Technical Report TR39, July 2008
Technical Report: OSU-CISRC-7/08-TR39
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210-1277, USA
Web site: **http://www.cse.ohio-state.edu/research/techReport.html**
Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2008**
File in pdf format: **TR39.pdf**

Musical Sound Separation Based on Binary Time-Frequency Masking

Yipeng Li* and DeLiang Wang

Abstract

The problem of overlapping harmonics is particularly acute in musical sound separation, and has not been addressed adequately. We propose a monaural system based on binary time-frequency masking with an emphasis on robust decisions in time-frequency regions where harmonics from different sources overlap. Our computational auditory scene analysis system exploits the observation that sounds from the same source tend to have similar spectral envelopes. Quantitative results show that utilizing spectral similarity helps binary decision making in overlapped time-frequency regions and significantly improve separation performance.

I. Introduction

Monaural musical sound separation has received significant attention recently. Analyzing a musical signal is difficult in general due to the polyphonic nature of music, but extracting useful information from monophonic music is considerably easier. Therefore a musical sound separation system would be a very useful processing step for many audio applications, such as automatic music transcription, automatic instrument identification, music information retrieval, and object-based coding. A particularly interesting application of such a system is signal manipulation. After a polyphonic signal is decomposed to individual sources, modifications, such as pitch shifting and time stretching, can then be applied to each source independently. This provides infinite ways to alter the original signal and creates new sound effects [20].

An emerging approach for general sound separation exploits the knowledge from the human auditory system. In an influential book, Bregman proposed that the auditory system employs a process called *auditory scene analysis* (ASA) to organize an acoustic mixture into different

Y. Li* is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210-1277, USA. Phone: 1-614-292-7402; Fax: 1-614-292-2911; Email: li.434@osu.edu

D. L. Wang is with the Department of Computer Science and Engineering & Center of Cognitive Science, The Ohio State University, Columbus, OH, 43210-1277, USA. Phone: 1-614-292-6827; Fax: 1-614-292-2911; Email: dwang@cse.ohio-state.edu.

perceptual streams which correspond to different sound sources [3]. The perceptual process is believed to involve two main stages: The segmentation stage and the grouping stage [3]. In the segmentation stage, the acoustic input is decomposed into T-F segments, each of which mainly originates from a single source [22]. In the grouping stage, segments from the same source are grouped according to a set of grouping principles. Grouping has two types: *Primitive* grouping and *schema-based* grouping. The principles employed in primitive grouping include proximity in frequency and time, harmonicity/pitch, synchronous onset and offset, common amplitude/frequency modulation, and common spatial information. Human ASA has inspired researchers to investigate *computational auditory scene analysis* (CASA) for sound separation [22]. CASA exploits the intrinsic properties of sounds for separation and makes relatively minimal assumptions about specific sound sources. Therefore it shows considerable potential as a general approach to sound separation. Recent CASA-based speech separation systems have shown promising results in separating target speech from interference [22]. However, building a successful CASA system for musical sound separation is challenging, and a main reason is the problem of overlapping harmonics.

In a musical recording, sounds from different instruments are likely to have a harmonic relationship in pitch. Fig. 1 shows the score of the first four measures of a piece by J. S. Bach. The pitch intervals of pairs between the two lines in the first measurement are a minor third, a perfect fifth, a major third, a major sixth, and a major sixth. The corresponding pitch ratios are 6:5, 3:2, 5:4, 5:3, and 5:3, respectively. As can be seen, the two lines are in harmonic relationship most of the time. Since many instruments can produce relatively stable pitch, such as a piano, harmonics from different instruments may therefore overlap for some time. This is in sharp contrast with co-channel speech. In co-channel speech, harmonics from two speakers rarely overlap [19]. Even if they do, overlapping tends to occur briefly due to the dynamic nature of voicing. When frequency components from different sources cross each other, some time-frequency (T-F) units will have significant energy from both sources. A T-F unit is an element of a T-F representation, such as a spectrogram. In this case, existing CASA systems utilize the temporal continuity principle, or the “old plus new” heuristic [3], to estimate the contribution of individual overlapped frequency components [4]. Based on this principle, which states that the temporal and spectral changes of natural sounds are gradual, these systems obtain the properties of individual components in an overlapped T-F region, i.e., a set of contiguous T-F units where two or more harmonics overlap, by linearly interpolating the properties in neighboring non-overlapped regions. The temporal continuity principle works reasonably well when overlapping is brief in time. However, it is not suitable when overlapping is relatively long as in music. Moreover, temporal continuity is not applicable in cases when harmonics of two sounds overlap completely from onset to offset.

As mentioned, overlapping harmonics are not as common in speech mixtures as in polyphonic music. This problem has not received much attention in the CASA community. Even those CASA systems specifically developed for musical sound separation [16], [5] do not address the problem explicitly.

In this paper, we present a monaural CASA system that explicitly addresses the problem of overlapping harmonics for 2-source separation. Our goal is to determine in overlapped T-F regions which harmonic is dominant and make binary pitch-based labeling accordingly. Therefore we follow a general strategy in CASA that allocates T-F energy to individual sources exclusively. More specifically, our system attempts to estimate the ideal binary mask (IBM) [8], [7]. For a T-F unit, the IBM takes value 1 if the energy from target source is greater than that from interference

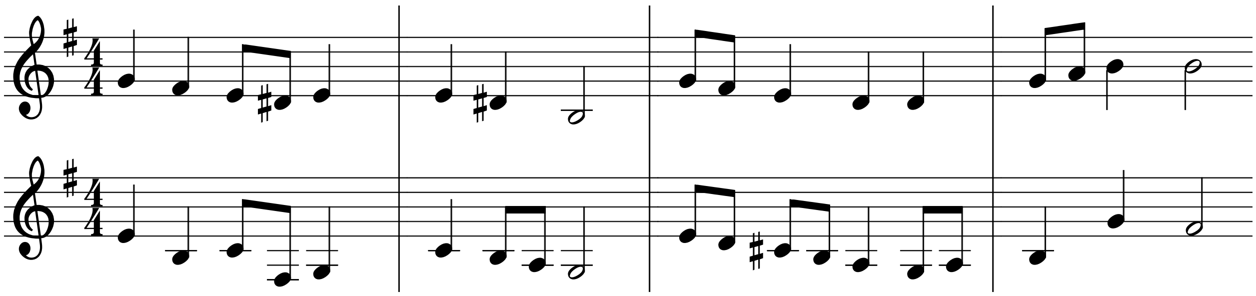

Fig. 1. Score of a piece by J. S. Bach. The first four measures are shown.

