_{j=1}^{J} a_{i,j} s_j, \quad i = 1, ..., I, \tag{1}$$

where $a_{i,j}$ is the weight of j^{th} source in the i^{th} observation signal.

Both the source vectors and the weights are assumed unknown. The sources are obtained by multiplying the observation matrix by the estimate of an *unmixing* matrix. The main assumption in sparse coding techniques is that the sources are non-active most of the time, which means that the mixing matrix has to be sparse. The estimation can be done using a cost function that minimizes the reconstruction error and maximizes the sparseness of the mixing matrix. More specifically, this method is called

convolutive sparse coding because the source model is formulated as the convolution of a source spectrogram and an onset vector. The suitability of this model over-covers the case of respective transient sources.

The commonly used model fitting criterion that consists of the sum of squared elements of the reconstruction error emphasizes music signal aspects that differ from the human sound perception. In order to obtain higher perceptual quality of separated sources, the CSC algorithm uses compression, as explained in the following Section 2.1.1.

2.1.1 Loudness Criterion :

The human auditory system allows humans to perceive very low-amplitude sounds. The large dynamic range of the human auditory system is mainly caused by the nonlinear response of the auditory cells, which can be modeled as a separate compression of the input signal at each auditory channel (Virtanen, 2004). In our system, the compression is modeled by calculating a weight for each frequency bin in each frame. The weights are selected so that the sum of squared magnitudes be equal to the estimated loudness, since the separation algorithm uses the squared error criterion as a fitting criterion. This way "quantitative significance" corresponds to "perceptual significance."

Specifically, in our system, 24 separate bands are spaced uniformly on Bark scale and denoted by disjoint sets F_b, $b = 1...24$. The fixed response of the outer and middle ear is taken into account by multiplying each bin of spectrum by the corresponding response. In the CSC algorithm, the term *loudness index* is used for the loudness estimate in a frame within a critical band. The loudness index in frame t in critical band b is denoted by $L_{b,t}$ and given as

$$L_{b,t} = [\sum_{f \epsilon F_b} (h_b x_{f,t})^2 + \varepsilon_b^2]^\nu - \varepsilon_b^{2\nu}, \qquad (2)$$

where h_b is the fixed response of the outer and middle ear within band b, ε_b is a fixed scalar with value 0.23 and is the (fixed) threshold of hearing on band b. In practice, ε_b is not known, but it can be estimated from the input signal, e.g. by calculating the average level of the signal, and scaling down 30 dB. This procedure has resulted into the value $\varepsilon_b = 0.23$.

2.1.2 The iterative algorithm :

Each part of the input signal is represented with a magnitude spectrogram, which is calculated as follows: first, the time domain input signal is divided into frames and windowed with a fixed 40 ms Hamming window with 50% overlap between frames. Next, each frame is transformed into the frequency domain by computing its discrete Fourier transform (DFT) of length equal to the window size. Only positive frequencies are retained and phases are discarded by keeping only the magnitude of the DFT spectra. This results in a magnitude spectrogram $x_{f,t}$, where f is a discrete frequency index and t is a frame index. A two-dimensional magnitude spectrogram is used to characterize one event of a source at discrete frequency f, t frames as the onset varies between 0 and D.

The magnitudes $x_{f,t}$ and weights $w_{f,t}$ are calculated. The number of sources N is predefined. N should be equal to the number of clearly distinguishable instruments. If the spectrum of one source varies significantly, for example because of accentuation, one may have to use more than one component per source. The model considers the different fundamental frequencies of each instrument as separate sources. Initialize $a_1...a_n$ with the absolute values of Gaussian noise.
Iteration:

1. Update $s_{f,t}$ using the multiplicative step

$$s^{\{p+1\}} = s^{\{p\}} .* \left(A^T W_f^T W_f x_f\right) ./ \left(A^T W_f^T W_f A s^{\{p\}}\right) \qquad (3)$$

   where the $s^{\{p+1\}}$ is the updated $s^{\{p\}}$ for p^{th} iteration given the A, W_f.

2. Calculate $\nabla a_n = \frac{\partial c_{tot}(\lambda)}{\partial a_n}$.
3. Update $a_n \leftarrow a_n - \mu_\kappa \nabla a_n$. Set the negative a_n elements to zero. μ_κ is the step size, which is adaptively set.
4. Evaluate the cost function.
5. Repeat Steps 1-4 until the value of the cost function remains unchanged.

