Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 86369, 10 pages
doi:10.1155/2007/86369

Research Article

Using Pitch, Amplitude Modulation, and Spatial Cues for Separation of Harmonic Instruments from Stereo Music Recordings

John Woodruff[1] and Bryan Pardo[2]

[1] *Music Technology Program, School of Music, Northwestern University, Evanston, IL 60208, USA*
[2] *Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA*

Received 2 December 2005; Revised 30 July 2006; Accepted 10 September 2006

Recommended by Masataka Goto

Recent work in *blind source separation* applied to anechoic mixtures of speech allows for improved reconstruction of sources that rarely overlap in a time-frequency representation. While the assumption that speech mixtures do not overlap significantly in time-frequency is reasonable, music mixtures rarely meet this constraint, requiring new approaches. We introduce a method that uses spatial cues from anechoic, stereo music recordings and assumptions regarding the structure of musical source signals to effectively separate mixtures of tonal music. We discuss existing techniques to create partial source signal estimates from regions of the mixture where source signals do not overlap significantly. We use these partial signals within a new demixing framework, in which we estimate *harmonic masks* for each source, allowing the determination of the number of active sources in important time-frequency frames of the mixture. We then propose a method for distributing energy from time-frequency frames of the mixture to multiple source signals. This allows dealing with mixtures that contain time-frequency frames in which multiple harmonic sources are active without requiring knowledge of source characteristics.

1. INTRODUCTION

Source separation is the process of determining individual source signals, given only mixtures of the source signals. When prior analysis of the individual sound sources is not possible, the problem is considered *blind source separation* (BSS). In this work, we focus on the BSS problem as it relates to recordings of music. A tool that can accomplish blind separation of musical mixtures would be of use to recording engineers, composers, multimedia producers, and researchers.

Accurate source separation would be of great utility in many music information retrieval tasks, such as music transcription, vocalist and instrument identification, and melodic comparison of polyphonic music. Source separation would also facilitate post production of preexisting recordings, sample-based musical composition, multichannel expansion of mono and stereo recordings, and structured audio coding.

The following section contains a discussion of related work in source separation, with an emphasis on current work in music source separation. In Section 3 we present a new source separation approach, designed to isolate multiple simultaneous instruments from an anechoic, stereo mixture of tonal music. The proposed method incorporates existing statistical BSS techniques and perceptually significant signal features utilized in computational auditory scene analysis to deal more effectively with the difficulties that arise in recordings of music. Section 4 provides a comparison of our algorithm to the DUET [1] source separation algorithm on anechoic, stereo mixtures of three and four harmonic instruments, and a discussion of the advantages and limitations of using our approach. Finally, in Section 5 we summarize our findings and discuss directions for future research.

2. CURRENT WORK

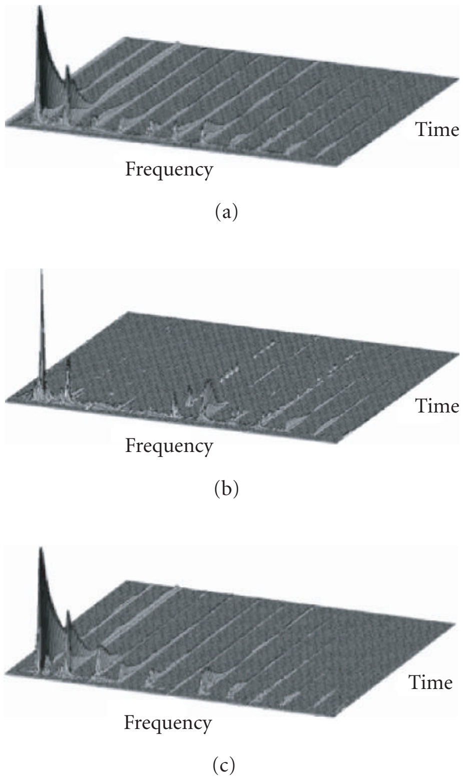
Approaches to source separation in audio are numerous, and vary based on factors such as the number of available mixture channels, the number of source signals, the mixing process used, or whether prior analysis of the sources is possible. *Independent component analysis* (ICA) is a well-established technique that can be used in the BSS problem when the number of mixtures equals or exceeds the number of source signals [2–5]. ICA assumes that source signals are statistically independent, and iteratively determines time-invariant demixing filters to achieve maximal independence between sources. When fewer mixtures than sources are available (i.e., stereo recordings of three or more instruments), the problem is considered the *degenerate* case of BSS and traditional ICA approaches cannot be used.

Researchers have proposed *sparse* statistical methods to deal more effectively with the degenerate case [1, 6–8]. Sparse methods assume that in a time-frequency representation, most time-frequency frames of individual source signals will have magnitude near zero. In speech, if sources are also independent (in terms of pitch and amplitude), the assumption that at most one source signal has significant energy in any given time-frequency frame is made [9]. Given this assumption, binary time-frequency masks can be constructed based on cross-channel amplitude and phase differences in an anechoic stereo recording and multiplied by the mixture to isolate source signals [1, 6]. The DUET algorithm, which we discuss in more detail in a later section, operates in this manner.

Tonal music makes extensive use of multiple simultaneous instruments, playing *consonant intervals*. When two harmonic sources form a consonant interval, their fundamental frequencies are related by a ratio that results in significant overlap between the *harmonics* (regions of high-energy at integer multiples of the fundamental frequency) of one source and those of another source. This creates a problem for DUET and other binary time-frequency masking methods that distribute each mixture frame to only one source signal. The resulting music signal reconstructions can have audible gaps and artifacts, as shown in Figure 1.

To deal with overlap of source signals in a time-frequency representation, researchers have incorporated heuristics commonly used in *computational auditory scene analysis* (CASA). CASA systems seek to organize audio mixtures based on known principles governing the organization of sound in human listeners [10, 11]. Perceptually significant signal features such as pitch, amplitude and frequency modulation, and common onset and offset are used in CASA systems to identify time-frequency regions of the mixture that result from the same sound source [12–14]. While the goal of many CASA researchers is to create a symbolic representation of a sound scene in terms of individual sources, CASA heuristics can be used within source separation algorithms to both identify mixture regions in which source signals overlap and to guide the reconstruction of source signals in overlap regions [2, 12, 14–19].

