Model-based Approach to Separating Instrumental Music from Single Track Recordings

Sintiani D. Teddy and Edmund M-K. Lai

School of Computer Engineering
Nanyang Technological University, Singapore 639798.
E-mail: sdt@pmail.ntu.edu.sg, asmklai@ntu.edu.sg

Abstract

The objective of audio source separation is to separate sound mixtures into individual streams based on the sources. It has many potential applications, one of which is in a system for perceptually-based search and retrieval of audio data from multimedia databases.

The task of audio source separation is very difficult if all audio data are mixed into a single track. In this paper, we restrict the single track recordings to instrumental music. Hence we attempt to construct separate streams of data each consisting of the sound of a single instrument. A model-based approach is used. The architecture of the system is based on a cerebellar-based (CMAC) fuzzy neural network. The subjective test results of our experiments on the separation of 2-source audio mixtures shows that our approach is promising.

Keywords: Single Channel, Audio Source Separation, Musical Data Mining, CMAC.

1 Introduction

Audio Source Separation is one of the first steps of Computational Auditory Scene Analysis (CASA), which attempts to build a computer program to understand and analyze an auditory scene. While the general problem in Audio Scene Analysis [1] is to understand and explain the complex phenomenon in audition, Audio Source Separation is about separating mixtures of audio signals into individual streams of audio data based on the source.

Separation of mixture of audio from various sources into its individual sound stream will allows for a construction of a more structured representation of audio data. This structured representation can be used for perceptual-based search and retrieval of audio data from multimedia databases to be performed. The work in this paper can be considered as the front-end processing in the creation of a musical database, by performing intelligent musical data mining of instrumental music. Audio source separation can be used to separate, classify and therefore analyze the content of the digital audio, and therefore providing more information in performing a more human-understandable and meaningful search.

There are three main approaches to audio source separation: model-based, prediction-based and blind source separation (BSS). Model-Based source separation make use of some form of phychoacoustic model in the separation process. Research involved here typically pay close attention to the result from the psychoacoustic and perceptual study of human audition. Examples are audio content analysis for signal classification [2, 3, 4, 5], and automatic transcription of music [6, 7, 8].

Prediction-based audio separation, on the other hand, is motivated by the perceptual illusion phenomena in human auditory processing. Examples include the continuity illusion and phonemic restoration phenomena which show that the brain is able to use a wide range of knowledge drawn from past experiences for the interpretation of obscured or complex sound mixtures. A comprehensive introduction on this material can be found in [9]. Examples of systems making use of this approach include [10, 11, 12, 13].

Blind source separation (BSS) attempts to obtain a decorrelation matrix from the mixtures. It is blind the technique does not make any assumption about the source signals and the mixing functions. There is no prior knowledge about the statistics of the source in general. It is most effective in situations where multiple spatially separated sensors (microphones), and hence multiple track recordings, are available.

The scope of our research reported in this paper is restricted to single track recordings of the audio mixture which is a substantially more difficult task compared to that for multiple track data. This is motivated by the fact that in most circumstances, only a single track record is available. A model-based architecture employing a fuzzy CMAC (Cerebellar Model Articulation Controller) neural networks has been developed. The system is used to perform audio segregation of 2-source mixtures of various combination of piano, flute, clarinet and trumpet sound.

The subjective test results of our experiments showed promising accuracy in identifying the instruments' type and pitch inside the mixtures and the quality of the separated streams is good.

The architecture of our system is described in Section 2. In 3, we discuss the auditory features used and the techniques involved. The experiments and subjective test results are presented in 4, with a discussion of the advantages of our approach given in 5. Finally, we conclude in 6 with a brief outline of further work that is currently being pursued.

2 Model Based Approach to Source Separation

The general architecture of our model-based approach to source separation is outlined in Figure 1. It consists of three major components: the pitch estimator block, signal reconstruction block and signal cancellation block. The experimental system assumes two-source mixtures input, $y = x_1 + x_2$, whereby one of the sound component is of known type, which is referred to in this paper as the *primary component* x_1.

