Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation

Yipeng Li, John Woodruff, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural musical sound separation has been extensively studied recently. An important problem in separation of pitched musical sounds is the estimation of time–frequency regions where harmonics overlap. In this paper, we propose a sinusoidal modeling-based separation system that can effectively resolve overlapping harmonics. Our strategy is based on the observations that harmonics of the same source have correlated amplitude envelopes and that the change in phase of a harmonic is related to the instrument's pitch. We use these two observations in a least squares estimation framework for separation of overlapping harmonics. The system directly distributes mixture energy for harmonics that are unobstructed by other sources. Quantitative evaluation of the proposed system is shown when ground truth pitch information is available, when rough pitch estimates are provided in the form of a MIDI score, and finally, when a multipitch tracking algorithm is used. We also introduce a technique to improve the accuracy of rough pitch estimates. Results show that the proposed system significantly outperforms related monaural musical sound separation systems.**

***Index Terms*—Common amplitude modulation (CAM), musical sound separation, sinusoidal modeling, time–frequency masking, underdetermined sound separation.**

I. Introduction

MUSICAL sound separation attempts to isolate individual instruments from a polyphonic mixture. In recent years, this problem has attracted attention as the demand for automatic analysis, organization, and retrieval of a vast amount of online music data has exploded. A solution to this problem allows for efficient audio coding, accurate content-based analysis, and sophisticated manipulation of musical signals [18], [25], [31], [39]. In this paper, we address the problem of monaural musical sound separation, where multiple harmonic instruments are recorded by a single microphone or mixed to a single channel.

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

Sinusoidal modeling assumes a sound to be a linear combination of sinusoids with time-varying frequencies, amplitudes, and phases. Consequently, the task of sound separation becomes estimating these parameters for each sound source in the mixture [11], [31]. Sinusoidal modeling is often used for separating harmonic sounds when the pitch contour of each source is known *a priori* or can be estimated accurately.

Statistical techniques for musical sound separation generally assume certain statistical properties of sound sources. Independent subspace analysis (ISA) [9] extends independent component analysis, which assumes statistical independence among sources, to single-channel source separation. Sparse coding assumes that a source is a weighted sum of bases from an overcomplete set. The weights are assumed to be mostly zero, i.e., most of the bases are inactive most of the time [2]. Nonnegative matrix factorization (NMF) attempts to find a mixing matrix and a source matrix with non-negative elements such that the reconstruction error is minimized. It implicitly requires the mixing weights to be sparse [20]. Several recent systems [29], [32] have demonstrated the applicability of these techniques to musical sound separation.

Computational auditory scene analysis (CASA) [36] is inspired by auditory scene analysis [5], an influential perceptual theory which attempts to explain the remarkable capability of the human auditory system in selective attention. CASA aims to build computational systems for general sound separation [36], but several CASA systems have been developed specifically for monaural musical sound separation [6], [13], [22], [24].

A central problem in separation of pitched musical sounds is overlapping harmonics. Two harmonics of different instruments overlap when their frequencies are the same or close. Since Western music favors the twelve-tone equal temperament scale [7], common musical intervals have pitch relationships very close to simple integer ratios ($\approx 3/2$, 4/3, 5/3, 5/4, etc.). As a consequence, a large number of harmonics of a given source may be overlapped by another source in a mixture. For example, in a perfect fifth relationship (3/2), one source has every second harmonic overlapped while the other has every third overlapped. An adequate musical separation system must reliably handle overlapping harmonics.

Almost all music separation systems transform a mixture to some time–frequency (T–F) representation such as a spectrogram. Overlapping harmonics result in T–F regions that contain significant energy from multiple sources. Existing CASA-based separation systems [6], [13], [22], [24] allocate energy

Manuscript received November 10, 2008; revised March 23, 2009. Current version published July 17, 2009. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-08-1-0155) and in part by the National Science Foundation under Grant IIS-0534707. A preliminary version of this work was presented at the 9th International Conference on Music Information Retrieval [37]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Chen.

Y. Li and J. Woodruff are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210-1277 USA (e-mail: li.434@osu.edu; woodrufj@cse.ohio-state.edu).

D. L. Wang is with the Department of Computer Science and Engineering and Center of Cognitive Science, The Ohio State University, Columbus, OH, 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2009.2020886

1558-7916/$25.00 © 2009 IEEE

exclusively to one source and make no attempt to separate overlapping harmonics. Therefore, their separation performance for musical sounds is limited. Systems based on ISA, sparse coding, or NMF, handle overlapping harmonics implicitly. These systems operate in the magnitude domain and rely on the observed magnitudes in overlapped T–F regions to recover individual harmonics [9], [2], [29], [32]. Such systems are not expected to achieve optimal performance because they ignore the relative phases of the overlapping harmonics, which play a critical role in the observed magnitude spectrum. For example, assume that two overlapping harmonics have the same frequency and the amplitudes of the individual harmonics are a_1 and a_2, respectively. In this case, the amplitude of the observed harmonic is $a = |a_1 + a_2 e^{i\Delta\phi}|$, where $\Delta\phi$ is the relative phase of the two harmonics. If $\Delta\phi = 0$, then $a = |a_1 + a_2|$. However, if $\Delta\phi = \pi$, the observed amplitude, $a = |a_1 - a_2|$, is significantly different. Consequently, the observed magnitude spectrum in the overlapped region will be different depending on the relative phase; thus, the phase information must be considered in order to accurately recover individual harmonics from the overlapped regions.

One of the earliest systems that explicitly addresses the problem of overlapping harmonics was proposed by Parsons [26] for co-channel speech separation. Harmonics of each speech signal correspond to spectral peaks in the frequency domain. The Parsons system first identifies composite peaks where two peaks overlap and then uses the spectral shape of the analysis window to reconstruct one of the peaks, assuming that one harmonic is dominant. The other peak is recovered by subtracting the reconstructed peak from the composite peak. The system performs separation solely based on the observed magnitudes in overlapped regions and therefore is subject to the aforementioned phase problem. It also fails when the overlapping harmonics have the same frequency or close amplitudes.

Realizing that the information in overlapped regions is unreliable, several recent systems attempt to utilize the information of the neighboring non-overlapped harmonics. These systems assume that the spectral envelope of instrument sounds is smooth [19]. Based on this assumption, the amplitude of an overlapped harmonic can be estimated from the amplitudes of neighboring non-overlapped harmonics of the same source. For example, Virtanen and Klapuri [33] estimated an overlapped harmonic through nonlinear interpolation of neighboring harmonics. Every and Szymanski [11] used linear interpolation instead. Recently, Virtanen [31] proposed a system that directly imposes spectral smoothness by modeling amplitudes of harmonics as a weighted sum of fixed basis functions having smooth spectral envelopes. However, for real instrument sounds, the spectral smoothness assumption is often violated.

Another way to deal with overlapping harmonics is to use instrument models that contain the relative amplitudes of harmonics [3]. However, instrument models of this nature are limited because harmonic amplitude relationships are not consistent between recordings of different pitches, playing styles, and even different builds of the same instrument type.