In the one-channel case, multiple researchers [14, 15, 17, 18] assume that source signals are harmonic in order to determine time-frequency regions of source signal overlap based on the pitch of the individual sources. Virtanen and Klapuri [17, 18] use multipitch estimation to determine instrument pitches. Time-frequency overlap regions are resolved by assuming that the magnitude of each source signal's harmonics decreases as a function of frequency. Signals are then reconstructed using additive synthesis. Published results based on this method have been shown only for cases when pitches were determined correctly, so it is difficult to assess the robustness of this approach. Reconstructing signals based solely on additive synthesis also ignores *residual*, or nonharmonic energy in pitched instrument signals [20].



(a)

(b)

(c)

Figure 1: (a) The spectrogram of a piano playing a C (262 Hz). (b) The DUET source estimate of the same piano tone when extracted from a mixture with a saxophone playing G and French horn playing C. (c) The source estimate of the same piano tone extracted from the same mixture using the proposed source separation algorithm.

Every and Szymanski [15] assume that pitches are known in advance. Overlap regions are identified based on instrument pitch and resolved by linearly interpolating between neighboring harmonics of each source and applying spectral-filtering to the mixture. This approach resolves the limitations imposed by additive synthesis in [17, 18], but the assumption that linear interpolation between the amplitudes of known harmonics can be used to determine the amplitude of unknown harmonics is somewhat unrealistic.

In the two-channel case, Viste and Evangelista [19] show that they can perform iterative source separation by maximizing the correlation in amplitude modulation of frequency bands in the reconstructed source signals. Although this is a promising framework for demixing overlapping signals, the current approach cannot be applied to mixtures where more than two signals overlap. Stereo recordings of three or more instruments frequently violate this constraint.

Vincent [16] proposes demixing stereo recordings with two or more instruments by incorporating CASA heuristics, spatial cues, and time-frequency source signal priors to cast the demixing problem into a Bayesian estimation framework.

This approach is designed to handle reverberant recordings, but requires significant prior knowledge of each source signal in the mixture, making it unsuitable for mixtures where the acoustic characteristics of each source are not known beforehand.

3. THE PROPOSED ALGORITHM

In this section, we present a new musical source separation algorithm. The proposed method is designed to separate anechoic, stereo recordings of any number of harmonic musical sources without prior analysis of the sources and without knowledge of the musical score. This method is similar to recent approaches in that it incorporates signal features commonly associated with CASA to achieve separation of signals that overlap in time-frequency. Our technique differs from existing methods in that it is designed to work when the number of sources exceeds the number of mixtures, the score is unknown, and prior modeling of source signals is not possible. Since we use an existing time-frequency masking approach for initial source separation, we require a portion of the time-frequency frames in the mixture contain energy from only one source signal. This requirement is, however, substantially reduced when compared to existing time-frequency masking techniques.

3.1. Overview

Assume that **N** sources are recorded using two microphones. If the sound sources are in different locations, the distance that each source travels to the individual microphones will produce a unique amplitude and timing difference between the two recorded signals. These differences, often called spatial cues or mixing parameters, provide information about the position of the sources relative to the microphones. The first step in numerous BSS methods is the determination of mixing parameters for each source signal. Once mixing parameters are determined, they can be used to distribute time-frequency frames from the mixture to individual source signals. In our approach, we assume that mixing parameters can be determined using the DUET [1] algorithm (Section 3.2), or from known source locations.

In assigning energy from a time-frequency frame in a pair of anechoic mixtures to a set of sources, we note three cases of interest. The first case is where at most one source is active; we call these *one-source frames*. In this case, the full energy from one mixture may be assigned directly to an estimate of the source **j**, denoted by $\hat{\mathbf{S}}_j$. The second case is where exactly two sources are active; *two-source frames*. In this case, we can explicitly solve for the correct energy distribution to each active source using the system of equations provided by (1). The third case is where more than two sources are active; *multisource frames*. Since there are at least three unknown complex values, we cannot solve for the appropriate source energy and must develop methods to estimate this energy.

We approach source separation in three stages, corresponding to the three cases described above. Figure 2 provides a diagram of the three stages of analysis and reconstruction in the proposed algorithm. In the first stage (Section 3.3), we create initial signal estimates using the *delay and scale subtraction scoring* (DASSS) method [21], which identifies time-frequency frames from the mixture that contain energy from only one source. If we assume that sources are harmonic and monophonic, there is often sufficient information in these initial signal estimates to determine the fundamental frequency of each source.

If fundamental frequencies can be determined, we can estimate the time-frequency frames associated with each source's harmonics, which lets us categorize additional mixture frames as one-source, two-source, or multisource. Two-source frames are then distributed, further refining the source estimates. This is the second stage of source reconstruction (Section 3.4).

In the final stage (Section 3.5) we analyze the amplitude modulation of the partially reconstructed sources to inform the estimation of source energy in multisource frames. The remainder of this section describes the implementation of the proposed source separation algorithm in greater detail.

3.2. Mixing parameter estimation

In this section, we give a brief overview of mixing parameter estimation using DUET. A more thorough discussion of parameter estimation and the demixing approach taken in DUET is provided in [1].

Let $\mathbf{X}_1(\boldsymbol{\tau},\boldsymbol{\omega})$ and $\mathbf{X}_2(\boldsymbol{\tau},\boldsymbol{\omega})$ represent the short-time Fourier transforms of two signal mixtures containing **N** source signals, $\mathbf{S}_j(\boldsymbol{\tau},\boldsymbol{\omega})$, recorded by two, omni-directional microphones,

$$\begin{aligned} X_1(\tau,\omega) &= \sum_{j=1}^{N} S_j(\tau,\omega), \\ X_2(\tau,\omega) &= \sum_{j=1}^{N} a_j e^{-i\omega\delta_j} S_j(\tau,\omega). \end{aligned} \tag{1}$$

Here, $\mathbf{a}_j$ is the amplitude scaling coefficient and $\boldsymbol{\delta}_j$ is the time-shift between the two microphones for the jth source, $\boldsymbol{\tau}$ represents the center of a time window, and $\boldsymbol{\omega}$ represents a frequency of analysis used in the STFT. Given these mixture models, parameter estimation is simply associating a particular amplitude scaling and time-shift value with each source.