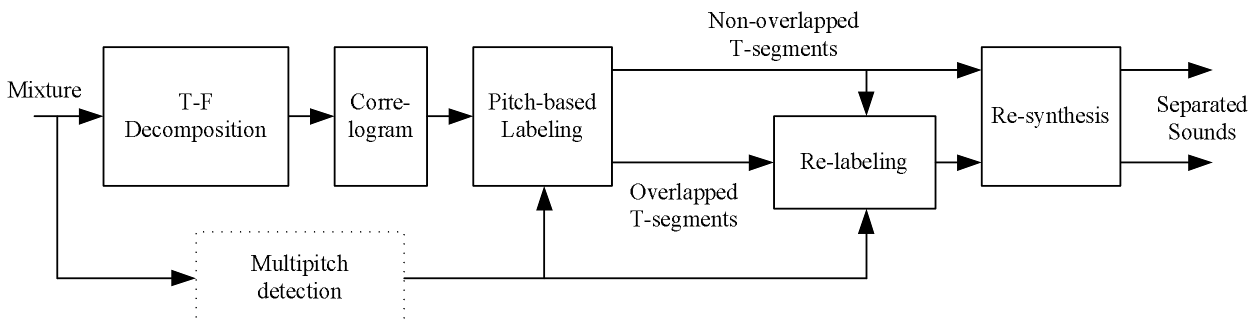

Fig. 2. Schematic diagram of the proposed CASA system for musical sound separation.

and 0 otherwise. The IBM was originally proposed as a main goal of CASA [21] and it is optimal in terms of signal-to-noise ratio gain among all the binary masks under certain conditions [15]. Compared to non-overlapped regions, making reliable binary decisions in overlapped regions is considerably more difficult. The key idea in the proposed system is to utilize contextual information available in a musical scene. Harmonics in non-overlapped regions, called non-overlapped harmonics, contain information that can be used to infer the properties of overlapped harmonics, i.e., harmonics in overlapped regions. Contextual information is extracted temporally, i.e., from notes played sequentially.

This paper is organized as follows. Section II provides the detailed description of the proposed system. Evaluation and comparison are presented in Section III. Section IV concludes the paper.

II. System Description

Our proposed system is illustrated in Fig. 2. The input to the system is a monaural polyphonic mixture consisting of two instrument sounds. In the 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, resulting in a matrix of T-F units. The next stage computes correlogram from the filter outputs. At the same time, the pitch contours of different instrument sounds are detected in the multipitch detection module. Multipitch detection for musical mixtures is a difficult problem because of the harmonic relationship of notes and huge variations of spectral shapes in instrument sounds [13]. Since the main focus of this study is to investigate the performance of pitch-based separation in music, we do not perform multiple pitch detection (indicated by the dashed box); instead we supply the system with pitch contours detected from premixed instrument sounds. In the pitch-based labeling stage, pitch points, i.e., pitch values at each frame, are used to determine which instrument each T-F unit should be assigned to. This creates a temporary binary mask for each instrument. After that, each T-segment, to be explained in Section II-C, is classified as overlapped or non-overlapped. Non-overlapped T-segments are directly passed to the resynthesis stage. For overlapped T-segments, the system exploits the information obtained from non-overlapped T-segments to decide which source is stronger and relabel accordingly. The system outputs instrument sounds resynthesized from the corresponding binary masks. The details of each stage are explained in the following subsections.

A. Time-Frequency Decomposition

In this stage, the input sampled at 20 kHz is first decomposed into its frequency components with a filterbank consisting of 128 gammatone filters (also called as channels). The impulse response of a gammatone filter is

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

where $l = 4$ is the order of the gammatone filter, f is the center frequency of the filter, and b determines the bandwidth of the filter [18] (see also [22]).

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 the above equation 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 speech energy is concentrated.

In most speech separation tasks, the bandwidth b of a fourth-order gammatone filter is usually set to be 1 ERB, i.e.,

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

where $\text{ERB}(f) = 24.7 + 0.108f$ is the equivalent rectangular bandwidth of a filter. This bandwidth is adequate when the intelligibility of separated speech is the main concern. However, for musical sound separation, the 1-ERB bandwidth appears too wide for analysis and resynthesis, especially in the high frequency range. We have found that using narrower bandwidths, which provide better frequency resolution, can significantly improve the quality of separated sounds. In this study we set the bandwidth to a quarter ERB. The center frequencies of channels are spaced from 50 to 8000 Hz. Hu [9] showed that a 128-channel gammatone filterbank with the bandwidth of 1 ERB per filter has a flat frequency response within the range of passband from 50 to 8000 Hz. Similarly, it can be shown that a gammatone filterbank with the same number of channels but the bandwidth of $1/4$ ERB per filter still provides a fairly flat frequency response over the same passband.

After auditory filtering, the output of each channel is divided into frames of 20 ms with a frame shift of 10 ms.

B. Correlogram

After T-F decomposition, the system computes a correlogram, $A(c, m, \tau)$, a well known mid-level auditory representation [22]. Specifically, $A(c, m, \tau)$ is computed as:

$$A(c,m,\tau) = \sum_{t=-T/2+1}^{T/2} r(c, m\frac{T}{2}+t) r(c, m\frac{T}{2}+t+\tau), \tag{4}$$

where r is the output of a filter. T is the frame length and $T/2$ is the frame shift. τ is the time lag. Similarly, a normalized correlogram, $\hat{A}(c, m, \tau)$, can be computed for T-F unit u_{cm} as:

$$\hat{A}(c,m,\tau) = \frac{\sum_{t=-T/2+1}^{T/2} r(c, m\frac{T}{2}+t) r(c, m\frac{T}{2}+t+\tau)}{\sqrt{\sum_{t=-T/2+1}^{T/2} r^2(c, m\frac{T}{2}+t)}\sqrt{\sum_{t=-T/2+1}^{T/2} r^2(c, m\frac{T}{2}+t+\tau)}}. \tag{5}$$

The normalization converts correlogram values to the range of $[-1, 1]$ with 1 at the zero time lag.

Several existing CASA systems for speech separation have used the envelope of filter outputs for autocorrelation calculation in the high frequency range, with the intention of encoding the beating phenomenon resulting from unresolved harmonics in high frequency (e.g. [7]). A harmonic is called resolved if there exists a frequency channel that primarily responds to it. Otherwise it is unresolved [7]. However, due to the narrower bandwidth used in this study, different harmonics from the same source will unlikely activate the same frequency channel. Fig. 3 plots the bandwidth corresponding to 1 ERB and $1/4$ ERB with respect to the channel number. From Fig. 3 we can see that the bandwidths of most filter channels are less than 100 Hz, smaller than the pitch ranges of most instruments. As a result, the envelope extracted would correspond to either the fluctuation of a harmonic's amplitude or the beating created by the harmonics from different sources. In both cases, the envelope information would be misleading. Therefore we do not extract envelope autocorrelation.