Although in general, the absolute value of the amplitude of a harmonic with respect to its neighboring harmonics is difficult to model, the amplitude envelopes of different harmonics of the same source tend to be similar. This is known as common amplitude modulation (CAM) and it is an important organizational cue in human auditory perception [5] and has been used in CASA-based systems [35]. It also has a long history in musical instrument synthesis [27]. Although it has been utilized for stereo musical sound separation [34], [38], to our knowledge, this cue has not been applied in existing monaural musical sound separation systems. In this paper, we demonstrate how CAM can be used to resolve overlapping harmonics. In the proposed separation system, we use CAM within a sinusoidal model and show that both the amplitudes and phases of overlapping harmonics can be accurately estimated in a least-squares framework. Non-overlapping harmonics are estimated using a binary masking approach that directly distributes mixture energy.

This paper is organized as follows. In Section II, we present the sinusoidal model of harmonic instruments, provide empirical evidence for the CAM assumption, and discuss how pitch information is used to estimate the change in phase of sinusoidal parameters. Section III presents the detailed description of the proposed separation system. In Section IV, we provide quantitative evaluation of our system. Section V concludes the paper.

II. Background

A. Sinusoidal Modeling

Sinusoidal modeling is a well established technique in audio synthesis and signal processing [23], [28]. It models a sound source as the summation of individual sinusoidal components. Specifically, within an analysis frame with index m where the frequencies and amplitudes of the sinusoids are assumed constant, the sinusoidal model of a signal can be written as

$$x_i^{(m)}[n] = \sum_{h_i=1}^{H_i} a_i^{h_i}(m) \cos\left(2\pi f_i^{h_i}(m) n T_n + \phi_i^{h_i}(m)\right) \qquad (1)$$

where $a_i^{h_i}(m)$ and $f_i^{h_i}(m)$ are the amplitude and frequency, respectively, of sinusoidal component h_i of source i within time frame m, and $\phi_i^{h_i}(m)$ is the phase of sinusoidal component h_i of source i at the beginning of time frame m. H_i denotes the number of sinusoidal components in source i and T_n denotes the sampling period in seconds. The sinusoidal model of $x_i^{(m)}[n]$ can be transformed to the time-frequency domain by the discrete Fourier transform (DFT) using an analysis window $w[n]$. The DFT of $x_i^{(m)}[n]$, windowed by $w[n]$, at frequency bin k is

$$\begin{aligned} X_i(m,k) = \sum_{h_i=1}^{H_i} \frac{a_i^{h_i}(m)}{2} \Big(& e^{j\phi_i^{h_i}(m)} W\left(k f_b - f_i^{h_i}(m)\right) \\ & + e^{-j\phi_i^{h_i}(m)} W\left(k f_b + f_i^{h_i}(m)\right)\Big) \end{aligned} \qquad (2)$$

where W is the discrete-time Fourier transform (DTFT) of the analysis window, or

$$W(f) = \sum_{n=0}^{N-1} w[n] e^{-j2\pi(f/f_s)n}. \qquad (3)$$

In (2), $f_b = f_s/N$ is the frequency resolution of the DFT, where f_s denotes the sampling frequency and N is the length of the DFT.

For a perfectly harmonic sound, $f_i^{h_i}(m) = h_iF_i(m)$, where $F_i(m)$ denotes the pitch of source i at time frame m. If we assume that $|W(f)| \approx 0$ for $|f| > \theta_1$, where θ_1 is a threshold in Hz, then $|W(kf_b + f_i^{h_i}(m))| \approx 0$ provided that $F_i(m) > \theta_1$ for all sources i and time frames m. Further, if $F_i(m) > 2\theta_1$, then $\left|W\left(kf_b - f_i^{h_i}(m)\right)\right| > 0$ for at most one harmonic of each source, allowing us to drop the summation over harmonics from (2). We discuss how θ_1 is set in Section III-B, but for now let us assume that harmonic h_i is the only harmonic with appreciable energy in frequency bin k. Given the above assumptions, we can simplify (2) as

$$X_i(m,k) \approx \frac{a_i^{h_i}(m)}{2} e^{j\phi_i^{h_i}(m)} W(kf_b - h_iF_i(m)). \qquad (4)$$

Assuming that the mixing process is linear, the sinusoidal model of a mixture of I harmonic sound sources in the time–frequency domain can be written as

$$Z(m,k) = \sum_{i=1}^{I} X_i(m,k). \qquad (5)$$

This model treats a polyphonic mixture as a collection of harmonic components from multiple sound sources. Given the pitch contour of each source, the task of musical sound separation is to estimate $\left\{a_i^{h_i}(m), \phi_i^{h_i}(m)\right\}$ for all the harmonic components of the I sources. As discussed in Section I, this is a challenging problem for harmonics that are overlap in the mixture.

B. Common Amplitude Modulation

CAM assumes that the amplitude envelopes of spectral components from the same source are correlated. In this section, we show empirical evidence that suggests this assumption holds most of the time for harmonics, especially the ones with strong energy, of many instrument sounds. We calculate the correlation between harmonics of 100 individual instrument note samples selected at random from the University of Iowa instrument database [1]. Instruments contained in the selected portion of the database were alto saxophone, bassoon, b-flat clarinet, e-flat clarinet, flute, French horn, oboe, soprano saxophone, trombone, and trumpet. Because the onset times of different harmonics from the same instrument note are likely to be similar, we first remove the attack portion of each note before performing the correlation analysis. This removes the possibility of an upward bias in correlation values for cases where harmonics start at the same time, but otherwise do not exhibit similar modulation trends.

To remove the attack and isolate the sustained portion of each note we employ a simple method of onset detection on the time-domain waveform by searching for the maximum value in the derivative of the signal's envelope, where the envelope is calculated by squaring and low-pass filtering the signal. We measure correlation of harmonics at the time frame level because as Section III describes, the CAM assumption is utilized at the STFT frame level in our proposed system. We consider the sustained portion of the signal to be all time frames after the frame that contains the onset to the final frame of the signal. After transforming each individual instrument signal to the STFT domain (where all parameters are set as described in Section IV-A) we associate frequency bins with each harmonic according to (9) and calculate each harmonic's amplitude values using (16).

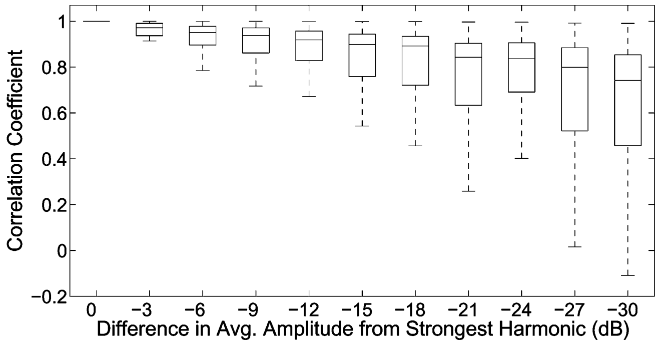


Fig. 1. Box plots of correlation coefficients [see (6)] measured between the strongest harmonic and other harmonics of individual instrument notes for the sustained portions of each note, plotted as a function of amplitude difference between harmonics. Results are calculated using 100 note samples. The upper and lower edges of each box represent the upper and lower quartile ranges, the middle line shows the median value and the whiskers extend to the most extreme values within 1.5 times the interquartile range.

