MUSICAL SOUND SEPARATION USING PITCH-BASED LABELING AND BINARY TIME-FREQUENCY MASKING

Yipeng Li and DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
{liyip, dwang}@cse.ohio-state.edu

ABSTRACT

Monaural musical sound separation attempts to segregate different instrument lines from single-channel polyphonic music. We propose a system that decomposes an input into time-frequency units using an auditory filterbank and utilizes pitch to label which instrument line each time-frequency unit is assigned to. The system is conceptually simple and computationally efficient. Systematic evaluation shows that, despite its simplicity, the proposed system achieves a competitive level of performance.

Index Terms— musical sound separation, computational auditory scene analysis, pitch-based labeling

1. INTRODUCTION

As the demand for automatically analyzing, organizing, and retrieving a vast amount of online music data explodes, musical sound separation has attracted significant attention in recent years. Monaural separation that attempts to recover each source/instrument line from single-channel polyphonic music is a particularly challenging problem. On the other hand, a system possessing such functionality allows more efficient audio coding, accurate content-based analysis, and sophisticated manipulation on musical signals [1].

In music multiple instruments often play simultaneously. The polyphonic nature of music creates unique problems for monaural musical sound separation. One such problem is overlapping harmonics where a harmonic of one note has a frequency that is the same as or close to the frequency of a harmonic from another concurrent note. The phenomenon of overlapping harmonics is common since Western music favors notes that are harmonically related—pitches are in a simple integer ratio [2]. It is in general difficult to recover each individual harmonic without instrument-specific knowledge. The interplay of different instrument lines also makes the independence assumption of sound sources invalid.

Broadly speaking, existing monaural musical sound separation systems are either based on traditional signal processing techniques (mainly sinusoidal modeling), statistical techniques (such as sparse coding and nonnegative matrix factorization), or psychoacoustic studies (computational auditory scene analysis).

The basic idea of sinusoidal modeling is to model a sound as a linear combination of sinusoids with time-varying frequencies, amplitudes, and phases. The task of musical sound separation is to estimate these parameters for each source from music. Virtanen [1] estimates parameters of sinusoids using least mean square estimation and applies spectral envelope modeling to address the problem of overlapping harmonics. Every and Szymanski [3] extract sinusoids using a technique called spectral filtering. Sinusoidal modeling is primarily used for pitched sounds. When the pitch of each sound source can be accurately estimated or known *a priori*, sinusoidal modeling usually gives good results for low polyphonies.

Statistical approaches to musical sound separation generally assume certain statistical properties of sources. Sparse coding assumes that a source is a weighted sum of bases from an overcomplete set. The weights are assumed to be zero with high probability, i.e., most of the bases are inactive most of the time [4]. Although nonnegative matrix factorization attempts to find a mixing matrix and a source matrix with nonnegative elements such that the reconstruction error is minimized, it implicitly requires the mixing weights and source to be sparse [5]. Several recent systems [6, 1] have demonstrated the applicability of statistical approaches. However the effectiveness of such approaches on a broad range of tasks remains to be seen.

Computational auditory scene analysis (CASA) is inspired by auditory scene analysis (ASA) [7], an influential perceptual theory which attempts to explain the remarkable capability of the human auditory system in sound separation. CASA aims to build computational systems for general sound separation. Several CASA systems have been developed for musical sound separation. Mellinger's work [8] represents the first CASA attempt to the task. Brown and Cooke [9] proposed a system which includes organizational cues such as pitch for separation. The system by Godsmark and Brown [10] uses a blackboard architecture to reconcile different organizational cues. However, the performance of existing CASA-based separation systems is limited.

Pitch has been shown to be very effective in organizing frequency components from a source. In this paper, we propose a pitch-based CASA system to separate music with two concurrent pitched sounds. Section 2 provides a detailed description of the proposed system. Evaluation and comparison

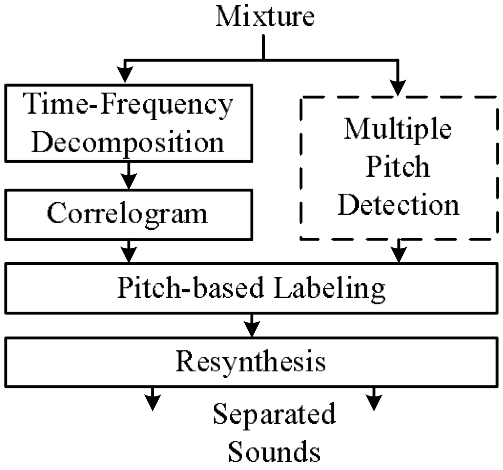

Fig. 1. Schematic diagram of the proposed system.

are presented in Section 3. The last section concludes the paper.