DUET assumes that signals are approximately *window-disjoint orthogonal*, meaning that most time-frequency frames in the mixture contain energy from no more than one source [1, 9]. Any frame that meets this requirement should match the amplitude scaling, $\mathbf{a}_j$, and time-shift, $\boldsymbol{\delta}_j$, properties resulting from one source's physical location relative to the microphones. Finding the most common pairs of amplitude scaling and time-shift values between the two mixtures provides a means of estimating the mixing parameters of each source.

In the rest of this work we assume that the amplitude scaling, $\mathbf{a}_j$, and time-shift, $\boldsymbol{\delta}_j$, can be estimated correctly for each source **j** using DUET's parameter estimation. Alternate approaches that simulate binaural hearing in humans have

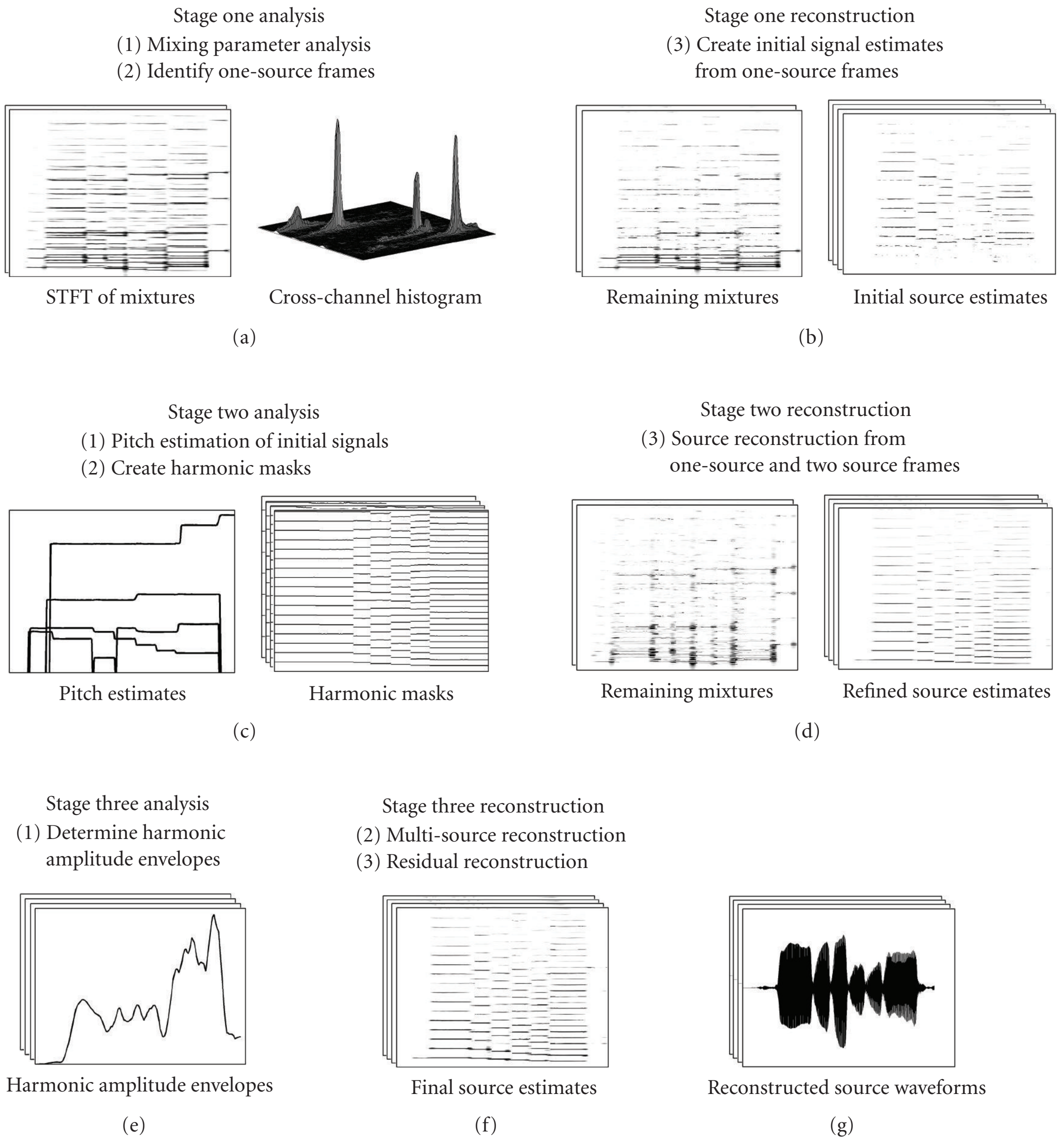

Stage one analysis
(1) Mixing parameter analysis
(2) Identify one-source frames
STFT of mixtures
Cross-channel histogram
(a)
Stage one reconstruction
(3) Create initial signal estimates from one-source frames
Remaining mixtures
Initial source estimates
(b)
Stage two analysis
(1) Pitch estimation of initial signals
(2) Create harmonic masks
Pitch estimates
Harmonic masks
(c)
Stage two reconstruction
(3) Source reconstruction from one-source and two source frames
Remaining mixtures
Refined source estimates
(d)
Stage three analysis
(1) Determine harmonic amplitude envelopes
Harmonic amplitude envelopes
(e)
Stage three reconstruction
(2) Multi-source reconstruction
(3) Residual reconstruction
Final source estimates
(f)
Reconstructed source waveforms
(g)

Figure 2: An illustration of the three stages of the proposed source separation algorithm.

been proposed to localize and separate source sounds with significant overlap or in reverberant environments [22–24], however in this work we assume that recordings are made with a stereo pair of omni-directional microphones.

3.3. Stage one: DASSS analysis and initial source reconstruction

The DUET algorithm allows for successful demixing when sources do not simultaneously produce energy at the same frequency and time. The DASSS method [21] was developed to determine which time-frequency frames of the mixture satisfy this condition, allowing reconstruction of sources from only the *disjoint*, or one-source frames. Our approach uses DASSS in the first stage to create partial signal estimates from the single source frames. These estimates are then analyzed to provide guidance in further distribution of mixture frames.