As will be shown in Section III-E, we utilize the CAM principle to estimate the ratio between amplitude values in different time frames of an overlapped harmonic from the ratio between amplitude values of a non-overlapped harmonic. Accordingly, let us introduce the notation $r_{m^*\to m}^{h_i} = a_i^{h_i}(m)/a_i^{h_i}(m^*)$. Thus, $r_{m^*\to m}^{h_i}$ is the amplitude change (in terms of a ratio) of harmonic h_i from frame m^* to m. Given this definition, we calculate the correlation coefficient between the strongest harmonic, denoted by h_i^*, and another harmonic g_i over a note segment with time frames from m_0 to m_1 as

$$C\left(h_i^*, g_i\right) = \frac{\sum\limits_{l=m_0}^{m_1} r_{m^*\to l}^{h_i^*} r_{m^*\to l}^{g_i}}{\sqrt{\left(\sum\limits_{l=m_0}^{m_1}\left(r_{m^*\to l}^{h_i^*}\right)^2\right)\left(\sum\limits_{l=m_0}^{m_1}\left(r_{m^*\to l}^{g_i}\right)^2\right)}} \qquad (6)$$

where $m^* = \arg\max_{m\in[m_0,\ldots,m_1]}\left(a_i^{h_i^*}(m)\right)$. We select m^* using this method to avoid scaling distortions when $a_i^{h_i^*}(m) \approx 0$.

Fig. 1 shows box plots of the correlation coefficients between harmonics for the sustained portions of the notes. The plots are shown as a function of the difference in average amplitude between harmonics (rounded to 3-dB increments), where amplitude values are averaged over all time frames being used in the correlation measure. The upper and lower edges of each box represent the upper and lower quartile ranges, the middle line shows the median value and the whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range.

We can see that the correlation is very high for harmonics with energy close to that of the strongest harmonic and tapers off as the energy in a harmonic decreases. At roughly 30 dB below the

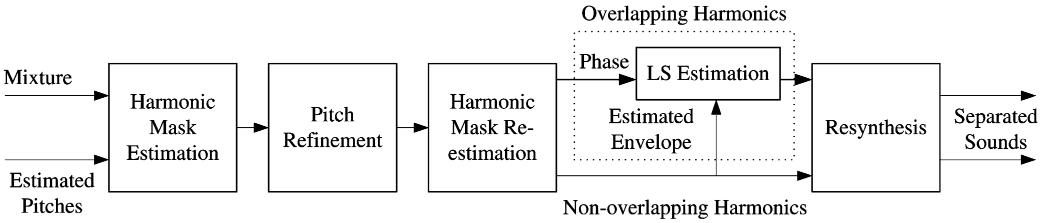


Fig. 2. System diagram. LS stands for least squares.

strongest harmonic, the median correlation value of 0.74 is still relatively high. The data shown in Fig. 1 suggests that strong harmonics are highly correlated with one another, and thus the amplitude envelope of relatively strong harmonics can be approximated by that of another strong, non-overlapped harmonic of the same source. Approximation of a low-energy harmonic using a non-overlapped harmonic of the same source may be less accurate during sustained portions of a note, but the perceptual degradation of the signal should be less severe than poor approximation of a strong harmonic. Correlation scores were slightly higher when the attack of each note was not removed (a 95% confidence interval for improvement in correlation coefficient is $[0.06, 0.1]$ using a two-sided, paired t test), suggesting that correlation during the onset portion of harmonics is also quite high.

C. Phase Change Estimation Using Pitch

As discussed in the introduction, both the amplitudes and phases of the harmonics must be considered for good estimation. The CAM assumption allows us to estimate how the amplitudes of overlapping harmonics change over time. With an estimate of the harmonics' change in phase over time, we show in Section III-E how the observed mixture can be used to resolve the overlapping harmonics in an efficient least squares framework. A harmonic's change in phase is related to the instantaneous frequency of a sinusoid as follows:

$$\phi_i^{h_i}(m+1) - \phi_i^{h_i}(m) = 2\pi f_i^{h_i}(m) T_m \tag{7}$$

or equivalently

$$\begin{aligned}\Delta\phi_i^{h_i}(m) &= 2\pi f_i^{h_i}(m) T_m \\ &= 2\pi h_i F_i(m) T_m.\end{aligned} \tag{8}$$

Here, T_m denotes the hop size of the STFT in seconds and $\Delta\phi_i^{h_i}(m) = \phi_i^{h_i}(m+1) - \phi_i^{h_i}(m)$. The relationship gives us the progression of a harmonic's phase from the signal's pitch contour, provided the signal adheres to the harmonic sinusoidal model, the frequency is stable over the duration of the time frame and the pitch estimate is accurate. Similar to the CAM assumption, the signals will not adhere to these constraints absolutely, but performance suggests that the assumptions are reliable enough to provide good separation.

III. System Description

A. System Overview

Our proposed separation system is illustrated in Fig. 2. The input to the system is a polyphonic, single-channel mixture and rough pitch estimates of each source. The pitch contour information can be in the form of a time-aligned MIDI score or from a multipitch detection algorithm. In the first stage, after the input is decomposed using the short-time Fourier transform (STFT), the pitch estimates are used to derive a harmonic mask for each source and identify T–F regions containing non-overlapped or overlapped harmonics. In the pitch refinement stage, we utilize the phase information in the T–F regions of non-overlapped harmonics to obtain more accurate pitch estimates. Using the refined pitch estimates, the system derives a new harmonic mask for each source and reidentifies T–F regions containing non-overlapped or overlapped harmonics. For T–F regions containing non-overlapped harmonics, the values in the mixture STFT are retained and passed to the resynthesis stage via a binary mask. The system also estimates the amplitude envelopes of the non-overlapped harmonics from these regions. For T–F regions containing overlapped harmonics, the system uses the refined pitch data to estimate the instantaneous frequency of overlapped harmonics, which yields the time dynamics of the phase parameters. The amplitude and phase contours (dynamics over time) are then used in a least-squares framework to estimate the amplitude and phase values of the overlapped harmonics. The resulting amplitude and phase parameters are used to estimate the STFT values for each source in the overlapped T–F regions and these values are passed to the resynthesis stage and added to the STFT values distributed from the binary masks. Finally, the overlap-add method is used to convert the estimated STFT of each signal to a time-domain estimate of each instrument.

B. Harmonic Mask Estimation

As mentioned in Section III-A, the first processing stage takes as input a polyphonic mixture and pitch estimates for each source. This stage first transforms the input using the STFT and uses pitch estimates to generate a harmonic mask for each source by identifying the frequency bins associated with each harmonic at each time frame. A frequency bin k at time frame m is associated with harmonic h_i if

$$\left| k f_b - f_i^{h_i}(m) \right| < \theta_1 \tag{9}$$

where θ_1 is a threshold. We denote the set of frequency bins associated with h_i as $\mathbf{K}_i^{h_i}(m)$. We can define overlapped and non-overlapped harmonics similarly. Harmonic h_i is overlapped by some other harmonic h_p of source p at time frame m if

$$\left| f_i^{h_i}(m) - f_p^{h_p}(m) \right| < \theta_2 \tag{10}$$

where θ_2 is also a threshold. If no other harmonic has a frequency within θ_2 of harmonic h_i, we call h_i non-overlapped and denote the set of non-overlapped harmonics for source i in frame m as $\tilde{\mathbf{H}}_i(m)$. A harmonic mask is simply a collection of T–F units associated with non-overlapped harmonics. In Section III-G, we describe how this set of time-frequency units can be used to directly distribute energy from the mixture to a signal estimate as a binary mask would, hence the decision to describe this set of T–F units as a "mask."