C. Pitch-based Labeling

After the correlogram is computed, we label each T-F unit u_{cm} using single-source pitch points detected from premixed sound sources. Since we are concerned only with 2-source separation, we consider at each T-F unit the values of $\hat{A}(c, m, \tau)$ at time lags that correspond to the pitch periods, d_1 and d_2, of the two sources. A natural choice is to compare $\hat{A}(c, m, d_1)$ and $\hat{A}(c, m, d_2)$ and assign the T-F unit accordingly, i.e.,

$$M_{cm} = \begin{cases} 1, & \text{if } \hat{A}(c,m,d_1) > \hat{A}(c,m,d_2), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Intuitively if source 1 has stronger energy at u_{cm} than source 2, the autocorrelation value at d_1 would be expected to be higher than that at d_2. However, due to the nonlinearity of the autocorrelation function and its sensitivity to the relative phases of harmonics, this intuition may not hold all the time. Nonetheless, empirical evidence shows that this labeling is reasonably accurate. It has been reported that when both pitch points are used for labeling as in Equation (6) for co-channel speech separation, the results are better compared to when only one pitch point is used for labeling [9]. Fig. 4 shows the percentage of correctly labeled T-F units for each

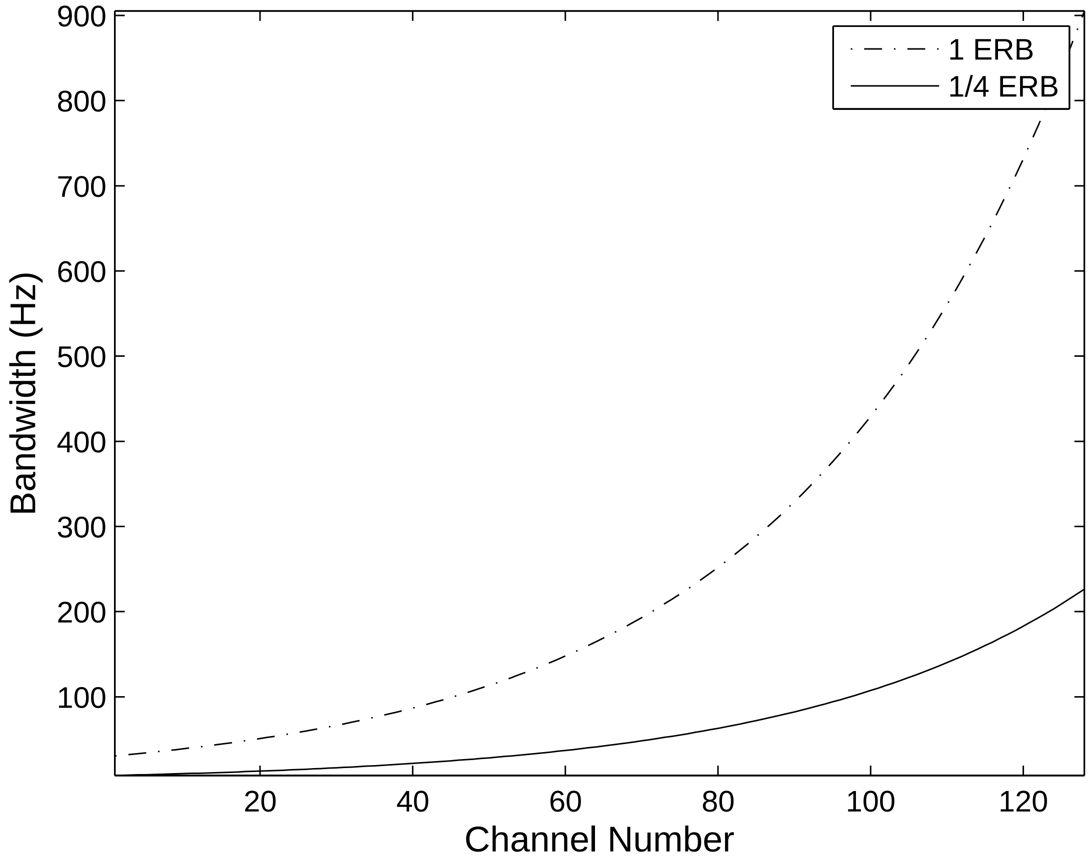

Fig. 3. Bandwidth in Hertz of gammatone filters in the filterbank. The dashed line indicates the 1 ERB bandwidth while the solid line indicates the 1/4 ERB bandwidth of the filters.

channel. We consider a T-F unit correctly labeled if labeling based on Equation (6) is the same as in the IBM. The plot is generated by comparing pitch-based labeling using Equation (6) to that of the IBM for all the musical pieces in our database (see Section III). It can be seen that labeling is well above the chance level for most of the channels. The poor labeling accuracy for channel numbers below 10 is due to the fact that the instrument sounds in our database have pitch higher than 125 Hz, which roughly corresponds to the center frequency of channel 10. The low-numbered channels contain little energy therefore labeling is not reliable.

Fig. 5 plots the percentage of correctly labeled T-F units according to Equation (6) with respect to the local energy ratio obtained from the same pieces as in Fig. 4. The local energy ratio is calculated as $|10\log_{10}\frac{E_1(c,m)}{E_2(c,m)}|$, where $E_1(c,m)$ and $E_2(c,m)$ are the energies of the two sources at u_{cm}. The local energy ratio is calculated using premixed signals. Note that the local energy ratio is measured in decibels and $|10\log_{10}\frac{E_1(c,m)}{E_2(c,m)}| = |10\log_{10}\frac{E_2(c,m)}{E_1(c,m)}|$. Hence the local energy ratio definition is symmetric with respect to the two sources. When the local energy ratio is high, one source is dominant and pitch-based labeling gives excellent results. A low local energy ratio indicates that two sources have close values of energy at u_{cm}. Since harmonics with