3.3.1. Finding one-source frames

To determine which frames in a stereo mixture correspond to a single source, define a function, $\mathbf{Y}_j$, for each pair of mixing parameters, $(\mathbf{a}_j, \boldsymbol{\delta}_j)$, associated with a source signal $\mathbf{j}$,

$$Y_j(\tau,\omega) = X_1(\tau,\omega) - \frac{1}{a_j} e^{i\omega\delta_j} X_2(\tau,\omega). \tag{2}$$

If only one source is active in a given time-frequency frame, $\mathbf{Y}_j(\boldsymbol{\tau},\boldsymbol{\omega})$ takes on one of two values. Equation (3) represents the expected values of the $\mathbf{Y}_j(\boldsymbol{\tau},\boldsymbol{\omega})$ functions, under the assumption that a single source, $\mathbf{g}$ (represented by the superscript ^{g}), was active,

$$\widehat{Y}_j^g(\tau,\omega) = \begin{cases} 0, & \text{if } j = g, \\ \left(1 - \dfrac{a_g}{a_j} e^{i\omega(\delta_j - \delta_g)}\right) X_1(\tau,\omega), & \text{if } j \neq g. \end{cases} \tag{3}$$

Equation (4) is a scoring function to compare the expected values in $\hat{Y}_j^g(\tau,\omega)$ to the calculated $\mathbf{Y}_j(\boldsymbol{\tau},\boldsymbol{\omega})$,

$$d(g,\tau,\omega)=\frac{\sum_{\forall j}\left|\hat{Y}_j^g(\tau,\omega)-Y_j(\tau,\omega)\right|}{\sum_{\forall j}\left|Y_j(\tau,\omega)\right|}. \tag{4}$$

As the function $\mathbf{d}(\mathbf{g},\boldsymbol{\tau},\boldsymbol{\omega})$ approaches zero, the likelihood that source $\boldsymbol{g}$ was the only active source during the time-frequency frame $(\boldsymbol{\tau},\boldsymbol{\omega})$ increases. A threshold value can then be used to determine which frames are one-source. These frames can be assigned directly to the estimate for source $\mathbf{g}$ [21].

3.3.2. Initial source reconstruction

We distribute the full energy from each one-source frame directly to the appropriate initial signal estimate, $\hat{\mathbf{S}}_g$, as shown in (5),

$$\hat{S}_g(\tau,\omega)=\begin{cases}X_1(\tau,\omega), & \mathit{if}\Big(d(g,\tau,\omega)<T\\ & \quad\wedge g=\arg\min_{\forall j}\big(d(j,\tau,\omega)\big)\Big)\\ 0, & \mathit{else}.\end{cases} \tag{5}$$

Here, $\mathbf{T}$ is a threshold value that determines how much energy from multiple sources a frame may contain and still be considered a one-source frame. When setting $\mathbf{T}$, we must both limit the error in $\hat{\mathbf{S}}_g$ and distribute enough frames to each source estimate so fundamental frequency estimation in stage two is possible. We have found that $\mathbf{T}=0.15$ balances these two requirements well [25]. Once an initial signal estimate is created for each source, the signals are analyzed and further source reconstruction is accomplished in stage two.

3.4. Stage two: source activity analysis and further source reconstruction

In this stage, we estimate the fundamental frequency of each source from the partially reconstructed signals. These estimates are used to create *harmonic masks*. The harmonic mask for a source indicates time-frequency regions where we expect energy from that source, given its fundamental frequency. We use these masks to estimate the number of active sources in important time-frequency frames remaining in the mixture. We then refine the initial source estimates by distributing mixture energy from additional mixture frames in which either one or two sources are estimated to contain significant energy.

3.4.1. Determining the active source count using harmonic masks

We first determine the fundamental frequency of each signal estimate using an auto-correlation-based technique described in [26]. We denote the fundamental frequency of signal estimate $\hat{\mathbf{S}}_g$ for time window $\boldsymbol{\tau}$ as $\mathbf{F}_g(\boldsymbol{\tau})$.

Since this estimation is based on partially reconstructed sources, we employ two rules to refine the fundamental frequency estimates of each source. The first eliminates spurious, short-lived variation in the $\mathbf{F}_g$ estimates. The second adjusts $\mathbf{F}_g$ values that we have low confidence in, based on the amount of energy distributed to the source estimate during stage one. Details on the refinement of the fundamental frequency estimates based on these rules are provided in [25].

Since we assume harmonic sound sources, we expect there to be energy at integer multiples of the fundamental frequency of each source. Accordingly, we create a *harmonic mask*, $\mathbf{M}_g(\boldsymbol{\tau},\boldsymbol{\omega})$, a binary time-frequency mask for each source. Each mask has a value of 1 for frames near integer multiples of the fundamental frequency and a value of 0 for all other time-frequency frames,

$$M_g(\tau,\omega)=\begin{cases}1, & \mathit{if}\ (\exists k\ \mathit{such\ that}\ |kF_g(\tau)-\omega|<\Delta_\omega),\\ 0, & \mathit{else}.\end{cases} \tag{6}$$

Here, $\mathbf{k}$ is an integer and Δ_ω is the maximal allowed difference in frequency from the kth harmonic. We set Δ_ω to 1.5 times the frequency resolution used in the STFT processing.

We use the harmonic masks to divide high-energy frames of the mixtures into three categories: one-source frames, two-source frames, and multisource frames. We do this by summing the harmonic masks for all the sources to create the *active source count* for each frame, $\mathbf{C}(\boldsymbol{\tau},\boldsymbol{\omega})$,

$$C(\tau,\omega)=\sum_{\forall g}M_g(\tau,\omega). \tag{7}$$

3.4.2. Further source reconstruction

Identification of one-source frames using DASSS is not perfect because two sources can interfere with each other and match the cross-channel amplitude scaling and time-shift characteristics of a third source. Also, we set the threshold in (5) to accept enough time-frequency frames to estimate $\mathbf{F}_g(\boldsymbol{\tau})$ for each source. We remove energy that might have been mistakenly given to each source in (8),

$$\hat{S}_g^{\text{two}}(\tau,\omega)=\hat{S}_g^{\text{one}}(\tau,\omega)M_g(\tau,\omega). \tag{8}$$