We set both θ_1 and θ_2 using the magnitude spectrum of the windowing function W. We associate frequency bins with a harmonic rather liberally and set θ_1 as half of the bandwidth at which W has dropped by 40 dB. We require more source interaction to label a harmonic as overlapped and set θ_2 as the more traditional bandwidth at which W has dropped by 6 dB. As discussed in Section II-A, the choice of θ_1 restricts the range of instrument pitches for which the approximation in (4) is valid. In our implementation, with parameters listed in Section IV-A, the instrument pitches are restricted to be above about 40 Hz. Of course, θ_1 and θ_2 can be tuned as necessary.

It should be noted that setting $\theta_1 > \theta_2/2$ means that a frequency bin could be assigned to multiple harmonics. We set these parameters as described with a simple idea in mind. When a harmonic is not close to any other harmonic, we would like to distribute as much of the mixture energy as possible. As a result, we need to take some care in how we assign frequency bins to the different harmonics. Thus, we define the set of frequency bins associated with harmonic h_i as

$$\mathbf{K}_i^{h_i}(m) = \left\{ k \mid \left|kf_b - f_i^{h_i}(m)\right| < \left|kf_b - f_p^{h_p}(m)\right| < \theta_1, \right. \\ \left. \forall\, p \neq i,\ \forall h_p \right\}. \quad (11)$$

The condition that $\left|kf_b - f_i^{h_i}(m)\right| < \left|kf_b - f_p^{h_p}(m)\right|$ is only necessary when $\theta_1 > \theta_2/2$.

C. Pitch Refinement

With the harmonic mask of each source generated using the initial pitches, we use frequency bins associated with non-overlapped harmonics to refine the pitch estimates. For each source, we first estimate the instantaneous frequency of each non-overlapped harmonic using the phase information from the mixture. Consider a non-overlapped harmonic h_i at frame m and its initial frequency estimate $f_i^{h_i}(m)$ given by $h_iF_i(m)$, where $F_i(m)$ is the initial pitch estimate. For a bin $k \in \mathbf{K}_i^{h_i}(m)$, denote the observed mixture phases at frame m and $m+1$ as $\varphi(m,k)$ and $\varphi(m+1,k)$, respectively. The observed phases have the same relationship to instantaneous frequency as the true phases of the underlying harmonic, as shown in (7), except that the observed phases are constrained between $-\pi$ and π. As a result, the term $2\pi z$, where z is an integer, is included in the instantaneous frequency estimate as

$$\tilde{f}_i^{h_i}(m) = \frac{1}{2\pi T_m}(\varphi(m+1,k) - \varphi(m,k) - 2\pi z). \quad (12)$$

In [23], it is shown that the integer z that correctly unwraps the phases can be calculated as

$$z = \left[\frac{1}{2\pi}(\varphi(m,k) - \varphi(m+1,k)) + kf_bT_m\right] \quad (13)$$

where $[\cdot]$ rounds the value inside the brackets to the nearest integer.

We select the strongest frequency bin associated with each harmonic to calculate $\tilde{f}_i^{h_i}(m)$. Formally, for each time frame m, and for each $h_i \in \tilde{\mathbf{H}}_i(m)$, we select

$$k^* = \max_{k \in \mathbf{K}_i^{h_i}(m)} (|Z(m,k)|) \quad (14)$$

and replace k with k^* in (12) to estimate the instantaneous frequency. Finally, we calculate the refined pitch estimate as the weighted average of the instantaneous frequencies of the harmonics divided by their harmonic number

$$\tilde{F}_i(m) = \frac{\sum\limits_{h_i \in \tilde{\mathbf{H}}_i(m)} \frac{\tilde{f}_i^{h_i}(m)}{h_i} E_i^{h_i}(m)}{\sum\limits_{h_i \in \tilde{\mathbf{H}}_i(m)} E_i^{h_i}(m)} \quad (15)$$

where $E_i^{h_i}(m) = |Z(m,k^*)|$.

D. Harmonic Mask Re-Estimation

Using the refined pitch estimate $\tilde{F}_i(m)$, we derive a new harmonic mask for each source by finding a new set of frequency bins $\mathbf{K}_i^{h_i}(m)$ associated with each harmonic using (11) and reidentify T–F regions for all the harmonics. We then estimate the amplitudes for all the non-overlapped harmonics by finding the amplitude $a_i^{h_i}(m)$ that minimizes the following:

$$\sum_{k \in \mathbf{K}_i^{h_i}(m)} \left(|Z(m,k)| - \frac{a_i^{h_i}(m)}{2} |W(kf_b - h_i\tilde{F}_i(m))| \right)^2. \quad (16)$$

The minimization of the above equation is

$$a_i^{h_i}(m) = \frac{2 \sum\limits_{k \in \mathbf{K}_i^{h_i}(m)} |Z(m,k)| \cdot |W(kf_b - h_i\tilde{F}_i(m))|}{\sum\limits_{k \in \mathbf{K}_i^{h_i}(m)} |W(kf_b - h_i\tilde{F}_i(m))|^2}. \quad (17)$$

In (17), and then in (23) and (27), we use the DTFT of $w[n]$ shown in (3) to calculate the value of $W(kf_b - h_i\tilde{F}_i(m))$.

E. Least-Squares Estimation

In many applications, the hop size of the STFT is in the tens of milliseconds (23 ms in our implementation), which tends to be shorter than the length of individual notes in many music recordings. As a result, overlap between harmonics often occurs in sequences of time frames as well as a series of frequency bins. Accordingly, we extend the idea of an overlapped harmonic to an overlapped T–F region. Let $\{h_{i_1}, \ldots, h_{i_P}\}$ be a set of P harmonics that overlap during time frames from m_0 to m_1. The overlapped T–F region for this set of harmonics is defined as

$$\begin{aligned}&\mathbf{D}(m_0, m_1; k_0, k_1)\\ &\quad = \{m, k \mid m \in \{m_0, \ldots, m_1\}; k \in \{k_0, \ldots, k_1\}\}. \qquad (18)\end{aligned}$$

where k_0 is the smallest $k \in \bigcup_{i=i_1}^{i_P} \bigcup_{l=m_0}^{m_1} \mathbf{K}_i^{h_i}(l)$ and k_1 is the largest. In other words, the overlapped region is the bounding box that includes the frequency bins associated with all of the overlapping harmonics. As a simple example, assume that h_{i_1} and h_{i_2} overlap during time frames 10 through 18, and that $\mathbf{K}_{i_1}^{h_{i_1}} = \{21, 22, 23, 24\}$ and $\mathbf{K}_{i_2}^{h_{i_2}} = \{23, 24, 25, 26\}$. Then the overlapped T–F region is $\mathbf{D}(10, 18; 21, 26)$.