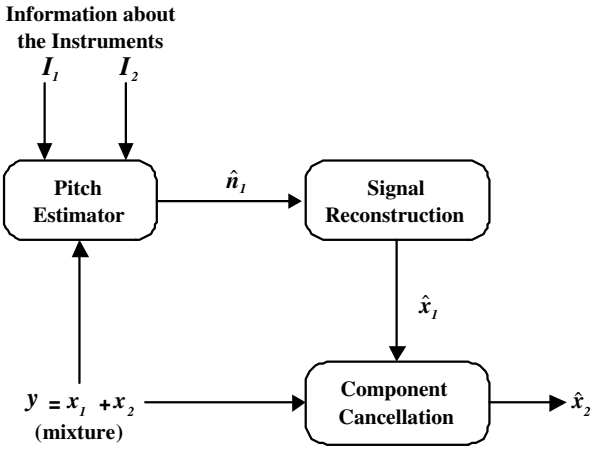


Figure 1: General Architecture of The Model-Based Approach to Musical Sound Separation from Single Channel Recordings

$y = x_1 + x_2$	the mixture input
I_1 & I_2	information about the instruments
$\hat{n}_1$	the estimated pitch of the primary component
$\hat{x}_1$	the reconstructed primary component signal
$\hat{x}_2$	the segregation result

2.1 Pitch Estimator

Every particular sound source can be identified and differentiated by its unique characteristics, i.e. the *timbre* of the sound. The mechanism behind the note estimator was motivated by our hypothesis on the mechanism of human auditory system. In this case, the human brain is considered as memory banks for storing the characteristics of various sounds (I_1&I_2 in Figure 1), which are acquired through repeated exposures to the particular sounds of interest. The features themselves are extracted from the sounds through various processing along the human auditory pathway [15]. The identification and thus separation of a particular sound source is then performed by human brain via pattern-correlation between the observed current features and the information stored in the memory.

The audio segregation method proposed in this paper works by elimination of the primary sound component - the component belonging to the known source type. It begins with the pitch estimation of the primary component, which in this particular implementation, is considered to be of piano's type. The human brain is modeled by a cerebellar-based (CMAC) fuzzy neural network described in [16], which stores pattern associations of the various sounds.

The CMAC network [14] emerged from the attempt to model a part of human brain called the cerebellar cortex. CMAC behaves like a memory, where the particular input to output mapping acts as the address decoder. In CMAC, each possible input address selects a unique set of memory location, the sum of whose content is the contents of the input address. This fact implies that any particular cell location can be selected by more than one input, and also ensures that whenever any two input vectors that are similar (close together in the input space defined by the state and input) will activate many of the same granule cells and thus output similar results. This property is know as the generalization of CMAC, is a property that is very important in this particular audio source separation applications, since it means that training is not required at every point of the input space in order for an approximately correct result to be obtained.

The pitch estimator block is depicted in Figure 2. The pitch estimation process consists of feature extraction, followed by feature segregation, and then feature recognition. The CMAC neural networks used are first trained to store the differentiating features, which will then serve as the knowledge-base of the system. A set of features is first extracted from the sound mixtures. Using the knowledge of the system, this feature set is clustered into two separate feature vectors, according to the sources. These vectors are then compared to the pattern stored in the CMAC to identify the primary component. Thereafter, another memory recall is performed for the estimation of the pitch ($\hat{n}_1$) of the primary component.

2.2 Primary Signal Reconstruction and Cancellation

The primary component's signal reconstruction ($\hat{x}_1$) and cancellation are performed simultaneously by filtering method, i.e. filtering out the primary component (x_1). Filtering is performed on the mixture, and the residual result represents the estimated unknown source's component ($\hat{x}_2$).

In this work, the filtering method chosen was the comb filtering method. The filter is designed to notch at every multiple of the fundamental frequency of the primary component. The filtered results are then interpolated between frames to yield the final result. The primary component is recovered by subtracting the recovered component from the original mixtures.