2. SYSTEM DESCRIPTION

Our proposed system is illustrated in Fig. 1. The input to the system is monaural polyphonic music. In the time-frequency (T-F) decomposition stage, the system decomposes the input into its frequency components using an auditory filterbank and divides the output of each filter into overlapping frames. We call an element indexed by frame and frequency a T-F unit. In the next stage, an auditory representation, called the correlogram, is computed. At the same time, the pitches of different instrument lines are detected in the multiple pitch detection module. Multiple pitch detection for music is a very difficult problem. Since the main focus of this study is to investigate the performance of pitch-based separation in music using auditory representations, we do not perform multiple pitch detection (indicated by the dashed box); instead we supply the system with ideal pitches detected from premixing instrument lines. In the pitch-based labeling stage, pitches are used to determine which instrument line each T-F unit should be assigned to. This creates a binary mask for each line. In this paper we do not attempt to separate overlapping harmonics and we leave it for future study. In the final stage of the system, the masks are used to resynthesize individual instrument lines. The details of each stage are explained in the following subsections.

2.1. Time-Frequency Decomposition

In this stage, the input sampled at 20 kHz is first decomposed into its frequency components with a 128-channel gammatone filterbank, which is a widely used auditory filterbank to model cochlear filtering [11]. The impulse response of a gammatone filter is

$$g(t) = \begin{cases} t^{l-1}\exp(-2\pi bt)\cos(2\pi ft), & t \geq 0 \\ 0, & \text{else,} \end{cases} \tag{1}$$

where $l = 4$ is the order of the filter, f is the center frequency of the filter, and b determines the bandwidth of the filter.

Usually the center frequencies of the filters are linearly distributed on the so-called "ERB-rate" scale, $E(f)$, which is related to frequency by

$$E(f) = 21.4\log_{10}(0.00437f + 1). \tag{2}$$

It can be seen from (2) that the center frequencies of the filters are approximately linearly spaced in the low frequency range while logarithmically spaced in the high frequency range. Therefore more filters are placed in the low frequency range where most sound energy is concentrated.

The bandwidth b of a fourth-order gammatone filter is usually set to be

$$b(f) = 1.019\,\mathrm{ERB}(f), \tag{3}$$

where $\mathrm{ERB}(f) = 24.7 + 0.108f$ is the equivalent rectangular bandwidth of auditory filters. This bandwidth is adequate for most speech separation tasks where the intelligibility of separated speech is the main concern. However for music sound separation, $b(f)$ defined as (3) appears too wide for accurate analysis and resynthesis, especially in the low frequency range. We have found that using narrower bandwidth can significantly improve the quality of separated sounds. In this study we set the bandwidth to $b(f) = 1.019\,\mathrm{ERB}(f)/4$, a quarter of the conventionally used value. The center frequencies of channels are spaced from 50 to 8000 Hz.

After auditory filtering, the output of each channel is divided into frames of 20 ms with 50% overlap.

2.2. Correlogram

After T-F decomposition, the system computes an autocorrelation for each T-F unit. This produces a correlogram, the well-known mid-level auditory representation [11]. Specifically, for each T-F unit u_{cm}, where c is the frequency channel index and m the frame index, we calculate autocorrelation as following:

$$A(c,m,\tau) = \sum_{k=0}^{K-1} h(c, mT-k)h(c, mT-k-\tau)w(k). \tag{4}$$

$h(c,\cdot)$ is the output of frequency channel c and K is the frame length. T is the frame shift, τ the time delay, and w a rectangular window with length K.

2.3. Pitch-based Labeling

After the correlogram is computed, we label each u_{cm} using ideal pitches. If there is only one source at frame c, all u_{cm}'s of the frame are assigned to the active source.

For frames with two sources, we consider the values of $A(c,m,\tau)$ at time lags that correspond to two different pitch periods, τ_1 and τ_2. If in u_{cm} only one source is present, e.g., the first one, then $A(c,m,\tau_1) > A(c,m,\tau_2)$ unless τ_1 and τ_2 are in harmonic relation. If two harmonics overlap in u_{cm}, one should assign the u_{cm} to the source with higher energy.