According to the model developed in (4) and (5), the observed STFT value $Z(m, k)$ can be written as

$$\begin{aligned}Z(m,k) &= \sum_i \underbrace{\frac{a_i^{h_i}(m_0)}{2} e^{j\phi_i^{h_i}(m_0)}}_{1}\\ &\quad \cdot \underbrace{r_{m_0 \to m}^{h_i} e^{j\sum_{l=m_0}^{m} \Delta\phi_i^{h_i}(l)}}_{2}\\ &\quad \cdot \underbrace{W(kf_b - h_i\tilde{F}_i(m))}_{3} \qquad (19)\\ &= \sum_i S_i^{h_i}(m_0) R_i(m,k). \qquad (20)\end{aligned}$$

The first term in (19) represents the amplitude and the phase of harmonic h_i of source i at the starting frame m_0 of the region. The second term models the amplitude and the phase change of the same harmonic from frame m_0 to frame m. As in Section II-B, $r_{m_0 \to m}^{h_i}$ denotes the amplitude ratio between the two frames. Since $r_{m_0 \to m}^{h_i}$ is unknown, we use the CAM principle to approximate $r_{m_0 \to m}^{h_i}$ by $r_{m_0 \to m}^{h_i^*}$, where h_i^* denotes a non-overlapped harmonic with strong energy from source i. A discussion of how to select h_i^* is provided in Section III-F. The last term in (19) accounts for the effect of the analysis window. The summation is over all the sources that have a harmonic contributing energy to the region $\mathbf{D}(m_0, m_1; k_0, k_1)$. Note that with the approximation of $r_{m_0 \to m}^{h_i}$ by $r_{m_0 \to m}^{h_i^*}$, and the relationship between instantaneous frequency and change in phase shown in (8), only the first term of (19) is unknown. Therefore, we can express (19) more concisely as in (20), where $S_i^{h_i}(m_0)$ is term 1 and $R_n(m, k)$ is the multiplication of term 2 and 3.

We can write (20) for all the T–F units within the region $\mathbf{D}(m_0, m_1; k_0, k_1)$. We define matrix $\mathbf{R}$ and vectors $\mathbf{S}$, $\mathbf{Z}$ as

$$\mathbf{R} = \begin{pmatrix} R_{i_1}(m_0,k_0) & \ldots & R_{i_P}(m_0,k_0) \\ \vdots & & \vdots \\ R_{i_1}(m_0,k_1) & \ldots & R_{i_P}(m_0,k_1) \\ \vdots & & \vdots \\ R_{i_1}(m,k) & \ldots & R_{i_P}(m,k) \\ \vdots & & \vdots \\ R_{i_1}(m_1,k_0) & \ldots & R_{i_P}(m_1,k_0) \\ \vdots & & \vdots \\ R_{i_1}(m_1,k_1) & \ldots & R_{i_P}(m_1,k_1) \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} S_{i_1}^{h_{i_1}}(m_0) \\ \vdots \\ S_{i_P}^{h_{i_P}}(m_0) \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} Z(m_0,k_0) \\ \vdots \\ Z(m_0,k_1) \\ \vdots \\ Z(m,k) \\ \vdots \\ Z(m_1,k_0) \\ \vdots \\ Z(m_1,k_1) \end{pmatrix} \qquad (21)$$

where

$$\mathbf{RS} = \mathbf{Z}. \qquad (22)$$

Thus, we have a set of matrix equations with the initial states (amplitude and phase) of each harmonic as the only unknowns. The coefficient matrix $\mathbf{R}$ can be constructed using the estimated amplitude envelope and phase change information, $\mathbf{S}$ is a vector of unknowns and $\mathbf{Z}$ is a vector with the observed STFT values. The least-squares estimation of $\mathbf{S}$ is given by

$$\mathbf{S} = (\mathbf{R}^H\mathbf{R})^{-1}\mathbf{R}^H\mathbf{Z} \qquad (23)$$

where H denotes conjugate transpose. After $\mathbf{S}$ is estimated, the complex sinusoidal parameter of each harmonic contributing to the overlapped region can be estimated as

$$S_i^{h_i}(m) = S_i^{h_i}(m_0) r_{m_0 \to m}^{h_i^*} e^{j\sum_{l=m_0}^{m} \Delta\phi_i^{h_i}(l)} \qquad (24)$$

Fig. 3 shows the effectiveness of the least-squares estimation in recovering two overlapping harmonics. In this case, the third harmonic of one source overlaps with the fourth harmonic of a second source. Fig. 3(c) shows the magnitude spectrum of the mixture in the overlapped region. Note that the amplitude modulation results from the relative phase of the two harmonics. The estimated magnitude spectra of two sources are shown in Fig. 3(d) and (e). For comparison, the magnitude spectra of the two sources obtained from premixed signals are shown in Fig. 3(a) and (b).

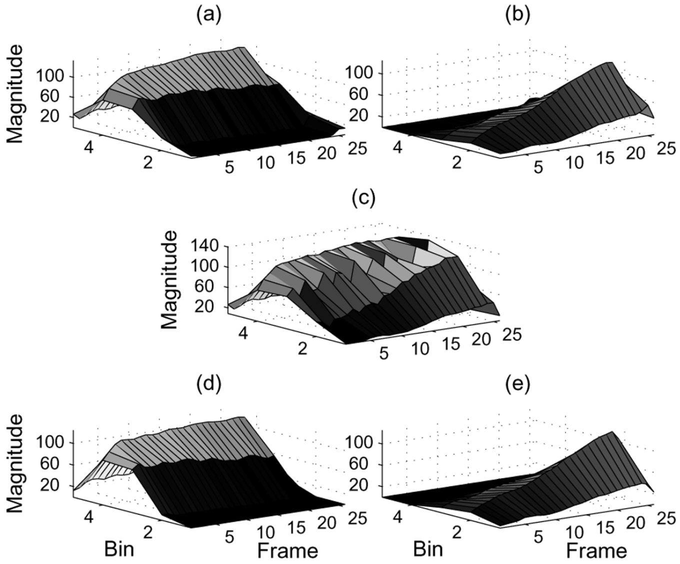


Fig. 3. Least-squares estimation of overlapping harmonics. (a) The magnitude spectrum of a harmonic of the first source in the overlapped T–F region. (b) The magnitude spectrum of a harmonic of the second source in the same T–F region. (c) The magnitude spectrum of the mixture. (d) The estimated magnitude spectrum of the harmonic from the first source. (e) The estimated magnitude spectrum of the harmonic from the second source.

F. Selection of a Non-Overlapping Harmonic

The previous section described how the amplitude envelopes of non-overlapped harmonics are used in the least squares framework to estimate the sinusoidal parameters of overlapped harmonics from the same source. Numerous approaches could be employed within the proposed framework to approximate $r^{h_i}_{m_0\to m}$. In this paper, we take a simple approach. For each harmonic h_i that is overlapped in a T–F region, we select h_i^* as the strongest harmonic of source i that is non-overlapped for the entire sequence of frames m_0 to m_1. Formally, we select h_i^* as follows:

$$h_i^* = \max_{g\in\cap_{l=m_0}^{m_1}\tilde{\mathbf{H}}_i(l)} \left\{ \sum_{l=m_0}^{m_1} a_i^g(l) \right\}. \tag{25}$$

Some alternatives that were explored include selection of a harmonic that is both non-overlapped during the entire sequence of frames and closest in frequency to h_i, taking an average of nearby harmonics or an average of all non-overlapped harmonics. We found average separation performance to be best with selection of h_i^* according to (25).