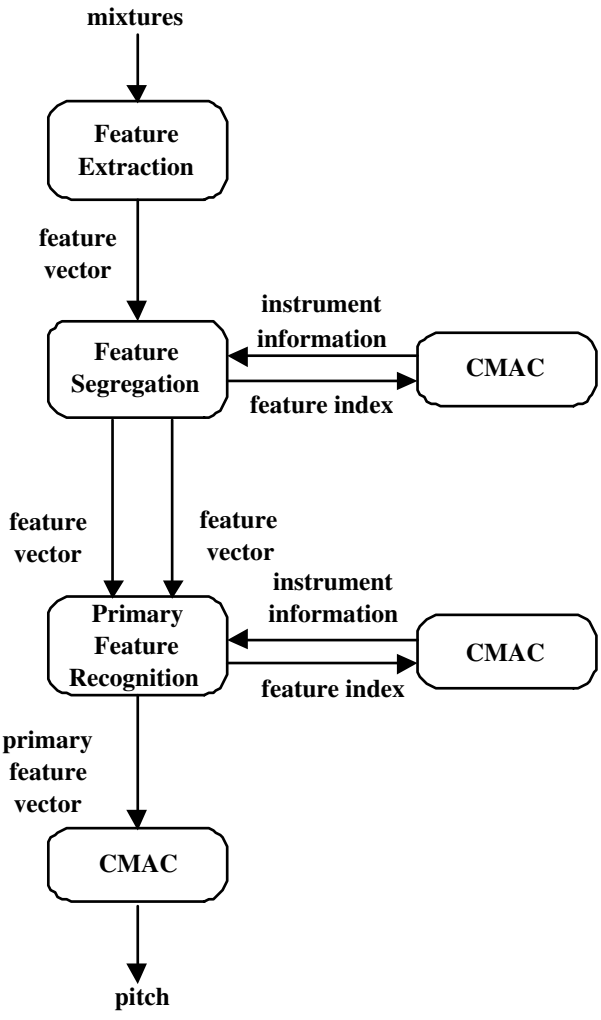


Figure 2: The Pitch Estimator System

3 The Differentiating Features

The source separation system presented above was designed as a general framework to allow it to be used with any combinations of instrument differentiating features. In general, the combinations of differentiating factors have to follow certain rules:

1. the features have to be relatively stable throughout the duration of the same sound, despite the periodic variations in the waveform of the musical notes.
2. the features have to have relatively little variations across pitches of the same instrument or instruments with similar timbre.
3. the features have to be relatively different across instruments with highly perceptually distinguishable timbre.

In this particular implementation, the set of features used is obtained using the Auditory Image Model (AIM) [17, 18] developed by the Center for The Neural Basis of Hearing (CNBH), University of Cambridge. The MATLAB implementation code of the AIM processing is available at [19]. This AIM is modelled from the mechanism of human auditory processing [15]. The processing of sound based on the AIM consists of of several stages [19]:

1. Pre-cochlear processing (PCP).
   The PCP is the processing of the physical sound wave up to the stage of the oval window inside the inner ear. There are several versions of the PCP available. In this paper, the "equal loudness contour" [20] processing is used.
2. Basilar Membrane Motion (BMM).
   The BMM processing module consists of the auditory filterbank, in particular, Gammatone filterbank, which emulate the operation of the cochlea. Gammatone filterbank is a set of bandpass filters, whose bandwidth increases with the center frequencies. The output of the Gammatone filterbank is a set of filtered waves simulating the motion of the basilar membrane as a function of time.
3. Neural Activity Pattern (NAP).
   The NAP is the representation of a model of neural firings' patterns in the auditory nerves transduced by the inner haircells mounted along the edge of the basilar membrane.
4. Temporal Strobing.
   The strobing process is intended to mark important time points in the signal, and is used as a basis of performing temporal integration in creating the final "Auditory Image".
5. Auditory Image Construction.
   Periodic sounds give rise to static, rather than oscillating perceptions, indicating that temporal integration is applied in the process of perception of sound. The result of the strobe temporal integration process is a relatively stable

auditory image or Stabilized Auditory Image (SAI), and it is the signal representation used in our experiments.