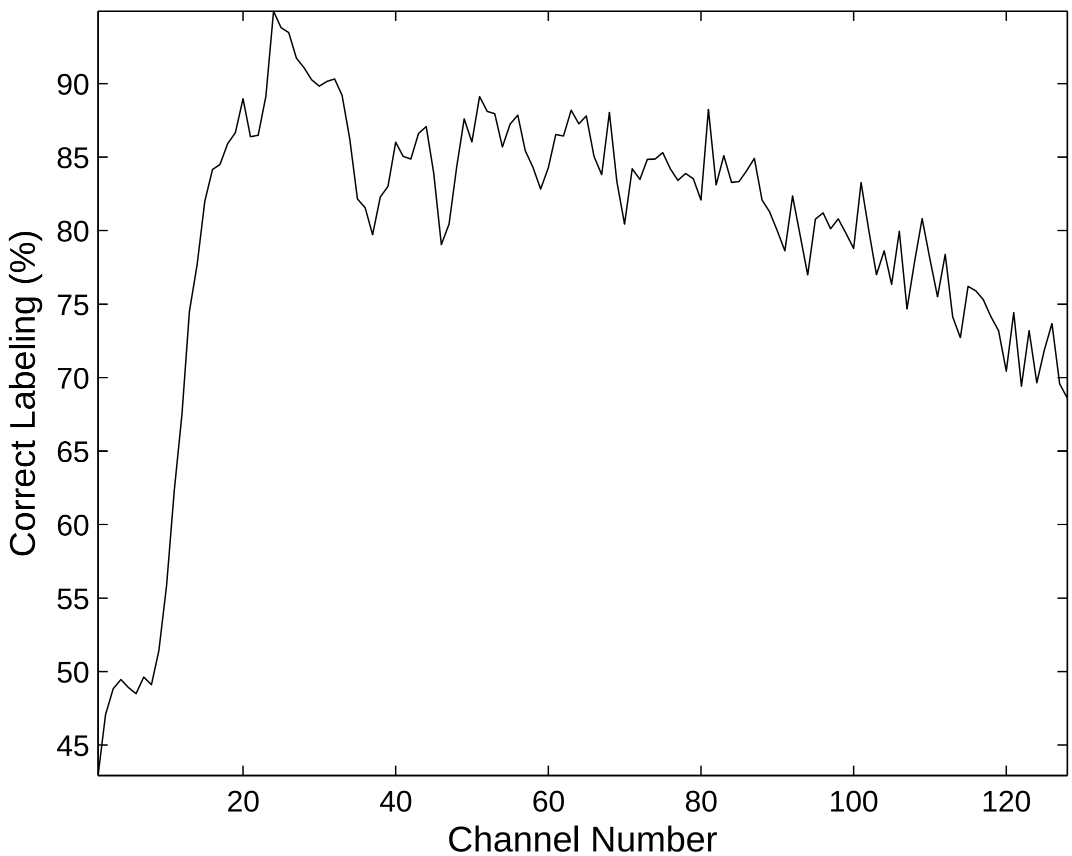

Fig. 4. The percentage of correctly labeled T-F units at each frequency channel

sufficiently different frequencies will not have close energy in the same frequency channel, a low local energy ratio also implies that in u_{cm} harmonics from two different sources have close (or the same) frequencies. As a result, the autocorrelation function will likely have close values at both pitch periods. In this case, the decision becomes unreliable and therefore the percentage of correct labeling is low.

Although this pitch-based labeling (Equation (6)) works well, it has two problems. The first problem is that the decision is made locally: The labeling of each T-F unit is independent of the labeling of its neighboring T-F units. Studies have shown that labeling on a larger auditory entity, such as a T-F segment, can often improve the performance. In fact, the emphasis of segmentation is considered as a unique aspect of CASA systems [22]. The second problem is overlapping harmonics. As mentioned, in T-F units where two harmonics from different sources overlap spectrally, unit labeling breaks down and the decision becomes unreliable. To address the first problem, we construct T-segments and find ways to make decisions based on T-segments instead of individual T-F units. For the second problem, we exploit the observation that sounds from the same source tend to have similar spectral envelopes.

The concept of T-segment is introduced in [9] (see also [10]). A segment is a set of contiguous

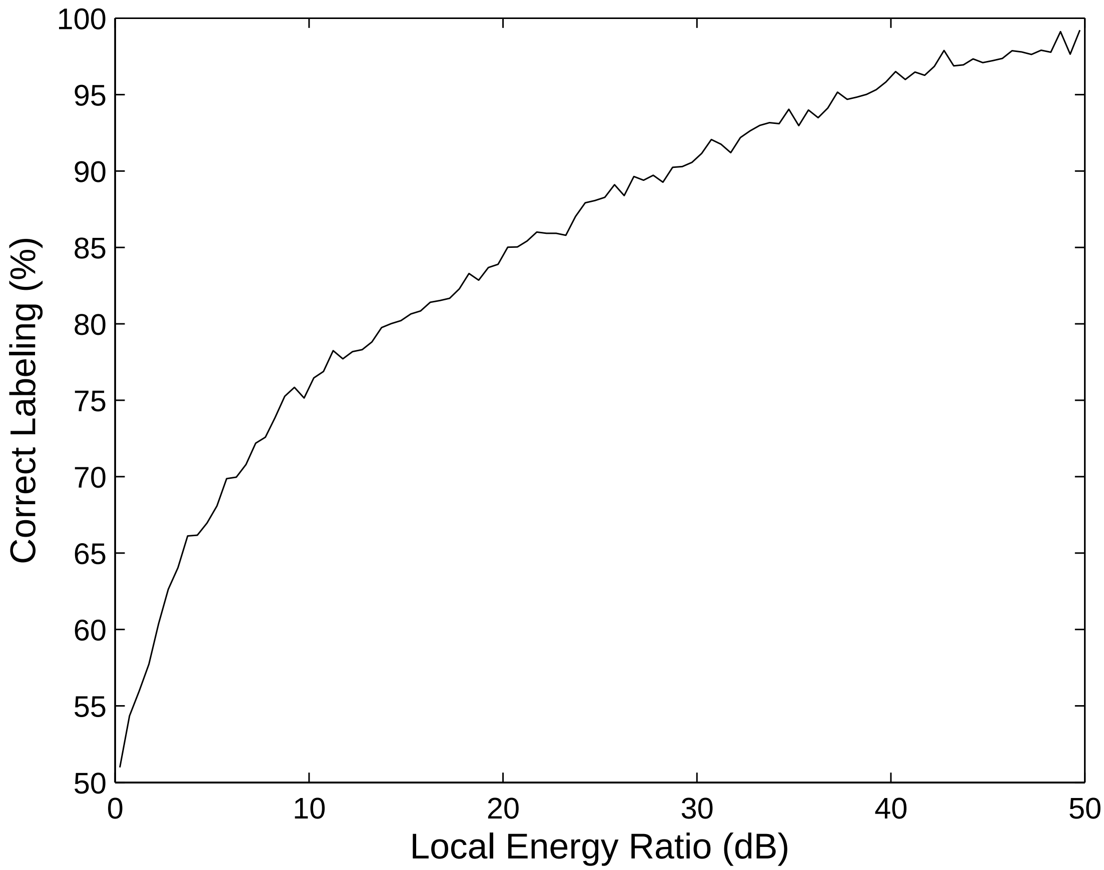

Fig. 5. The percentage of correctly labeled T-F units with respect to local energy ratio.