In the synthesis mode, the convolutions are evaluated to get frame-wide magnitudes of each source. To get the complex spectrum, phases are obtained from the phase spectrogram of the original mixture signal. The time-domain signal is obtained by inverse discrete Fourier transform and overlap-add. This procedure has been found to produce best quality. The use of the original phases allows the synthesis without abrupt changes in phase.

3 FEATURE EXTRACTION

We transform an audio signal at a certain level of information granularity. Information granules refer to a collection of data that contain only essential information. Such granulation allows more efficient processing for extracting features and computing numerical representations that characterize a music signal. As a result, the large amount of detailed information in a signal is reduced to a collection of features. Each feature captures some aspects of signal and gives the essential information of that. In our system, we used a 30-dimensional objective feature vector which was originally proposed by Tzanetakis (Tzanetakis and Cook, 2002, 2000) and used in other works (Tzanetakis, 2002; Lampropoulos et al., 2004c; Lampropoulos and Tsihrintzis, 2004a,b; Lampropoulos et al., 2005a,b; Foote, 1999; M. Welsh et al., 1999). For the extraction of the feature vector, we used MARSYAS, 0.1 a public software framework for computer audition applications (Tzanetakis and Cook, 2000). The feature vector consists of three different types of features rhythm related (Beat), timbral texture (musical surface: STFT, MFCCs) and pitch content related.

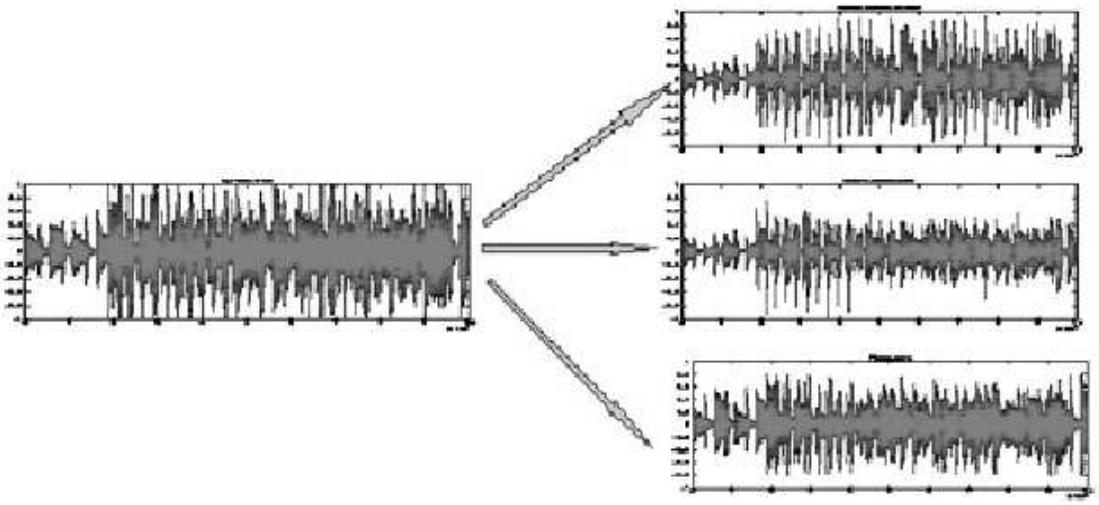

Figure 3: Source Separation: 3 component signals

3.1 Rhythmic Features

Rhythmic features characterize the variation of music signals over time and contain such information as regularity of the tempo. The feature set for representing rhythm is based on detecting the most silent periodicities of the signal. Rhythm is extracted from beat histograms, that is curves describing the beat strength as a function of tempo values and the complexity of the beat in the music. The regularity of the rhythm, the relation of the main beat to subbeats and the relative strength of subbeats to the main beat, are used as some of the features in musical genre recognition systems. The Discrete Wavelet Transform (DWT) is used to divide the signal into octave bands and, for each band, full-wave rectification, low pass filltering, downsampling and mean removal are performed in order to extract an envelope. The envelopes of each band are summed up and the autocorrelation is calculated to capture the periodicities in the signal's envelope. The dominant peaks in the autocorrelation function are accumulated over the whole audio signal into a beat histogram.