An attractive aspect of the proposed estimation approach is that one could easily substitute a note model for a particular instrument or devise an alternative scheme for approximating $r^{h_i}_{m_0\to m}$. The approximation chosen in the proposed system is attractive because no prior knowledge is needed. A shortcoming is that if there are no non-overlapped harmonics available for the duration m_0 to m_1, then the overlapped harmonic h_i cannot be estimated using the proposed approach. In the current study, we simply ignore reconstruction in this case.

G. Resynthesis

In the final estimation of the STFT of each source signal, we combine estimates from the non-overlapped and overlapped regions. First, let $\mathbf{K}_i^{no}(m) = \bigcup_{h_i\in\tilde{\mathbf{H}}_i(m)} \mathbf{K}_i^{h_i}(m)$ be the set of frequency bins associated with non-overlapped harmonics in time frame m, and let $\mathbf{K}_i^{o}(m) = \bigcup_{h_i\notin\tilde{\mathbf{H}}_i(m)} \mathbf{K}_i^{h_i}(m)$ be the set of frequency bins associated with overlapped harmonics in time frame m. For the bins associated with non-overlapped harmonics we directly distribute the mixture STFT to the source estimate

$$\hat{X}_i^{no}(m,k) = Z(m,k) \quad \forall\, k \in \mathbf{K}_i^{no}(m). \tag{26}$$

For the bins associated with overlapped harmonics, we utilize the sinusoidal model and calculate the STFT using

$$\hat{X}_i^{o}(m,k) = S_i^{h_i}(m) W(kf_b - h_i\tilde{F}_i(m)) \quad \forall\, k \in \mathbf{K}_i^{o}(m). \tag{27}$$

Finally, the overall source STFT is $\hat{X}_i = \hat{X}_i^{no} + \hat{X}_i^{o}$ and we use the overlap-add method to obtain the time domain estimate $\hat{x}_i[n]$ for each source.

IV. Evaluation and Comparison

A. Database and Parameter Settings

We test the proposed system on a database of 20 Bach quartets. Audio signals are generated from four-part MIDI files (soprano, alto, tenor, bass) by first selecting I instrument parts (e.g., alto and tenor). Each part is randomly assigned one of four instruments: clarinet, flute, violin, or trumpet. For each note event in the I parts we select an audio sample of the specified instrument that matches the specified pitch and place it at the specified onset time. In the case where the audio sample is longer than the MIDI event (as it typically is), we truncate the audio sample to match the length of the note event. When the MIDI note event is longer than the available audio sample, we use the audio sample as is and alter the length of the note event in the MIDI data. Instrument samples are drawn from the RWC music instrument database [14]. Mixtures are formed with either two instruments, where we select the alto and tenor musical parts, or three instruments, where we select the soprano, alto, and tenor parts. In all cases, instruments are mixed with equal average power over the duration of the signals. More details about the procedure can be found in [21].

Although the mixtures formed are at best an approximation of a real performance, they exhibit realistic pitch variations due to both vibrato and instances where the audio samples are slightly sharp or flat. This characteristic better tests our pitch refinement stage and is perhaps more true to real recordings than the use of a synthesizer or sampler to generate signals from the MIDI data, which have their own drawbacks.

In our implementation, we use a frame length of 4096 samples with sampling frequency 44.1 kHz. No zero-padding is used in the DFT. The hop size is 1024 samples. We choose $\theta_1 = 2.5f_b$, about half of the 40-dB bandwidth of the Hamming window, and $\theta_2 = 1.5f_b$, which is approximately the 6-dB bandwidth of the Hamming window [15]. The number of harmonics for each source H_i is chosen such that $f_i^{H_i}(m) < f_s/2$ for all time frames, where f_s denotes the sampling frequency.

B. Pitch Refinement

We first evaluate the effectiveness of the pitch refinement stage. Evaluation is performed on 5-s excerpts from the 20 two-instrument mixtures described above. We consider three different categories of rough pitch estimates. First is when a time-aligned MIDI score is available. The second and third cases are when pitch contours are detected from the multi-pitch detection algorithm by Klapuri [19]. Since Klapuri's system does not sequentially group detected pitch values from the same source, the second case assumes ideal sequential grouping of the detected pitches (i.e., each detected pitch is matched to the source with the closest ground truth pitch). In the third case, we group the detected pitch values sequentially using a heuristic grouping rule that states that pitch contours of different instrument lines should not cross each other. This rule has theoretical foundation in music composition and perception [16]. Studies have shown that this simple sequential grouping rule works very well for Bach's work [10], [17] and is likely a reasonable choice for a large body of musical works. In our evaluation, we simply arrange the pitch values at any given frame from high to low and then group the detected pitch values for separation. The results for this final case indicate the performance our system can achieve when applied to real recordings.

In the analysis, we consider two types of pitch errors, *gross* pitch errors and *fine* pitch errors. We define a gross pitch error as a frequency deviation from ground truth pitch greater than half of a semitone. We obtain ground truth pitch contours from the clean source signals prior to mixing using a program based on Praat [4]. In Table I, we show the percentage of time frames that contain gross pitch errors for the three cases, where "MIDI" denotes the MIDI pitch contour case, "DPI" denotes the detected pitch with ideal sequential grouping case, and "DP" denotes the detected pitch with heuristic grouping rule case. The first row of the table shows the gross error percentages of the rough pitch estimates prior to pitch refinement, while the second shows the error percentages after pitch refinement. We include this table to show that the pitch refinement stage does not significantly increase the number of gross pitch errors. While the refinement stage is not able to reduce gross errors since the initial harmonic masks must align with at least some of the instrument's harmonics in order to be successful, it is possible that the refinement could introduce gross pitch errors. Only when MIDI pitch contours are used, which have 0 gross pitch errors to begin with, are a handful of gross pitch errors introduced. On closer analysis, these cases primarily occur on note transitions when the assumption of stable instantaneous frequency over a time frame is violated.

TABLE I
PERCENTAGE OF GROSS PITCH ERRORS. THE COLUMNS DENOTE THE CASES WHEN MIDI PITCH CONTOURS ARE AVAILABLE (MIDI), DETECTED PITCHES ARE IDEALLY GROUPED (DPI), AND DETECTED PITCHES ARE HEURISTICALLY GROUPED (DP)

	MIDI	**DPI**	**DP**
Gross Error %	0	3.8	6.9
Gross Error (with refinement) %	0.3	3.8	6.9

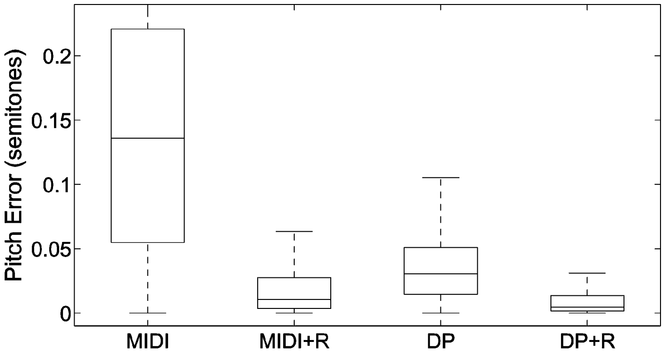


Fig. 4. Box plots of fine pitch error in semitones relative to ground truth pitch. Results are shown for pitch contours provided by MIDI (MIDI), MIDI with pitch refinement (MIDI + R), detected pitches grouped heuristically (DP) and detected pitches with refinement (DP + R).