T-F units that are supposed to mainly originate from the same source. A T-segment is a segment in which all the T-F units have the same center frequency. Hu noted that using T-segments gives a better balance on rejecting energy from a target source and accepting energy from the interference than T-F segments [9]. In other words, compare to T-F segments, T-segments achieve a good compromise between false rejection and false acceptance. Since musical sounds tend to be stable, a T-segment naturally corresponds to a frequency component from its onset to offset. To get T-segments, we use pitch information to determine onset times. If the difference of two consecutive pitch points is more than one semitone, it is considered as an offset occurrence for the first pitch point and an onset occurrence for the second pitch point. The set of all the T-F units between an onset/offset pair of the same channel defines a T-segment.

For each T-segment, we first determine if it is overlapped or non-overlapped. If harmonics from two sources overlap at channel c, $\hat{A}(c,m,d_1) \approx \hat{A}(c,m,d_2)$. A T-F unit is considered overlapped if at that unit $|\hat{A}(c,m,d_1) - \hat{A}(c,m,d_2)| < \theta$, where θ is chosen to be 0.05. If half of the T-F units in a T-segment are overlapped, then the T-segment is considered overlapped; Otherwise, the T-segment is considered non-overlapped. With overlapped T-segments, we can also determine which harmonics of each source are overlapped. Given an overlapped T-segment at channel c, the frequency of the overlapping harmonics can be roughly approximated by the center frequency of the channel. Using the pitch contour of each source, we can identify the harmonic number of each overlapped harmonic. All other harmonics are considered non-overlapped.

Since each T-segment is supposedly from the same source, all the T-F units within a T-segment should have the same labeling. For each T-F unit within a non-overlapped T-segment, we perform labeling as follows:

$$M_{cm} = \begin{cases} 1, & \text{if } \sum_{u_{cm'} \in \mathbf{U}_1} A(c, m', 0) > \sum_{u_{cm'} \in \mathbf{U}_0} A(c, m', 0), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathbf{U}_1$ and $\mathbf{U}_0$ are the sets of T-F units previously labeled as 1 and 0, respectively, in the T-segment. The zero time lag of $A(c, m, \tau)$ indicates the energy of u_{cm}. Equation (7) means that, in a T-segment, if the total energy of the T-F units labeled as the first source is stronger than that of the T-F units labeled as the second source, all the T-F units in the T-segment are labeled as the first source; otherwise, they are labeled as the second source. Although this labeling scheme works for non-overlapped T-segments, it can not be extended to overlapped T-segments because the labeling of T-F units in an overlapped T-segment is not reliable.

D. Relabeling

To make binary decisions for an overlapped T-segment, it is helpful to know the energies of the two sources in that T-segment. One possibility is to use the spectral smoothness principle [12] to estimate the amplitude of an overlapped harmonic by interpolating its neighboring non-overlapped harmonics. However, the spectral smoothness principle does not hold well for many real instrument sounds. Another way to estimate the amplitude of an overlapped harmonic is to use an instrument model, which may consist of templates of spectral envelopes of an instrument [1]. However, instrument models of this nature unlikely work due to enormous intra-instrument variations of musical sounds. When training and test conditions differ, instrument models would be ineffective.

Intra-instrument variations of musical sounds result from many factors, such as different makers of the same instrument, different players, and different playing styles. However, in the same musical recording, the sound from the same source is played by the same player using the same instrument with typically the same playing style. Therefore we can reasonably assume that the sound from the same source in a musical recording shares similar spectral envelopes. As a result, it is possible to utilize the spectral envelope of some other sound components of the same source to estimate overlapped harmonics. Concretely speaking, consider an instrument playing notes $\mathcal{N}_1$ and $\mathcal{N}_2$ consecutively. Let the h^{th} harmonic of note $\mathcal{N}_1$ be overlapped by some other instrument sound. If the spectral envelopes of note $\mathcal{N}_1$ and note $\mathcal{N}_2$ are similar and harmonic h of $\mathcal{N}_2$ is reliable, the overlapped harmonic of $\mathcal{N}_1$ can be estimated. By having similar spectral envelopes we mean

$$\frac{a^1_{\mathcal{N}_1}}{a^1_{\mathcal{N}_2}} \approx \frac{a^2_{\mathcal{N}_1}}{a^2_{\mathcal{N}_2}} \approx \frac{a^3_{\mathcal{N}_1}}{a^3_{\mathcal{N}_2}} \approx \cdots , \tag{8}$$

where $a^h_{\mathcal{N}_1}$ and $a^h_{\mathcal{N}_2}$ are the amplitudes of the h^{th} harmonics of note $\mathcal{N}_1$ and note $\mathcal{N}_2$, respectively. In other words, the amplitudes of corresponding harmonics of the two notes are approximately proportional. Fig. 6 shows the log-amplitude spectrogram of a 5-second clarinet piece. There are