3.2 Timbral Textrure

In short time audio analysis, the signal is broken into small, possibly overlapping temporal segments each segment is processed separately. These segments are called "analysis windows" and need to be short enough for the frequency characteristics of the magnitude spectrum to be relatively stable. The term "texture window" describes the longest window that is necessary to identify music texture. The timbral texture features are based on the Short Time Fourier Transform and calculated for every analysis windows. Means and standard deviations are calculated over the texture window.

3.3 Pitch Features

The pitch features describe melody and harmony information about a music signal. A pitch detection algorithm decomposes the signal into two frequency bands and amplitude envelops are extracted for each frequency band. Applying half-way rectification and low-pass filtering performs the envelope extraction. The envelopes are summed and an enhanced autocorrelation function is computed so that the effect of integer multiples of the peak of frequencies to multiple pitch detection be reduced. The dominant peaks of the autocorrelation function are accumulated into pitch histograms and the pitch content features extracted from the pitch histograms. The pitch content features typically include: the amplitudes and periods of maximum peaks in the histogram, pitch intervals between the two most prominent peaks and the overall sums of the histograms.

4 CLASSIFICATION

In order to evaluate our source separation-based music genre classification technique, we have tried different classifiers contained in the machine learning tool called WEKA (WEKA), which we have connected to our system.

In this work, we utilize genre classifiers based on multilayer perceptrons. The input to the artificial neural networks is the feature vector corresponding to the component signals produced by source separation. Specifically, the source separation process produced two or three component signals (see Figures 1 and 3, respectively) which correspond to instrument teams such as strings (bouzouki, guitar,etc), winds (greek clarinet, flute, bagpipes, etc) and percussion instruments (drums, tabor, etc). We constructed two different multilayer perceptrons, in which the artificial neural networks consisted of four (4) and ten (10) hidden layers of neurons, respectively. The number of neurons in the output layer is determined by the number of audio classes we want to classify into (four in this work: rebetico, dimotiko, laiko, entechno). The networks were trained with the back-propagation algorithm and their output estimates the degree of membership of the input fea-

ture vector in each of the four audio classes. Thus, the value at each output necessarily remains between 0 and 1. Classification results were calculated using 10-fold cross-validation evaluation, where the dataset to be evaluated was randomly partitioned so that 90% be used for training and 10% be used for testing. This process was iterated with different random partitions and the results were averaged. This ensured that the calculated accuracy was not biased because of the particular partitioning of training and testing. The specific data set we used consisted of 1049 music pieces from 4 genres of greek songs, namely Rebetico (396 pieces), Dimotiko (106 pieces), Laiko (414 pieces), and Entechno (133 pieces).

Table 1: Correctly Classified Instances without/with Source Separation

Classifier	**w/out SS**	**with SS**
Nearest-Neighbour Classifier	67.6	68.2
MLP 4 hidden layers	73.2	74.2
MLP 10 hidden layers	73.9	75.1

Table 2: Correctly Classified Instances with CSC and with improved CSC

Classifier	**CSC**	**imp CSC**
Nearest-Neighbour Classifier	68.2	69.2
MLP 4 hidden layers	74.2	75.8
MLP 10 hidden layers	75.1	75.9

As seen in Table 1 , the classification results after implementation of the source separation technique presented an improvement of 1% - 2%. This was due to the fact that the source separation technique revealed more information about timbral texture, rhythm and pitch (harmony) content, not only for the signal as a whole, but for a number of the separated instrument team sources. Moreover, as seen in Table 2, after implementation of the improved source separation technique we had a 0.5% - 1% improvement over a previous work of ours (Lampropoulos et al., 2005a), in which the CSC algorithm was applied to the entire signal. Thus, the total improvement in our present approach is about 2% - 2.5% over previous genre classification methods. Finally, the present CSC approach not only results in better genre classification, but is faster than the existing CSC algorithms, as it is applied in parallel to small portions of duration of 30 - 50 sec and not on an entire audio signal of duration of 3 - 3.5 min.