An analysis of fine pitch errors is provided in Fig. 4. To generate these box plots, where mid-line, box edges and whiskers are defined as in Fig. 1, we first ignore all gross pitch errors and measure the fine pitch error in terms of musical semitones. Since the primary difference between the ideally grouped detected pitches and the heuristically grouped detected pitches has to do with gross errors, we only show results for the MIDI pitch contours and the heuristically grouped detected pitch contours. In order to provide more detail on refined cases, the plot does not show the full extent of the MIDI pitch errors without refinement, but the whisker extends to 0.47 semitones in that case. In this figure, we again denote the MIDI contours with "MIDI" and the heuristically grouped detected pitches with "DP." The notation "$+R$" indicates the cases in which pitch refinement has been used. As can be seen, the refined pitch estimates are substantially more accurate than the rough pitch estimates. Median error (in semitones) drops from 0.14 to 0.01 for the MIDI case, and from 0.03 to 0.005 in the detected pitch case. In both cases, we can see that the spread of the accuracy is also much smaller when refinement is included.

C. Sound Separation

In this section, we provide the sound separation performance of the proposed system and compare it to existing sinusoidal-model based monaural separation systems. We first define the signal-to-noise ratio (SNR)

$$\mathrm{SNR}_{\mathrm{est}} = 10\log_{10}\frac{\sum_n x^2[n]}{\sum_n (x[n]-\hat{x}[n])^2} \tag{28}$$

$$\mathrm{SNR}_{\mathrm{mix}} = 10\log_{10}\frac{\sum_n x^2[n]}{\sum_n (x[n]-z[n])^2}. \tag{29}$$

Here, $x[n]$ is the clean source signal prior to mixing, $\hat{x}[n]$ is the estimated signal and $z[n]$ is the mixture signal. The SNR gain is then $\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{est}} - \mathrm{SNR}_{\mathrm{mix}}$. Evaluation is performed on the two-source condition where, as in Section IV-B, the tenor

TABLE II
AVERAGE ΔSNR OF THE PROPOSED SYSTEM AND EXISTING SYSTEMS ON 40 FIVE-SECOND INSTRUMENT SIGNALS FROM TWO-INSTRUMENT MIXTURES. RESULTS ARE SHOWN FOR THE PROPOSED SYSTEM WITH ONLY NON-OVERLAPPING HARMONICS INCLUDED (NON-OVER) AND THE FULL SYSTEM (PROPOSED), AND WITH $(+R)$ AND WITHOUT $(-R)$ PITCH REFINEMENT, AND FOR EXISTING SYSTEMS

	GTP	**MIDI**	**DPI**	**DP**
Non-Over - R	10.2	8.6	9.6	8.6
Non-Over + R	10.1	9.9	9.8	8.8
Proposed - R	14.5	8.3	9.0	7.9
Proposed + R	14.7	13.3	13.7	12.1
Virtanen (2006)	11.0	9.8	N/A	8.5
Parsons (1976)	10.7	5.8	9.8	8.8

TABLE III
AVERAGE SDR, SIR, AND SAR OF THE PROPOSED SYSTEM AND EXISTING SYSTEMS ON 40 FIVE-SECOND INSTRUMENT SIGNALS FROM TWO-INSTRUMENT MIXTURES. RESULTS ARE SHOWN FOR THE PROPOSED SYSTEM USING GROUND TRUTH PITCH (GTP), MIDI PITCH (MIDI), IDEALLY GROUPED DETECTED PITCH (DPI) AND HEURISTICALLY GROUPED DETECTED PITCH (DP)

	SDR	**SIR**	**SAR**
Proposed (GTP)	14.5	44.3	14.5
Proposed (MIDI)	13.0	42.7	13.0
Proposed (DPI)	13.4	42.3	13.5
Proposed (DP)	11.7	36.6	11.8
Virtanen (GTP)	10.7	30.8	10.8
Parsons (GTP)	10.2	25.7	10.4

and alto musical parts are selected from the 20 Bach quartets and signals are mixed with equal energy (i.e., $\mathrm{SNR}_{\mathrm{mix}} = 0$).

Overall separation performance is a factor of estimation of both non-overlapped harmonics and overlapped harmonics. Accordingly, we show results for cases in which the signal estimates only include non-overlapped harmonics $\left(\hat{X}_i = \hat{X}_i^{no}\right)$ and results using the entire reconstructed signal $\left(\hat{X}_i = \hat{X}_i^{no} + \hat{X}_i^{o}\right)$, both overlapped and non-overlapped harmonics. Table II shows the average ΔSNR of the 40 five-second instrument signals (two for each mixture in the database of 20) for the proposed system and a recent musical separation system [31], denoted by "Virtanen (2006)," as well as a classic separation system "Parsons (1976)" [26]. Signal estimates for the system in [31] were generated by the author. We provided the mixture database and pitch contours to him and he returned the separated instrument signals. The results for the system in [26] were generated using our own implementation. Results for the proposed system are shown with only non-overlapping harmonics included, "Non-Over," and the full system, "Proposed," and with "$+R$" and without "$-R$" pitch refinement. The separate columns are for the four different cases of pitch contours used: ground truth pitch "GTP," "MIDI," 'DPI," and "DP."

Results show the effectiveness of both novel aspects of the proposed system, the pitch refinement stage and reconstruction of overlapping harmonics. As one would expect from the results presented in Section IV-B, the pitch refinement greatly improves the average ΔSNR in all cases of rough pitch estimates. Comparing the third and fourth rows of Table II, the average improvement achieved by pitch refinement of rough estimates is over 4.6 dB. The results also show that, provided the pitch refinement has been enabled, reconstruction of the overlapping harmonics improves the estimation of the signal. Comparing the second and fourth rows of Table II, the average improvement through inclusion of the overlapping harmonics is 3.8 dB over all four pitch cases. We can also see that the estimation of overlapping harmonics is more strongly effected by pitch inaccuracies than the estimation of non-overlapping harmonics.

The performance of the full system on the ground truth pitch condition is 14.7 dB, while it degrades to 12.1 dB when we use detected pitches and simply group the higher ones with the alto line and the lower ones with the tenor line. When the detected pitches are ideally grouped with the correct instrument, the performance is 13.7 dB, indicating that the majority of the degradation is due to cases that violate the heuristic grouping rule. We found that 3 of the 20 mixtures violated this rule, and signals from those mixtures had an average ΔSNR of 5.1 dB.

In comparison to the existing separation systems, we can see that the proposed system provides an improvement in all cases. The improvement in ΔSNR over the most competitive system is 3.7 dB for the ground truth pitch case and 3.3 dB for the detected pitch and heuristic grouping case. It should be noted that results for the Virtanen system presented in [31] are given for single-note polyphonies (mixtures of multiple instruments each simultaneously playing only one note) rather than for note sequences, as are used in the test database presented here. It is possible that tuning of the Virtanen system on sequences of notes from multiple instruments gives better performance than the results presented here.