In (8) and (9) we add the superscripts "*one*" and "*two*" to clarify which stage of source reconstruction is specified. Thus, (8) eliminates time-frequency frames from the initial source estimates that are not near the predicted harmonics of that source. In time-frequency frames where the source count $\mathbf{C}(\boldsymbol{\tau},\boldsymbol{\omega})=1$ and the stage one estimate is zero, we add energy to the stage two estimates, as shown in (9),

$$\begin{gathered}\hat{S}_g^{\text{two}}(\tau,\omega)=X_1(\tau,\omega),\\ \mathit{if}\Big(C(\tau,\omega)=M_g(\tau,\omega)=1\wedge\hat{S}_g^{\text{one}}(\tau,\omega)=0\Big).\end{gathered} \tag{9}$$

In time-frequency frames where the source count $\mathbf{C}(\boldsymbol{\tau},\boldsymbol{\omega})=2$, we presume the frame has two active sources and use the system of equations in (10) and (11) to solve for the source values,

$$X_1(\tau,\omega)\approx S_g(\tau,\omega)+S_h(\tau,\omega), \tag{10}$$

$$X_2(\tau,\omega)\approx a_g e^{-i\omega\delta_g}S_g(\tau,\omega)+a_h e^{-i\omega\delta_h}S_h(\tau,\omega). \tag{11}$$

We can solve for source $\mathbf{g}$ as in (12), and use (10) to solve for source $\mathbf{h}$,

$$\hat{S}_g(\tau,\omega) = \frac{X_2(\tau,\omega) - a_h e^{-i\omega\delta_h} X_1(\tau,\omega)}{a_g e^{-i\omega\delta_g} - a_h e^{-i\omega\delta_h}}. \tag{12}$$

Once we have calculated the energy for both sources in the frame, we add this energy to the source signal estimates. Any time-frequency frames with $\mathbf{C}(\boldsymbol{\tau},\boldsymbol{\omega}) > 2$ are distributed in stage three.

3.5. Stage three: amplitude modulation analysis and final reconstruction

In this section we propose a method to estimate the energy contribution from each source in a multisource mixture frame, using the reconstructed source signals created during stages one and two as guides.

We first note that when instrument pitches are stable for even a short duration of time (20 milliseconds or so), overlap between source signals tends to occur in sequences of time-frequency frames. With this in mind, the proposed multisource estimation method deals with sequences of time frames at a particular frequency of analysis when possible.

Let $[\boldsymbol{\tau}_s, \boldsymbol{\tau}_{\mathbf{s+n}}]$ be a sequence of multisource frames at frequency of analysis $\boldsymbol{\omega}$. In order to estimate the energy in multiple sources over this sequence of time-frequency frames, we assume that each source signal's harmonics will have correlated amplitude envelopes over time. Although this is not precisely the case, this principle is used in instrument synthesis [20], and source separation [2, 14, 19]. CASA algorithms also commonly use correlated amplitude modulation as a grouping mechanism [11–13].

A *harmonic amplitude envelope* is an estimate of the amplitude modulation trend of a source, based on the harmonics reconstructed in stages one and two. We use these envelopes to estimate the energy for harmonics that could not be resolved in the first two stages, due to overlap with multiple sources. To do this for a sequence of multisource frames $[\boldsymbol{\tau}_s, \boldsymbol{\tau}_{\mathbf{s+n}}]$ at frequency $\boldsymbol{\omega}$ we require an estimate of $\hat{\mathbf{S}}_g(\boldsymbol{\tau}_s,\boldsymbol{\omega})$, the complex value of each active source at the beginning of the sequence. If we assume that each source's phase progresses linearly over the sequence, the harmonic amplitude envelopes let us estimate how each source's energy changes during the sequence. We can then appropriately assign energy to each active source $\mathbf{g}$ in frames $\hat{\mathbf{S}}_g(\boldsymbol{\tau}_{\mathbf{s+1}},\boldsymbol{\omega})$ through $\hat{\mathbf{S}}_g(\boldsymbol{\tau}_{\mathbf{s+n}},\boldsymbol{\omega})$.

We now describe our method to determine *harmonic amplitude envelopes*, and then proceed with a discussion of how to estimate $\hat{\mathbf{S}}_g(\boldsymbol{\tau}_s,\boldsymbol{\omega})$, the first complex value of each active source in the sequence of multisource frames.

3.5.1. Determining harmonic amplitude envelopes

To calculate the overall harmonic amplitude envelope for source $\mathbf{g}$, we first find the amplitude envelope of each harmonic in the signal estimate for $\mathbf{g}$, using (13). Here, $\mathbf{k}$ denotes the harmonic number and $\mathbf{A}_g(\boldsymbol{\tau},\mathbf{k})$ is the amplitude envelope for the kth harmonic. Equation (14) defines which time-frequency frames we include in the estimate of $\mathbf{A}_g(\boldsymbol{\tau},\mathbf{k})$. A frame is included if both the center frequency of the frame is within Δ_ω of the harmonic frequency (see (6)) and the source signal estimate from stage two contains energy in that frame,

$$A_g(\tau,k) = \text{mean}_{\forall\omega\in\Gamma(k)}\left(\left|\hat{S}_g(\tau,\omega)\right|\right), \tag{13}$$

$$\omega \in \Gamma(k) \text{ if } \left(\left|\omega - kF_g(\tau)\right| < \Delta_\omega \wedge \hat{S}_g(\tau,\omega) > 0\right). \tag{14}$$

Equation (15) normalizes each amplitude envelope so that each harmonic contributes equally to the overall amplitude envelope,

$$\tilde{A}_g(\tau,k) = \frac{A_g(\tau,k)}{\max_{\forall\tau}\left(A_g(\tau,k)\right)}. \tag{15}$$

Equation (16) is used to determine the overall harmonic amplitude envelope, which we denote, $\mathbf{H}_g(\boldsymbol{\tau})$. This equation simply finds the average amplitude envelope over all harmonics, and scales this envelope by the *short-term energy* of the signal estimate, as shown in (17). Here, $\mathbf{L}$ specifies a time window over which the signal energy is calculated. We include the amplitude scaling in (16) so the relative strength of each source's harmonic amplitude envelope corresponds to the overall loudness of each source during the time window $\mathbf{L}$,

$$H_g(\tau) = \text{mean}_{\forall k}\left(\tilde{A}_g(\tau,k)\right)E_g(\tau), \tag{16}$$

$$E_g(\tau) = \sum_{\lambda=-L/2}^{L/2}\sum_{\forall\omega}\left|\hat{S}_g(\tau+\lambda,\omega)\right|^2. \tag{17}$$