This auditory modeling technique was chosen based on the observations carried out on the outputs of the model corresponding to the instruments' sound segments. The observations showed a relatively stable and unique SAI representations for each of the instruments. An example of comparison between the SAI of C4 note of piano and C4 note of clarinet is given in Figure 3.

In our experiments, the frequency profile of the SAI was used as the input data. The features extracted were the peak locations and their relative magnitudes. In musical pitch identification, the locations of the peaks were used, since it was observed that the locations of the peaks were preserved for the same pitch across different instruments, while in the case of instrument type identification, the relative values of the peaks are used as the differentiating factors.

4 Experiments and Results

Experiments were carried out using musical notes generated by 4 musical instruments: Clarinet, Flute, Piano and Trumpet, on their C4–G4 range of notes. Each mixture consists of 2 musical notes, where one of them is the sound from the piano and the other is from one of the other instruments. All sound samples are taken from McGill University Master Samples' collection. Artificial mixtures of piano-flute, piano-clarinet and piano-trumpet are created by superposition and scaling of the individual sound waves.

In the separation process, the mixed signal is segmented into $40ms$ time-frames with 50% overlap. The component detection and pitch identification of the piano note are therefore performed on a frame-by-frame basis. Following the identification of components and estimation of the pitch of the piano component, the sound segment is comb-filtered to remove the piano component. The segments are then interpolated between frames to yield the final result. The piano component is recovered by subtracting the recovered component from the original mixtures. An example of recovery result is depicted in Figure 4, which shows the a mixture sound segment of piano and flute type (Fig 4(a)), and the recovered flute component (Fig 4(b)).

To assess the separation quality, a subjective quality assessment by hearing test were conducted on 20 subjects. Each of the test subject was presented with the both of the original sound segments, the mixture segment and the recovered segment. The percentage of results are shown in Table 1. In general, the test shows a satisfactory separation results.

5 Discussions

The main problem encountered in this source separation attempt was the missing peaks due to the overlapping frequency profiles of the components in the mixture. To alleviate this problem, instead of having a purely bottom-up information flow (from auditory peripheral to the brain), a top-down information flow [9] recalled from the present knowledge stored in the CMAC was also used to assist the feature extraction process. In other words, the knowledge recalled from the brain is used to approximate the locations of the missing peaks from the frequency profile.

Another problem faced was the noisy features' values due to shifted peak locations and varying peak amplitude level in the frequency profile of the mixtures. This problem is addressed by the generalization nature of the fuzzy neural networks (CMAC) which does not require training at every possible input values combinations in order for correct results to be obtained.

6 Conclusions

A new architecture for model-based approach to separating single channel instrumental recordings has been presented. The experimental results shows a promising research direction towards Audio Source Separation. The framework proposed represents a general and extensible structure that allows different feature extraction modules to be incorporated.

There are a lot of improvements to be embedded into the system to improve the quality of the separation result. One of them is the filtering method. Right now, the filtering method used is a simple comb filtering method whereby the notching response of the filter is fixed at multiples of fundamental frequency. However, in the real instrument case, the position of the harmonics frequencies are not always in exact multiples of the fundamental. In this case, a more exact filtering technique needs to be proposed and another CMAC network can be used to hold the filter coefficients.

Acknowledgements

The authors wish to acknowledge the fruitful discussions with Dr H. C. Quek, School of Computer Engineering, Nanyang Technological University regarding fuzzy neural networks.