Table 3: Confusion Matrix: MLP 4 hidden layers w/out SS 73.2%

	Rebetico	**Dimotiko**	**Laiko**	**Entechno**
R	**304**	9	73	10
D	20	**67**	18	1
L	88	6	**307**	13
E	17	2	24	**90**

Table 4: Confusion Matrix: MLP 4 hidden layers with SS CSC 74.2%

	Rebetico	**Dimotiko**	**Laiko**	**Entechno**
R	**325**	8	57	6
D	32	**62**	10	2
L	101	9	**296**	8
E	25	0	13	**95**

Table 5: Confusion Matrix: MLP 4 hidden layers with SS improved CSC 75.8%

	Rebetico	**Dimotiko**	**Laiko**	**Entechno**
R	**299**	8	80	9
D	22	**67**	15	2
L	67	7	**330**	10
E	19	0	15	**99**

To analyze further a musical genre classifier (e.g., the multilayer perceptron with 4 hidden layers), we also present the corresponding *confusion matrices*, as shown in Tables 3 (without Source Separation, classification accuracy of 73.2%), 4 (with Source Separation based on CSC, classification accuracy of 74.2%) and 5 (with Source Separation based on improved CSC, classification accuracy of 75.8%). In a confusion matrix, the columns correspond to the *actual* genre, while the rows correspond to the *predicted* genre. For example in Table 3, the cell in row 2 of column 4 has value 1, which means that 1 song (in a total of 106 songs) from the ”dimotiko” class was wrongly predicted as ”entechno”. Similarly, 20 and 18 songs from the ”dimotiko” class were predicted to be from the ”rebetiko” and ”laiko” classes, respectively. Therefore, the percentage of correct classification of songs from the ”dimotiko” class is computed to equal 67*100/106=63,2 % for this classifier. The correct classification percentages are, therefore, derived from the entries in the diagonal elements of a confusion matrix and the corresponding actual number of songs in the library.

5 CONCLUSIONS - FUTURE WORK

It has been observed that audio signals corresponding to music of the same genre share certain common characteristics as they are performed by similar types of instruments and have similar pitch distribution and rhythmic patterns (Aucoutier and Pachet, 2003). Motivated by this, we presented a novel approach based on classification of features extracted from component signals that corresponded to musical instrument teams (sources), as these sources have been identified by a source separation process. For source separation, we presented an improved algorithm based on convolutive sparse coding. Evaluation of the performance of our method showed clear improvement in classification accuracy and execution speed over our previous methods (Lampropoulos et al., 2005a).

Currently, we are in the process of improving further the classification efficiency of our system by considering additional low-level music features, specifically MPEG-7

low-level audio descriptors, incorporating other classifiers such as immune classification algorithms (Lampropoulos et al., 2005b) and additional classifiers included in the WEKA tool (WEKA). Another direction of our future work will be the identification of specific instruments from the separated component sources. This and related work is currently in progress and its results will be announced shortly.

REFERENCES

J. J. Aucoutier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

M. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. International Computer Music Conference, ICMA*, Berlin, August 2000.

W. J. Dowling and D. L. Harwood. *Music Congition*. Academic Press, 1996.

J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999.

Lame. http://lame.sourceforge.net.

A. S. Lampropoulos, P. S. Lampropoulou, and G. A. Tsihrintzis. Musical genre classification of audio data using source separation techniques. In *Proc. IEEE 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, The Slovak Republic, July 2005a.

A. S. Lampropoulos, D. N. Sotiropoulos, and G. A. Tsihrintzis. Artificial immune system-based music piece similarity measures and database organization. In *Proc. IEEE 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, The Slovak Republic, July 2005b.

A. S. Lampropoulos, D. N. Sotiropoulos, and G. A. Tsihrintzis. Individualization of music similarity perception via feature subset selection. In *Proc. IEEE International Conference on Systems, Man, and Cybernetics 2004*, The Hague, The Netherlands, October 2004c.

A. S. Lampropoulos and G. A. Tsihrintzis. Agglomerative hierarchical clustering for musical database visualization and browsing. In *Proc. 3rd Hellenic Conference on Artificial Intelligence*, Samos Island, Greece, May 2004a.

A. S. Lampropoulos and G. A. Tsihrintzis. Semantically meaningful music retrieval with content-based features and fuzzy clustering. In *Proc. 5th International Workshop on Image Analysis for Multimedia Interactive Services.*, Lisboa, Portugal, April 2004b.

N. M. Welsh, B. von Behren, and A. Woo. Querying large collections of music for similarity. Technical report UCB/CSD00-1096, Computer Science Department U.C. Berkeley, 1999.

K. Martin. *Sound-source recognition: A theory and computational mode*. PhD thesis, MIT, 1999.

M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6):603–627, 2002.

G. Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Princeton University, 2002.

G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organised Sound*, 4(3), 2000.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), July 2002.

T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proc. Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, SAPA*, 2004.

WEKA. http://www.cs.waikato.ac.nz/ml/weka.