A natural choice is to compare $A(c, m, \tau_1)$ to $A(c, m, \tau_2)$ to see which one has a higher value. Intuitively if one harmonic has a higher amplitude than the other, the peak in the autocorrelation function would be closer to τ_1 than τ_2. We have conducted simulations to check this intuition. For each frequency channel, we create a cosine with the center frequency of the channel. We then create another cosine whose frequency ratio to the first one is uniformly distributed from 1 to 1.03, which corresponds to a semitone. The relative amplitude and phase of the second cosine to the first are uniformly distributed in the range of [0, 0.5] and [0, 2π], respectively. We find that, although our intuition is not strictly true, it is valid most of the time. Therefore we use two pitch periods to directly label each T-F unit. Specifically, u_{cm} is labeled 1 if $A(c, m, \tau_1) > A(c, m, \tau_2)$, and 0 otherwise.

2.4. Resynthesis

With the masks obtained in the pitch labeling stage, each individual instrument line can be resynthesized from the outputs of the gammatone filterbank. This is achieved using a technique introduced by Weintraub [12] (see also [11]). The resynthesis pathway allows the quality of separated lines to be assessed by human listeners or measured by signal-to-noise ratio (SNR) before and after separation.

3. EVALUATION

To evaluate the proposed system, we constructd a database consisting of 20 pieces of quartet composed by J. S. Bach. Since it is difficult to obtain multi-track signals where different instruments are recorded in different tracks, we generate audio signals from MIDI files. For each MIDI file, we use the tenor and the alto line for synthesis since we focus on separating two concurrent instrument lines. Audio signals could be generated from MIDI data using MIDI synthesizers. But such signals tend to have stable spectral contents, which are very different from real music recordings. In this study, we use recorded note samples from the RWC music instrument database [13] to synthesize audio signals based on MIDI data. First, each line is randomly assigned to one of the four instruments: a clarinet, a flute, a violin, and a trumpet. After that, for each note in the line, a note sound sample with the closest average pitch is selected from the samples of the assigned instrument and used for that note. Details about the synthesis procedure can be found in [14]. Admittedly, the audio signals generated this way are a rough approximation of real recordings. But they show realistic spectral and temporal variations. Different instrument lines are mixed to 0 dB SNR for separation. The first 5-second signal of each piece is used for testing. The pitches of each instrument line are detected using Praat [15].

Fig. 2 shows an example of separated instrument lines. The top panel is the waveform of a mixture, created by mixing the clarinet line in Fig. 2(b) and the trumpet line in Fig. 2(d). Fig. 2(c) and Fig. 2(e) are the corresponding separated lines. The second separated line is very close to the original one while the first separated line has some noticeable differences for several notes. Sound demos can be found at http://www.cse.ohio-state.edu/~liyip/Research/Publication/2007/mss_demo.htm.

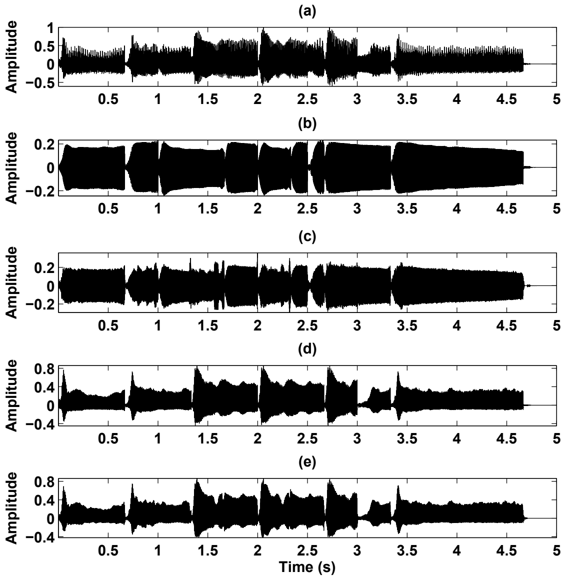

Fig. 2. An separation example. (a). A mixture. (b). The first line by a clarinet. (c). The separated first line. (d). The second line by a trumpet. (e). The separated second line.

We calculate SNR gain before and after separation to quantify the system's performance. To compensate for distortions introduced in the resynthesis stage, we pass a premixing signal through an all-one mask and use it as the reference signal for SNR calculation. Table 1 shows the SNR gains for different systems as well as when prior information is available.