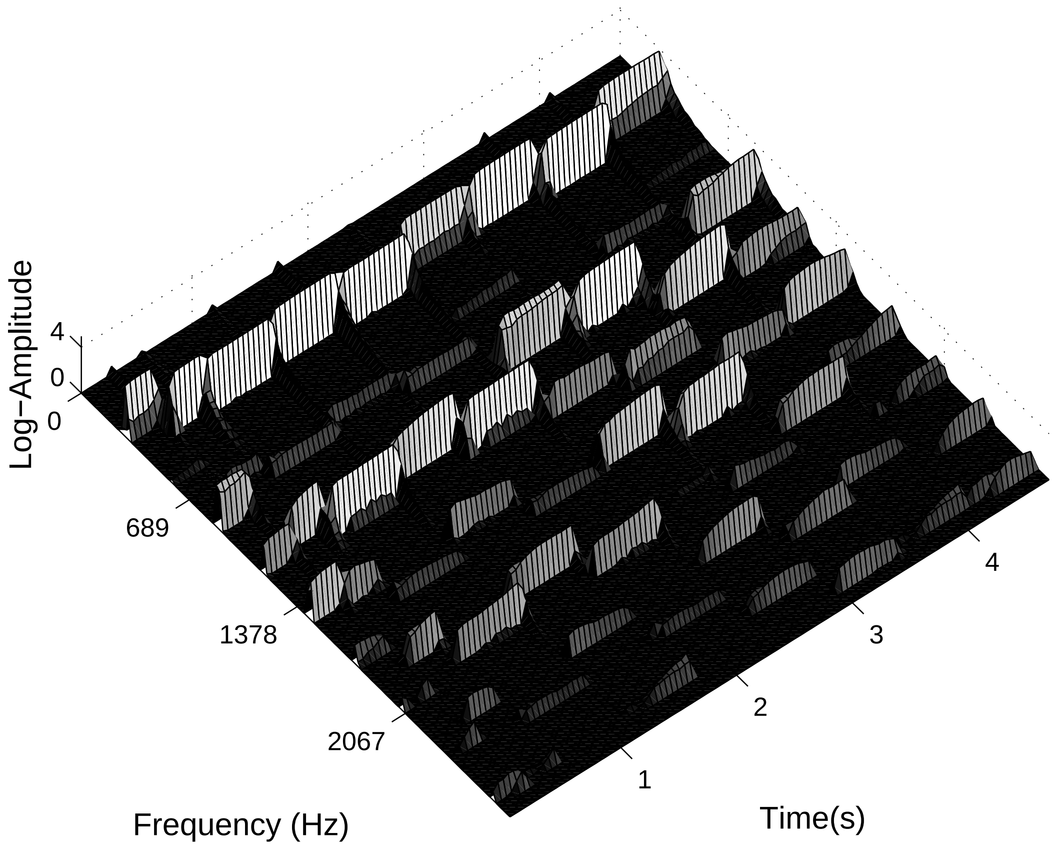

Fig. 6. Log-amplitude spectrogram of a 5-second clarinet performance.

9 notes in the piece. As can be seen, the average spectral shapes over the durations of individual notes are similar.

If the h^{th} harmonic of $\mathcal{N}_1$ is overlapped while the same-numbered harmonic of $\mathcal{N}_2$ is not, using Equation (8), we can estimate the amplitude of harmonic h of $\mathcal{N}_1$ as

$$a^h_{\mathcal{N}_1} \approx a^h_{\mathcal{N}_2} \frac{a^1_{\mathcal{N}_1}}{a^1_{\mathcal{N}_2}}. \tag{9}$$

In the above equation, we assume that the first harmonics of both notes are not overlapped. If the first harmonic of $\mathcal{N}_1$ is also overlapped, then all the harmonics of $\mathcal{N}_1$ will be overlapped. Currently our system is not able to handle this extreme situation. If the first harmonic of note $\mathcal{N}_2$ is overlapped, we try to find some other note which has the first harmonic and harmonic h reliable. Note from Equation (9) that with an appropriate note, the overlapped harmonic can be recovered from the overlapped region without the knowledge of the other overlapped harmonic. In other words, using temporal contextual information, it is possible to extract the energy of only one source.

It can be seen from Equation (9) that the key to estimating overlapped harmonics is to find a note with a similar spectral envelope. Given an overlapped harmonic h of note $\mathcal{N}_1$, one approach to finding an appropriate note is to search the neighboring notes from the same source. If harmonic h of a note is non-overlapped, then that note is chosen for estimation. However, it has been shown that spectral envelopes are pitch dependent [11] and related to dynamics of an instrument nonlinearly. To minimize the variations introduced by pitch as well as dynamics and improve the accuracy of binary decisions, we search notes within a temporal window and choose the one with the closest spectral envelope. Specifically, consider again note $\mathcal{N}_1$ with harmonic h overlapped. Within a temporal window, we first identify the set of non-overlapped harmonics, denoted as $\tilde{\mathbf{H}}_{\mathcal{N}}$, for each note $\mathcal{N}$ from the same instrument as note $\mathcal{N}_1$. We then check every $\mathcal{N}$ and find the harmonics which are non-overlapped between notes $\mathcal{N}_1$ and $\mathcal{N}$. This is to find the intersection of $\tilde{\mathbf{H}}_{\mathcal{N}}$ and $\tilde{\mathbf{H}}_{\mathcal{N}_1}$. After that, we calculate the correlation of the two notes, $\rho(\mathcal{N},\mathcal{N}_1)$, based on the amplitudes of the non-overlapped harmonics. The correlation is obtained by

$$\rho(\mathcal{N},\mathcal{N}_1) = \frac{\sum_h a^h_{\mathcal{N}} a^h_{\mathcal{N}_1}}{\sqrt{\sum_h (a^h_{\mathcal{N}})^2 \sum_h (a^h_{\mathcal{N}_1})^2}}. \tag{10}$$

After this is done for each such note $\mathcal{N}$, we choose the note $\mathcal{N}^*$ that has the highest correlation with note $\mathcal{N}_1$ and whose h^{th} harmonic is non-overlapped. The temporal window in general should be centered on a note being considered, and long enough to include multiple notes from the same source. However, in this study, since each test recording is 5-second long (see Section III), the temporal window is set to be the same as the duration of a recording.

The above procedure is illustrated in Fig. 7. In the figure, the note under consideration, $\mathcal{N}_1$, has its fourth harmonic (indicated by an open arrowhead) overlapped with a harmonic (indicated by a dashed line with an open square) from the other source. To uncover the amplitude of the overlapped harmonic, the non-overlapped harmonics (indicated by filled arrowheads) of note $\mathcal{N}_1$ are compared to the same harmonics of the other notes of the same source in a temporal window using Equation (10). In this case, note $\mathcal{N}^*$ has the highest correlation with note $\mathcal{N}_1$.

After the appropriate note is identified, the amplitude of h of note $\mathcal{N}_1$ is estimated according to Equation (9). Similarly, the amplitude of the other overlapped harmonic, $a^{h'}_{\mathcal{N}'}$ (i.e., the dashed line in Fig. 7), can be estimated. As mentioned before, the labeling of the overlapped T-segment depends on the relative overall energy of overlapping harmonics h and h'. If the overall energy of harmonic h in the T-segment is greater than that of harmonic h', all the T-F units in the T-segment will be labeled as source 1. Otherwise, they will be labeled as source 2. Since the amplitude of a harmonic is calculated as the square root of the harmonic's overall energy (see below), we label all the T-F units in the T-segment based on the relative amplitudes of the two harmonics, i.e., all the T-F units are labeled as 1 if $a^h_{\mathcal{N}_1} > a^{h'}_{\mathcal{N}'}$ and 0 otherwise.

The above procedure requires the amplitude information of each non-overlapped harmonic. This can be obtained by using single-source pitch points and the activation pattern of gammatone filters. For harmonic h, we use the median pitch points of each note over the time period of a T-segment to determine the frequency of the harmonic. We then identify which frequency channel is most strongly activated. If the T-segment in that channel is not overlapped, then the harmonic amplitude is taken as the square root of the overall energy over the entire T-segment.