3.5.2. Estimating $\hat{S}_g(\tau_s,\omega)$

If, for each source g, the first value in the sequence, $\hat{\mathbf{S}}_g(\boldsymbol{\tau}_{\mathbf{s}},\boldsymbol{\omega})$, can be estimated, then (18) and (19) can be used to estimate the values of the sources in the remaining multisource frames, $[\boldsymbol{\tau}_{\mathbf{s+1}}, \boldsymbol{\tau}_{\mathbf{s+n}}]$. Here, we set $\boldsymbol{\tau}_{\mathbf{a}} = \boldsymbol{\tau}_{\mathbf{s}}$ and $\boldsymbol{\tau}_{\mathbf{b}} \in [\boldsymbol{\tau}_{\mathbf{s+1}}, \boldsymbol{\tau}_{\mathbf{s+n}}]$,

$$\left|\hat{S}_g(\tau_b,\omega)\right| = \frac{H_g(\tau_b)}{H_g(\tau_a)}\left|\hat{S}_g(\tau_a,\omega)\right|, \tag{18}$$

$$\angle\hat{S}_g(\tau_b,\omega) = \text{mod}\left(\angle\hat{S}_g(\tau_a,\omega) + (\tau_b-\tau_a)\omega, 2\pi\right). \tag{19}$$

3.5.3. Estimation from a prior example

The frame immediately before the start of the sequence of multisource frames in question is $(\boldsymbol{\tau}_{\mathbf{s-1}},\boldsymbol{\omega})$. If a source estimate was already given energy in this frame during stage one or two (i.e., if $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s-1}},\boldsymbol{\omega})| > 0$), we can use $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s-1}},\boldsymbol{\omega})$ to estimate $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}},\boldsymbol{\omega})$ using (18) and (19) by setting $\boldsymbol{\tau}_{\mathbf{a}} = \boldsymbol{\tau}_{\mathbf{s-1}}$ and $\boldsymbol{\tau}_{\mathbf{b}} = \boldsymbol{\tau}_{\mathbf{s}}$.

Since stage one and two only resolve one-source and two-source frames, no matter how many sources we are estimating in frame $\boldsymbol{\tau}_{\mathbf{s}}$, we can expect that $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s-1}},\boldsymbol{\omega})| > 0$ for at most two sources. We estimate $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\omega})|$ for the remaining active sources by assuming that the relationship between the amplitudes of two different sources' harmonics at frequency $\boldsymbol{\omega}$ will be proportional to the relationship between the two sources' average harmonic amplitude, or $\mathbf{H}_{\mathbf{g}}(\boldsymbol{\tau})$.

We denote a source whose amplitude was estimated using (18) as $\mathbf{n}$, and now estimate the amplitude of any remaining active source in frame $\boldsymbol{\tau}_{\mathbf{s}}$,

$$|\hat{S}_g(\tau_s, \omega)| = \frac{H_g(\tau_s)}{H_n(\tau_s)} |\hat{S}_n(\tau_s, \omega)|. \tag{20}$$

We set the phase of sources whose amplitudes are derived using (20) to a value of 0.

3.5.4. Estimation without a prior example

If after stage two, $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s-1}}, \boldsymbol{\omega})| = 0$ for all sources, we must use an alternate method of estimating $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\omega})$. In this case, we rely on the assumption that overlapping signals will cause amplitude *beating* (amplitude modulation resulting from interference between signals) in the mixture signals. The time frame with maximal amplitude in the mixture signals during the sequence $[\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\tau}_{\mathbf{s+n}}]$ corresponds to the frame in which the most constructive interference between active sources takes place. We assume that this point of maximal constructive interference results from all active sources having equal phase and call this frame $\boldsymbol{\tau}_{\mathbf{MaxInt}}$. With this assumption, (8), altered for the $\mathbf{N}$ active source case in frame $(\boldsymbol{\tau}_{\mathbf{MaxInt}}, \boldsymbol{\omega})$, yields (21), where Φ is the set of active sources in the multisource sequence, $[\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\tau}_{\mathbf{s+n}}]$, as determined by the harmonic masks,

$$|X_1(\tau_{\text{MaxInt}}, \omega)| \approx \sum_{\forall g \in \Phi} |S_g(\tau_{\text{MaxInt}}, \omega)|. \tag{21}$$

The amplitude of any active source $\mathbf{g}$ can then be determined using (22),

$$|\hat{S}_g(\tau_{\text{MaxInt}}, \omega)| = |X_1(\tau_{\text{MaxInt}}, \omega)| \frac{H_g(\tau_{\text{MaxInt}})}{\sum_{\forall h \in \Phi} H_h(\tau_{\text{MaxInt}})}. \tag{22}$$

To find $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\omega})|$ from $|\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{MaxInt}}, \boldsymbol{\omega})|$ we apply (18) with $\boldsymbol{\tau}_{\mathbf{a}} = \boldsymbol{\tau}_{\mathbf{MaxInt}}$ and $\boldsymbol{\tau}_{\mathbf{b}} = \boldsymbol{\tau}_{\mathbf{s}}$. We set the phase values of each active source during the first frame, $\angle\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\omega})$, to a default value of 0.

We now apply (18) and (19) to determine $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s+1}}, \boldsymbol{\omega})$ through $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s+n}}, \boldsymbol{\omega})$ from $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}_{\mathbf{s}}, \boldsymbol{\omega})$, and complete this process for each sequence of multisource frames determined by the source count, $\mathbf{C}(\boldsymbol{\tau}, \boldsymbol{\omega})$.