The first row in Table 1 gives the SNR gain by the Hu-Wang system, an effective CASA system for voiced speech separation. The Hu-Wang system has the same time-frequency decomposition as ours, but following that it implements the two stages of segmentation and grouping, and utilizes pitch and amplitude modulation as organizational cues for separation. The Hu-Wang system has a mechanism to detect the pitches of one source. For comparison purposes, we supply the system with ideal pitch and adjust the filter bandwidth to be the same as ours. Although the Hu-Wang system performs well on voiced speech separation [16], our experiment shows that it is inadequate for musical sound separation. The proposed system, the fourth row in Table 1, performs 2.1 dB better than theirs by utilizing both pitches.

Virtanen's system is based on sinusoidal modeling. At each frame, his system uses pitch information and least mean square estimation to simultaneously estimate the amplitudes and phases of harmonics of all instruments. His system also uses a so-called adaptive frequency-band model to recover each individual harmonic for overlapping harmonics [1]. To avoid inaccurate implementation of his system, we sent our test signals to him and he provided the output. Note that his

Methods	SNR Gain (dB)
Hu and Wang	9.1
Virtanen	11.1
Proposed	**11.2**
Ideal Binary Mask	15.3

Table 1. SNR Gains

results are also obtained using ideal pitch. The average SNR gain of his system is shown in the third row of Table 1. Despite our simple use of pitch information and lack of handling overlapping harmonics, our system achieves the same performance as his.

Since our system is based on binary masking, we calculate the SNR gain of ideal binary masks which can be constructed with premixing instrument lines as following. After T-F decomposition, the energy of each T-F unit is calculated. A T-F unit is labeled 1 if the energy of the first instrument line is stronger than that of the second, and 0 otherwise. Although overlapping harmonics are not separated by the ideal binary mask, it can still give high-quality sounds free of artifacts. The SNR gain of the ideal binary mask is shown in the last row of Table 1. Compared to the ideal binary mask, the proposed system performs 4.1 dB worse, which suggests room for further improvement.

4. CONCLUSION

In this paper, we have proposed a CASA system for monaural musical sound separation. We label each T-F unit solely based on the values of the autocorrelation function at time lags corresponding to two pitch periods. The SNR evaluation shows the proposed system is as effective as more complicated sinusoidal model-based systems. Besides auditory filtering, the main computation of our system is to obtain the values of autocorrelation at two time lags at each T-F unit. Note that the calculation of a full correlogram is unnecessary, i.e., the system does not need to calculate autocorrelation for all possible time lags. We believe there is considerable room to improve our system. For example, segmentation and grouping, the two stages widely adopted in CASA, can be applied to make unit labeling more reliable. One can also first identify T-F units that are reliably labeled and use those T-F units to further process unreliable T-F units. We will pursue these directions in our future study.

Acknowledgments. We thank T. Virtanen for his assistance in sound separation and comparison. This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an NSF grant (IIS-0534707).

5. REFERENCES

[1] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, Tampere University of Technology, 2006.

[2] T. L. Hubbard and D. L. Datteri, "Recognizing the component tones of a major chord," *American Journal of Psychology*, vol. 114, no. 4, pp. 569–589, 2001.

[3] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1845–1856, 2006.

[4] S. A. Abdallah, "Towards music perception by redundancy reduction and unsupervised learning in probablistic models," Ph.D. dissertation, King's College London, Department of Electronic Engineering, 2002.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[6] P. Smaragdis, "Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Blind Signal Separation*, 2004, pp. 494–499.

[7] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[8] D. K. Mellinger, "Event formation and separation in musical sound," Ph.D. dissertation, Stanford University, Department of Computer Science, 1991.

[9] G. J. Brown and M. P. Cooke, "Perceptual grouping of musical sounds: A computational model," *Journal of New Music Research*, vol. 23, pp. 107–132, 1994.

[10] D. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," *Speech Communication*, vol. 27, no. 4, pp. 351–366, 1999.

[11] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press, 2006.

[12] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University, Department of Electrical Engineering, 1985.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, 2003.

[14] Y. Li and D. L. Wang, "Pitch detection in polyphonic music using instrument tone models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. II.481–484.

[15] P. Boersma and D. Weenink. (2002) Praat: Doing phonetics by computer, version 4.0.26. [Online]. Available: http://www.fon.hum.uva.nl/praat

[16] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.