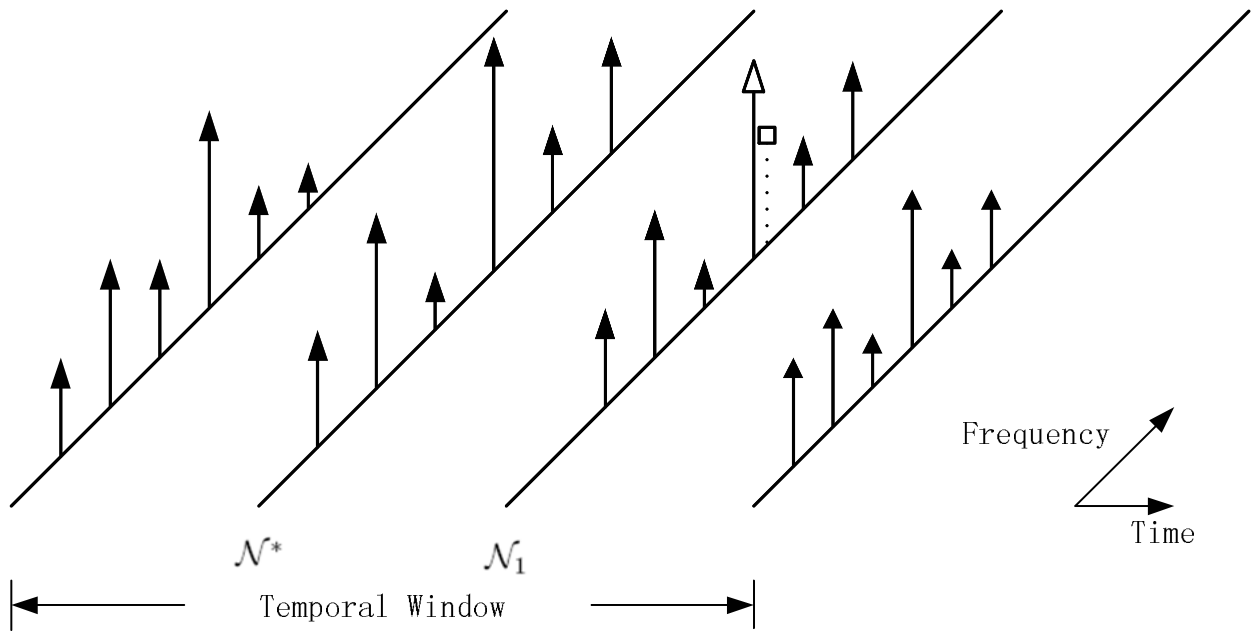

Fig. 7. Illustration of identifying the note for amplitude estimation of overlapped harmonics.

E. Resynthesis

The resynthesis is performed using a technique introduced by Weintraub [23] (see also [22]). During the resynthesis, the output of each filter is first phase-corrected and then divided into time frames using a raised cosine with the same frame size used in T-F decomposition. The responses of individual T-F units are weighted according to an obtained binary mask and summed over all the frequency channels and time frames to produce a reconstructed audio signal. The resynthesis pathway allows the quality of separated lines to be assessed quantitatively.

III. Evaluation and Comparison

To evaluate the proposed system, we construct 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 [6] 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 points is selected from the samples of the assigned instrument and used for that note. Details about the synthesis procedure can be found in [13]. 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 with equal energy. The first 5-second signal of each piece is used for testing. The pitch contour of each instrument line is detected using Praat [2].

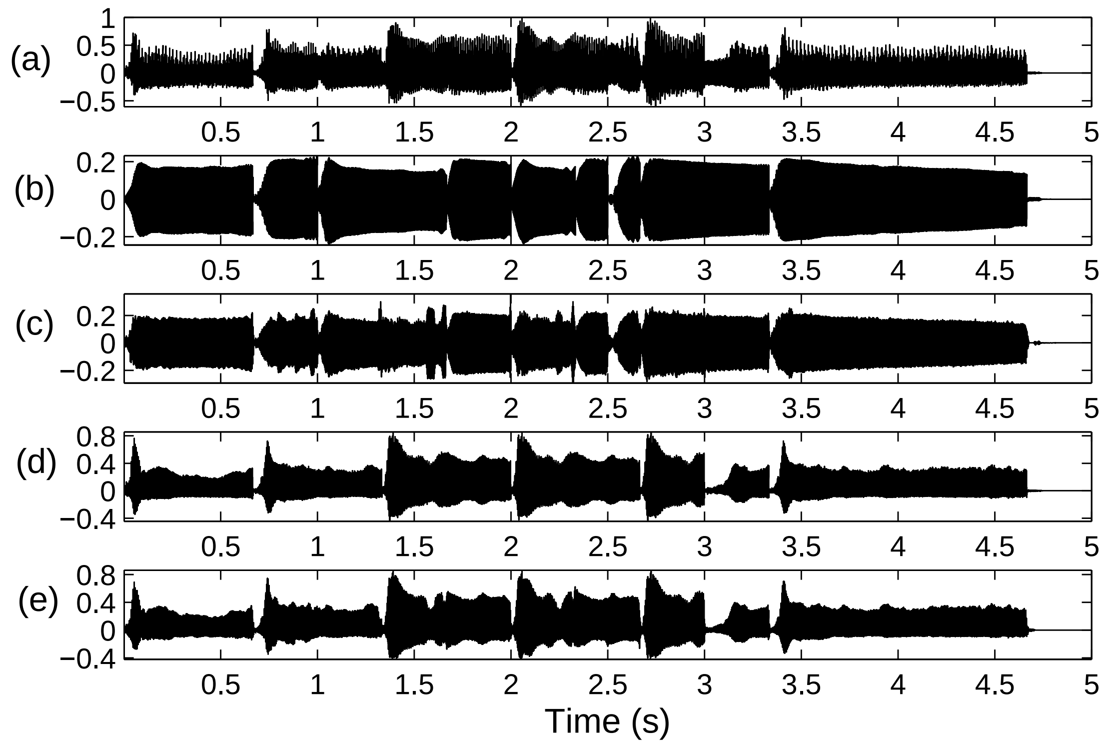

Fig. 8. An separation example. (a). A mixture. (b). The first line by a clarinet in the mixture. (c). The separated first line. (d). The second line by a trumpet in the mxiture. (e). The separated second line.

Fig. 8 shows an example of separated instrument lines. The top panel is the waveform of a mixture, created by mixing the clarinet line in Fig. 8(b) and the trumpet line in Fig. 8(d). Fig. 8(c) and Fig. 8(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/pnl/demo/LiBinary.html.