3.5.5. Distributing residual energy

Thus far, we have focused our attention on the harmonic regions of individual source signals. Even though we are assuming that source signals are harmonic, harmonic instrument signals also contain energy at nonharmonic frequencies due to factors such as excitation noise [20]. The nonharmonic energy in a harmonic signal is often called the *residual energy*. We take a simple approach to the distribution of residual energy in that we distribute any remaining time-frequency frame of the mixture to the most likely source using an altered version of (5), shown in (23),

$$\hat{S}_g(\tau, \omega) = \begin{cases} X_1(\tau, \omega), & \text{if } \big(g = \arg\min_{\forall j}(d(j, \tau, \omega))\big), \\ 0, & \text{else.} \end{cases} \tag{23}$$

Once the residual energy has been distributed, each source estimate, $\hat{\mathbf{S}}_{\mathbf{g}}(\boldsymbol{\tau}, \boldsymbol{\omega})$, is transformed back into the time domain using the overlap-add technique [27]. The result is a time domain waveform of each reconstructed source signal.

4. EXPERIMENTAL RESULTS

In this section we compare the performance of the proposed method and the DUET algorithm on three and four instrument mixtures. We chose to compare performance to DUET because our approach is designed with the same mixture models and constraints, making it a natural extension of time-frequency masking techniques such as DUET. In previous work [25, 28] we have called our approach the *active source estimation* algorithm. For convenience, we refer to our method as ASE in the discussion below.

4.1. Mixture creation

The instrument recordings used in the testing mixtures are individual long-tones played by alto flute, alto and soprano saxophones, bassoon, B-flat and E-flat clarinets, French horn, oboe, trombone, and trumpet, all taken from the University of Iowa musical instrument database [29].

Mixtures of these recordings were created to simulate the stereo microphone pickup of spaced source sounds in an anechoic environment. We assume omni-directional microphones, spaced according to the highest frequency we expect to process, as in [1]. Instruments were placed in a semicircle around the microphone pair at a distance of one meter. In the three-instrument mixtures, the difference in azimuth angle from the sources to the microphones was 90°. In the four-instrument case, it was 60°.

For each mixture, each source signal was assigned a randomly selected instrument and a randomly selected pitch from 13 pitches of the equal tempered scale, C4 through C5. We created 1000 three-instrument mixtures and 1000 four-instrument mixtures in this manner.

We wanted mixtures to realistically simulate a performance scenario in which instrument attacks are closely aligned. For this reason, each sample used was hand cropped so that the source energy is present at the beginning of the file. Although the instrument attack times vary to some extent, cropping samples in this manner ensures that the created mixtures contain each instrument in all time frames of analysis.

Each source was normalized to have unit energy prior to mixing. Mixtures were created at 22.05 kHz and 16 bits, and were 1 second in length. Mixtures were separated into reconstructed source signals by our method and the DUET algorithm, using a window length of 46 milliseconds and step size of 6 milliseconds for STFT processing.

Extracted sources were then compared to the original sources using the *signal-to-distortion ratio* (**SDR**) described in [30]. In (24), **s** represents the original time-domain source signal,

$$\mathrm{SDR} = 10\log_{10}\left(\frac{\left|\langle\hat{s}, s\rangle\right|^2}{\left|\langle\hat{s}, \hat{s}\rangle\right|^2 - \left|\langle\hat{s}, s\rangle\right|^2}\right). \qquad (24)$$

4.2. Results

In order to assess the utility of the multisource distribution stage proposed in Section 3.5, we compared performance results using the full algorithm as presented in Section 3 (denoted as ASE 1 in Table 1) and a simpler multisource distribution scheme. The alternate algorithm, denoted as ASE 2, is identical to ASE 1 until the multisource distribution stage from Section 3.5, where ASE 2 distributes multisource frames of the mixture, unaltered, to each active source.

Table 1 shows the median performance of ASE 1, ASE 2, and DUET on the testing data. The median performance is measured over the total number of source signals, 3000 in the three-instrument tests and 4000 in the four-instrument tests. Results of all mixtures containing consonant musical intervals are also shown. The ASE performance data is not normally distributed, thus we do not show means and standard deviations of the **SDR** data. In a nonparametric sign test performed over all mixtures, we found the median performance to be significantly different between ASE 1, ASE 2, and DUET, with $\mathbf{p} < 10^{-50}$ in all three comparisons.

The sole difference between ASE 1 and ASE 2 is in the method used to assign energy from time-frequency frames with energy from three or more sources. The results in Table 1 indicate that the multi-source energy assignment method in Section 3.5 improves performance, when compared to a simpler approach of simply assigning multisource energy evenly to each active source.

Table 1: Median signal-to-distortion ratio of the ASE and DUET algorithms on 1000 three-instrument mixtures (3000 signals) and 1000 four-instrument mixtures (4000 signals). The table also shows median performance on three- and four-instrument mixtures containing specific musical intervals: unison (2383 signals), octave (366 signals), perfect fifth (1395 signals), and perfect fourth (1812 signals). Higher values are better.

	ASE 1	ASE 2	DUET
All mixtures	13.77 dB	12.26 dB	10.22 dB
Three-instrument mixtures	18.63 dB	17.57 dB	14.12 dB
Four-instrument mixtures	10.22 dB	9.01 dB	8.13 dB
Unison	4.72 dB	3.63 dB	2.92 dB
Octave	8.79 dB	6.82 dB	6.38 dB
Fifth	13.36 dB	11.44 dB	8.13 dB
Fourth	13.99 dB	13.05 dB	10.45 dB

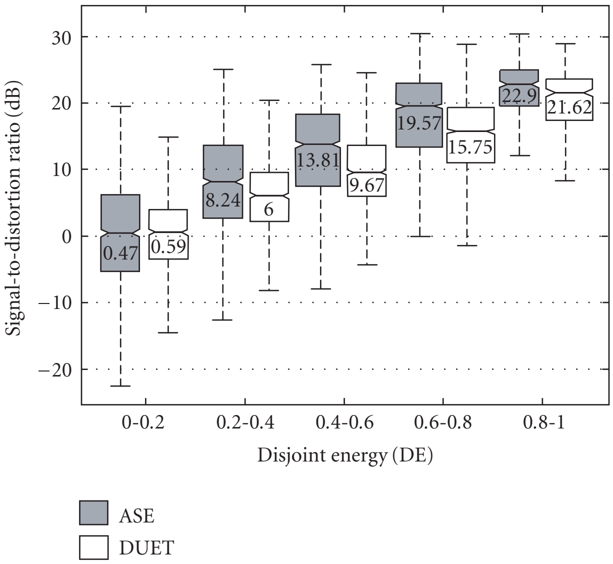


Figure 3: ASE 1 and DUET SDR performance over five groups of signals. Signals are grouped according to disjoint energy, **DE**. Median performance is shown in the lower half of each box. Higher values are better.