Alternative measures to SNR have been proposed [30] for evaluation of sound separation algorithms. The source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) measure overall distortion, energy from interfering sources, and artifacts introduced by the separation algorithm, respectively. Results from a preliminary study indicate that these measures may correlate more closely with human perception of signal similarity than SNR-based measures [12]. The results using these metrics are given in Table III. We show performance for the full proposed system (including overlapping harmonics and pitch refinement) on the four different pitch cases, and again for comparison, results using the existing systems with ground truth pitch, "GTP," Again, we see that the proposed system provides a significant improvement over the other systems.

We also test our system on ten 15-s mixtures of three instruments. Again, each instrument is mixed so that all three sources have equal average power over the duration of the signal. Using the ground truth pitch information, the average ΔSNR achieved was 14.2 dB, where $\mathrm{SNR}_{\mathrm{est}}$ was 11.2 dB and $\mathrm{SNR}_{\mathrm{mix}}$ was -3 dB. Additionally, we test the proposed system on reverberant recordings by generating impulse responses with 0.54-s reverberation time (T_{60}) using the Roomsim package [8]. The estimated signals for the two-instrument reverberant mixtures using the ground truth pitch contours yield an average ΔSNR of 13.6 dB, only 1.1 dB lower than in the anechoic case.

Sound demos of the proposed separation system can be found at http://www.cse.ohio-state.edu/~woodrufj/mmssTASLP.html.

V. Conclusion

In this paper, we have proposed a monaural musical sound separation system that explicitly resolves overlapping harmonics. Our approach is based on CAM and phase change estimation using pitch contours. Quantitative results show that when pitches can be estimated accurately, the separation performance is excellent. Even with rough pitch estimates, the proposed system still achieves good separation. In addition to a large increase in SNR, the perceptual quality of the separated signals is satisfactory in most cases. We have also shown that rough pitch estimates can be refined from a subset of a signal's harmonics, and that the proposed mechanism for refinement achieves decreased deviation from ground truth pitch and leads to improved signal separation.

Acknowledgment

The authors would like to thank the two anonymous reviewers for their constructive criticisms and suggestions. The authors would also like to thank T. Virtanen for his assistance in comparison.

References

[1] The University of IOWA Musical Instrument Sample Database. [Online]. Available: http//:theremin.music.uiowa.edu/

[2] S. A. Abdallah, "Towards Music Perception by Redundancy Reduction and Unsupervised Learning In Probablistic Models," Ph.D. dissertation, Dept. of Electron. Eng., King's College London, , London, U.K., 2002.

[3] M. Bay and J. W. Beauchamp, "Harmonic source separation using prestored spectra," *Indep. Compon. Anal. Blind Signal Separ.*, pp. 561–568, 2006.

[4] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Comput., Ver. 4.0.26," 2002. [Online]. Available: http://www.fon.hum.uva.nl/praat

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

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

[7] E. M. Burns, "Intervals, scales, and tuning," in *The Psychology of Music*, D. Deutsch, Ed. San Diego, CA: Academic, 1999.

[8] D. R. Campbell, "The ROOMSIM User Guide (v3.3)," 2004. [Online]. Available: http://media.paisley.ac.uk/ campbell/Roomsim/

[9] M. A. Casey and W. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Comput. Music Conf.*, 2000.

[10] E. Chew and X. Wu, "Separating voices in polyphonic music: A contig mapping approach," in *Computer Music Modeling and Retrieval*, ser. Lecture Notes in Computer Science. Berlin/Heideberg, Germany: Springer, 2005.

[11] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1845–1856, Sep. 2006.

[12] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proc. Int. Conf. Ind. Compon, Anal. Signal Separ.*, 2007.

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

[14] M. Goto, "Analysis of musical audio signals," in *Computational Auditory Scene Analysis*, D. L. Wang and G. J. Brown, Eds. New York: Wiley, 2006.

[15] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[16] D. Huron, "Tone and voice: A derivation of the rules of voice-leading from perceptual principles," *Music Perception*, vol. 19, no. 1, pp. 1–64, 2001.

[17] D. Huron, "The avoidance of part-crossing in polyphonic music: Perceptual evidence and musical practice," *Music Perception*, vol. 9, no. 1, pp. 93–104, 1991.

[18] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: improving sound source separation based on integrated harmonic and inharmonic models," in *Proc. Int. Conf. Music Inf. Retrieval*, 2008.

[19] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

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

[21] Y. Li and D. L. Wang, "Pitch detection in polyphonic music using instrument tone models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. II-481–II-484.

[22] Y. Li and D. L. Wang, "Musical sound separation using pitch-based labeling and binary time-frequency masking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 173–176.

[23] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[24] D. K. Mellinger, "Event Formation and Separation in Musical Sound," Ph.D. dissertation, Dept. of Comput. Sci., Stanford Univ., Stanford, CA, 1991.

[25] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. Int. Conf. Music Inf. Retrieval*, 2008.

[26] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, no. 4, pp. 911–918, 1976.

[27] J. Risset and D. Wessel, "Exploration of Timbre by Analysis and Synthesis," in *The Psychology of Music*, D. Deutsch, Ed. New York: Academic, 1982, pp. 26–58.

[28] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. Poli, Eds. Lisse, The Netherlands: Swets & Zeitlinger, 1997.

[29] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2003, pp. 177–180.

[30] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[31] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, Tampere Univ. of Technol., Tampere, Finland, 2006.

[32] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[33] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2001, pp. 83–86.

[34] H. Viste and G. Evangelista, "Separation of harmonic instruments with overlapping partials in multi-channel mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2003, pp. 25–28.

[35] D. L. Wang, "Feature-based speech segregation," in *Computational Auditory Scene Analysis*, D. L. Wang and G. J. Brown, Eds. New York: Wiley, 2006.

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

[37] J. Woodruff, Y. Li, and D. L. Wang, "Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation," in *Proc. Int. Conf. Music Inf. Retrieval*, 2008.

[38] J. Woodruff and B. Pardo, "Using pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007.

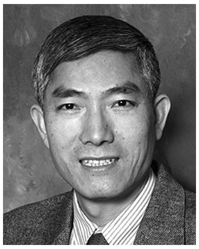
[39] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Conf. Music Inf. Retrieval*, 2006.

Yipeng Li received the B.S. degree in nuclear engineering from Tsinghua University, Beijing, China, in 2000 and the the M.S. degree in nuclear engineering and the Ph.D. degree in computer science and engineering from The Ohio State University, Columbus, in 2002 and 2008, respectively.

Since 2008, he has been with Microsoft.

John Woodruff (S'09) received the B.F.A. degree in performing arts and technology in 2002, the B.S. degree in mathematics from the University of Michigan, Ann Arbor, in 2004, and the M.Mus. degree in music technology from Northwestern University, Evanston, IL, in 2006. He is currently pursuing the Ph.D. degree in computer science and engineering at The Ohio State University, Columbus.

His research interests include computational auditory scene analysis, music and speech processing, auditory perception, and statistical learning. He is also an active recording engineer, electroacoustic composer, and songwriter.



DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1991, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. From October 2006 to June 2007, he was A Visiting Scholar at Oticon A/S, Denmark. His research interests include machine perception and neurodynamics.

Dr. Wang received the National Science Foundation Research Initiation Award in 1992, the Office of Naval Research Young Investigator Award in 1996, and the Helmholtz Award from the International Neural Network Society in 2008. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.