We calculate SNR gain by comparing the performance before and after separation to quantify the system's results. To compensate for possible distortions introduced in the resynthesis stage, we pass a premixed signal through an all-one mask and use it as the reference signal for SNR calculation [7]. In this case, the SNR is defined as:

$$\mathrm{SNR} = 10\log_{10}[\frac{\sum_t x^2_{ALL-ONE}(t)}{\sum_t (x_{ALL-ONE}(t) - \hat{x}(t))^2}], \tag{11}$$

where $x_{ALL-ONE}(t)$ is the signal after all-one mask compensation. In calculating the SNR after separation, $\hat{x}(t)$ is the output of the separation system. In calculating the SNR before separation, $\hat{x}(t)$ is the mixture resynthesized from an all-one mask. We calculate the SNR difference for each separated sound and take the average. Results are shown in the first row of Table I in terms of SNR gain after separation. Our system achieves an average SNR gain of 12.6 dB.

We compare the performance of our system with those of related systems. The second row in Table I gives the SNR gain by the Hu-Wang system, an effective CASA system designed for

Separation Methods	SNR Gain (dB)
Proposed System	12.6
Hu and Wang (2004)	9.1
Virtanen (2006)	11.0
Parsons (1976)	10.6
Ideal Binary Mask	15.3
2-Pitch Labeling	11.3
2-Pitch Labeling (Ideal Segmentation)	13.1
Spectral Smoothness	9.8

TABLE I
SNR GAIN (IN DECIBELS) OF THE PROPOSED CASA SYSTEM AND RELATED SYSTEMS

voiced speech separation. The Hu-Wang system has similar time-frequency decomposition to ours, 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 pitch contour of one voiced source. For comparison purposes, we supply the system with single-source pitch contours and adjust the filter bandwidths to be the same as ours. Although the Hu-Wang system performs well on voiced speech separation [7], our experiment shows that it is not every effective for musical sound separation. Our system outperforms theirs by 3.5 dB.

We also compare with Virtanen's system which is based on sinusoidal modeling [20]. At each frame, his system uses pitch information and least mean square estimation to simultaneously estimate the amplitudes and phases of the harmonics of all instruments. His system also uses a so-called adaptive frequency-band model to recover each individual harmonic from overlapping harmonics [20]. To avoid inaccurate implementation of his system, we sent our test signals to him and he provided the output. Note that his results are also obtained using single-source pitch contours. The average SNR gain of his system is shown in the third row of Table I. Our system's SNR gain is higher than Virtanen's system by 1.6 dB. In addition, we compare with a classic pitch-based separation system developed by Parsons [17]. Parsons's system is one of the earliest that explicitly addresses the problem of overlapping harmonics in the context of separating co-channel speech. Harmonics of each speech signal are manifested as spectral peaks in the frequency domain. Parsons's system separates closely spaced spectral peaks and performs linear interpolation for completely overlapped spectral peaks. Note that for Parsons's system we also provide single-source pitch contours. As shown in Table I the Parsons system achieves an SNR gain of 10.6 dB, which is 2.0 dB smaller than the proposed system.

Since our system is based on binary masking, it is informative to compare with the SNR gain of the IBM which is constructed from premixed instrument sounds. Although overlapping harmonics are not separated by ideal binary masking, the SNR gain is still very high, as shown in the fifth row of Table I. There are several reasons for the performance gap between the proposed system and the ideal binary mask. One is that pitch-based labeling is not error-free. Second, a T-segment can be mistaken, i.e., containing significant energy from two different sources. Also using contextual information may not always lead to the right labeling of a T-segment.

If we simply apply pitch-based labeling and ignore the problem of overlapping harmonics, the SNR gain is 11.3 dB as reported in [14]. The 1.3 dB improvement of our system over the previous one shows the benefit of using contextual information to make binary decisions.

We also consider the effect of segmentation on the performance. We supply the system with ideal segments, i.e, segments from the IBM. After pitch-based labeling, a segment is labeled by comparing the overall energy from one source to that from the other source. In this case, the SNR is 13.1 dB. This shows that if we had access to ideal segments, the separation performance could be further improved. Note that the performance gap between ideal segmentation and the IBM exists mainly because ideal segmentation does not help in the labeling of the segments with overlapped harmonics.

As the last comparison, we apply the spectral smoothness principle [12] to estimate the amplitude of overlapped harmonics from concurrent non-overlapped harmonics. We use linear interpolation for amplitude estimation and then compare the estimated amplitudes of overlapped harmonics to label T-segments. In this case, the SNR gain is 9.8 dB, which is considerably lower than that of the proposed system. This suggests that the spectral smoothness principle is not very effective in this case.

IV. Conclusion

In this paper, we have proposed a CASA system for monaural musical sound separation. We first label each T-F unit based on the values of the autocorrelation function at time lags corresponding to two underlying pitch periods. We adopt the concept of T-segments for more reliable estimation for non-overlapped harmonics. For overlapped harmonics, we analyze the musical scene and utilize the contextual information from notes of the same source. Quantitative evaluation shows the proposed system yields large SNR gain and performs better than related separation systems.

Acknowledgment

The authors would like to thank T. Virtanen for his assistance in sound separation and comparison, J. Woodruff for his help in figure preparation, and E. Fosler-Lussier for useful comments. This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

References

[1] M. Bay and J. W. Beauchamp, “Harmonic source separation using prestored spectra,” in *Independent Component Analysis and Blind Signal Separation*, 2006, pp. 561–568.

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

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

[4] M. P. Cooke and G. J. Brown, “Computational auditory scene analysis: Exploiting principles of perceived continuity,” *Speech Communication*, vol. 13, pp. 391–399, 1993.

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

[6] M. Goto, “Analysis of musical audio signals,” in *Computational Auditory Scene Analysis*, D. L. Wang and G. J. Brown, Eds. John Wiley & Sons, 2006.

[7] 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.

[8] ——, “Speech segregation based on pitch tracking and amplitude modulation,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[9] G. Hu, “Monaural speech organization and segregation,” Ph.D. dissertation, The Ohio State University, 2006.

[10] G. Hu and D. L. Wang, “Segregation of unvoiced speech from nonspeech interference,” *Journal of the Acoustical Society of America*, 2008, in press.

[11] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno, “Instrument identification in polyphonic music: feature weighting with mixed sounds, pitch-dependent timbre modeling, and use of musical context,” in *International Conference on Music Information Retrieval*, 2005, pp. 558–563.

[12] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[13] 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.

[14] ——, “Musical sound separation using pitch-based labeling and binary time-frequency masking,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[15] ——, “On the optimality of ideal binary time-frequency masks,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

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

[17] T. W. Parsons, “Separation of speech from interfering speech by means of harmonic selection,” *Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[18] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” MRC Applied Psychology Unit, Cambridge, U.K., Tech. Rep., 1988.

[19] Y. Shao and D. L. Wang, “Model-based sequential organization in cochannel speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 289–298, 2006.

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

[21] D. L. Wang, “On ideal binary masks as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, pp. 181–197.

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

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