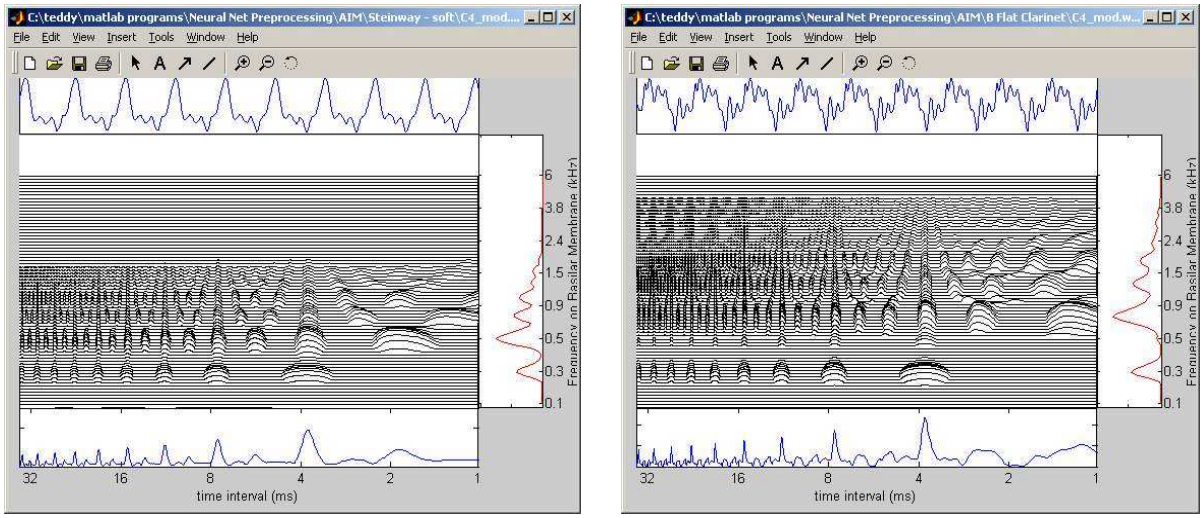
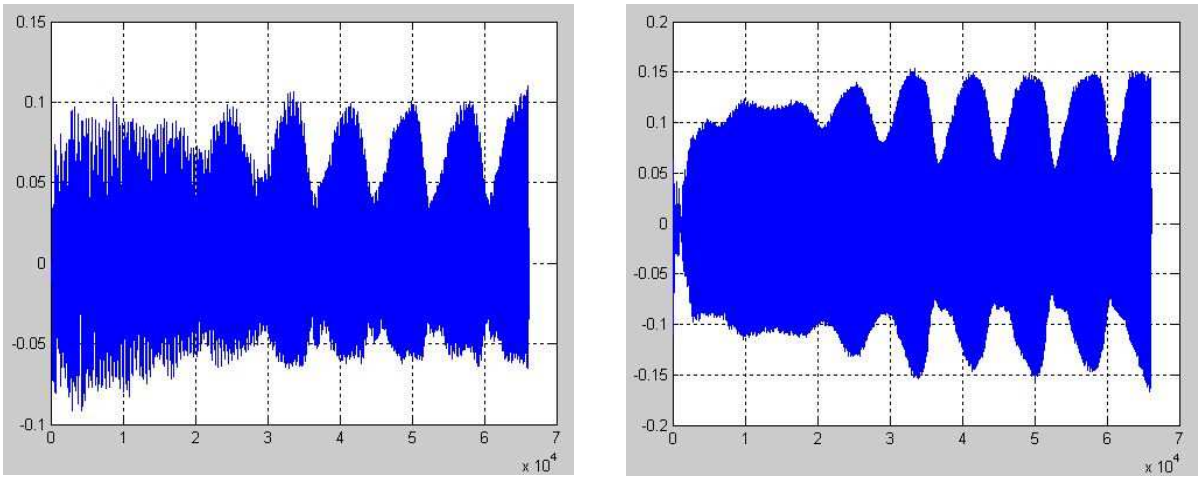


Figure 3: The SAI representation of C4 Piano (left) and C4 Clarinet (right)



(a) Mixture of piano and flute sound

(b) The Recovered Flute Segment

Figure 4: Example of Recovery Result

Mixture Type	**1 (very poor)**	**2**	**3**	**4**	**5 (very good)**
C4 Clarinet + G4 Piano		15%	25%	55%	5%
C4 Flute + G4 Piano			15%	50%	35%
C4 Trumpet + G4 Piano		20%	25%	55%	
D4 Piano + G4 Clarinet		15%	40%	30%	15%
D4 Piano + G4 Flute				55%	45%
D4 Piano + G4 Trumpet		15%	45%	40%	

Table 1: Results of Subjective Quality Assessment on the Separation Results

References

[1] A. Bregman, *Auditory Scene Analysis*. Cambridge, MIT Press, 1996.

[2] D. Gerhard, "Audio Signal Classification: An Overview," *Canadian Artificial Intelligence*, Vol. 45, No. 4–6, 2000.

[3] G. Tzanetakis and P. Cook., "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, 2002, pp 293–302.

[4] K. D. Martin, E. D. Schreirer, and B. L. Vercoe, "Music content analysis through models of audition," *Proceedings of ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*, Bristol UK, 1998.

[5] E. D. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proceedings of IEEE ICASSP*, 1997, pp.1331-1334.

[6] K. D Martin, "Automatic transcription of simple polyphonic music: Robust Front End Processing," *Perceptual Computing Techical Report*, No. 399, MIT Media Lab, Cambridge, 1996.

[7] K. D. Martin and E. D. Scheirer, "Automatic transcription of simple polyphonic music: Incorporation high-level knowledge," *Society for Music Perception and Cognition Conference*, 1997.

[8] A. Klapuri, T. Virtanen, A. Eronen, and J. Seppnen. "Automatic transcription of musical recordings," *Consistent & Reliable Acoustic Cues Workshop* (CRAC-01), Aalborg, Denmark, 2001.

[9] M. Slaney, "A Critique of Pure Audition", in *Computational Auditory Scene Analysis*. D. Rosenthal and H. Okuno (Eds.). Mahwah, NJ: Lawrence Erlbaum Assoc. 1997.

[10] D. P. W. Ellis, "Using Knowledge to Organize Sound: The Prediction-driven Approach to Computational Auditory Scene Analysis, and its application to speech/non-speech mixture," *Speech Communication*, Special Issue on Computational Auditory Speech Analysis. M. Cooke and H. Okuno, Guest Eds., 1997.

[11] K. D. Martin, *"Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing,"* Technical Report No. 399, Massachusetts Institute of Technology, 1996.

[12] J. P. Bello and M. Sandler, "Blackboard System and Top-Down Processing for the Transcription of Simple Polyphonic Music," *Proceedings of the COST G-6 Conference on Digital Audio Effects*, verona, Italy, 2000.

[13] G. Monti and M. Sandler, "Automatic Polyphonic Piano Note Extraction Using Fuzzy Logic in Blackboard System," *Proceedings of the 5th International Conference on Digital Audio Effects*, Hamburg, Germany, 2002.

[14] J. S. Albus, *Brains, Behavior, & Robotics*. McGraw-Hill, 1981.

[15] W. A Yost, *Fundamentals of Hearing*. 4th ed. Academic Press, 2000.

[16] K. K. Ang, and C. Quek, "Improved MCMAC with Momentum, Neighborhood and Averaged Trapezoidal Output," *IEEE Transactions on Systems, Man, and Cybernetics*. Part B: Cybernetics, Vol. 30, No.3 , 2000.

[17] R. D. Patterson, "Auditory images: How complex sounds are represented in the auditory system". *Journal of Acoustical Society of Japan (E)*, Vol. 21, No. 4, pp.183–190. 2000.

[18] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, Proceedings of the 9th International Symposium on Hearing*. Y. Cazals, L. Demany, K. Horner (Eds). Pergamon, Oxford, pp. 429–446, 1992.

[19] "An Implementation of Auditory Image Model in MATLAB," available at *http://www.mrc-cbu.cam.ac.uk/cnbh/aimmanual/Introduction/-Introductionframeset.htm*.

[20] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.