A primary goal of the ASE system was to reduce the reliance on nearly disjoint source signals, when compared to existing time-frequency masking techniques. To determine how both ASE and DUET perform as a function of interference from other sources, we use a measure of *disjoint energy*, **DE**. Disjoint energy represents the amount of energy in a source signal that *is not* heavily interfered with by other sources in the mix. We calculate **DE** as a simple ratio, where the energy in all time-frequency frames that are deemed disjoint (less than 1 dB error caused by interfering sources) in a particular mixture is divided by the total energy in the signal, resulting in a value between 0 and 1. A **DE** score of 0 reflects that all time-frequency frames of a source signal are distorted by at least 1 dB due to the other sources in the mixture, while a value of 1 reflects that interference from other sources is restricted to less than 1 dB in all time-frequency frames. We chose the error threshold of 1 dB because on informal tests, subjects were unable to detect random amplitude distortions of less than 1 dB when applied to all time-frequency frames of a signal independently. More details on the calculation of **DE** are provided in [25].

Figure 3 shows **SDR** performance for ASE 1 and DUET as a function of **DE**. We first divided the data set into five categories: source signals with $\mathbf{DE} \in (0, 0.2)$, $(0.2, 0.4)$, $(0.4, 0.6)$, $(0.6, 0.8)$, and $(0.8, 1)$. We show boxplots of the **SDR** performance by ASE 1 and DUET on all signals within these groupings. The lower and upper lines of each box show 25th and 75th percentiles of the sample. The line in the middle of each box is the sample median. The lines extending above and below the box show the extent of the rest of the sample, excluding outliers. Outliers are defined as points further from the sample median than 1.5 times the interquartile range and are not shown.

When disjoint energy is 0.8 or greater, both ASE and DUET do quite well in source separation and the performance improvement provided by our approach is moderate. As the disjoint energy in a source signal decreases, the improvement provided by ASE increases, as we can see on signals with **DE** between 0.2 and 0.8. This suggests that our approach can deal more effectively with partially obstructed source signals. Performance improvement is greatest for signals with **DE** between 0.4 and 0.6 (over 4 dB), or signals with roughly half of their energy unobstructed. As a source signal's **DE** falls below 0.2, the performance by both algorithms is poor, although only 17.56% of the signals in the mixtures created for this study had **DE** below 0.2.

It is also clear that as **DE** falls, the variability of ASE **SDR** performance increases. This results from the fact that ASE relies on fundamental frequency estimation of partial signals, created from only the disjoint (nonoverlapping) time-frequency frames of each signal. In cases where fundamental frequency is estimated correctly, performance of ASE is good despite significant source overlap. When fundamental frequencies are incorrect, reconstruction of signals can be degraded when compared to DUET. While this is a limitation of our approach, the data is promising in that more reliable fundamental frequency estimation techniques may provide significant performance improvements. We found that fundamental frequencies were estimated correctly in 89.42% of the total time frames in the three-instrument data set and in 84.3% of the time frames in the four-instrument data set. In other work, we have seen that using pitch information provided by an aligned musical score can lead to statistically significant **SDR** improvements averaging nearly 2 dB [28] on a corpus of four-part Bach chorales.

5. CONCLUSIONS AND FUTURE WORK

In this work we have presented a method to extend time-frequency disjoint techniques for blind source separation to the case where there are harmonic sources with significant time-frequency overlap. We showed our method's improvement over the DUET method at separating individual musical instruments from contexts which contain low amounts of disjoint signal energy.

We improve source reconstruction by predicting the expected time-frequency locations of source harmonics. These predictions are used to determine which sources are active in each time-frequency frame. These predictions are based on fundamental frequencies estimated from incomplete source reconstructions. In the future, we intend to develop methods to generate source templates from disjoint mixture regions that do not assume harmonic sources.

In this paper, we introduced an analytic approach to assign energy from two-source time-frequency frames. Our methods of assigning energy from frames with more than two sources make somewhat unrealistic assumptions. Despite this, source separation is still improved, when compared to systems that do not attempt to appropriately assign energy from time-frequency frames with three or more sources. In future work we will explore improved ways to determine source amplitude and phase in these cases.

The theme of this work and our future work will remain rooted in the idea of learning about source signals through partial output signals. Considering that in any truly blind algorithm we will have no a priori knowledge about the source signals, techniques such as these can provide the necessary means for deconstructing difficult mixtures.

Although there are still many obstacles which prevent robust, blind separation of real-world musical mixtures, the performance of our approach on anechoic mixtures provides promising evidence that we are nearing a tool that can effectively process real musical recordings.

REFERENCES

[1] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[2] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '00)*, pp. 215–220, Helsinki, Finland, June 2000.

[3] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, pp. 2129–2134, Houston, Tex, USA, June 1997.

[4] L. C. Parra and C. D. Spence, "Separation of non-stationary natural signals," in *Independent Component Analysis: Principles and Practice*, pp. 135–157, Cambridge University Press, Cambridge, Mass, USA, 2001.

[5] J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, MIT Press, Cambridge, Mass, USA, 2004.

[6] P. Aarabi, G. Shi, and O. Jahromi, "Robust speech separation using time-frequency masking," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '03)*, vol. 1, pp. 741–744, Baltimore, Md, USA, July 2003.

[7] R. Balan and J. Rosca, "Source separation using sparse discrete prior models," in *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS '05)*, Rennes, France, November 2005.

[8] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.

[9] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 1, pp. 529–532, Orlando, Fla, USA, May 2002.

[10] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Mass, USA, 1990.

[11] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1998.

[12] G. J. Brown and D. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 371–402, Springer, New York, NY, USA, 2005.

[13] D. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Media Laboratory, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1996.

[14] 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.

[15] M. Every and J. Szymanski, "A spectral-filtering approach to music signal separation," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx '04)*, pp. 197–200, Naples, Italy, October 2004.

[16] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[17] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 83–86, New Paltz, NY, USA, October 2001.

[18] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 2, pp. 1757–1760, Orlando, Fla, USA, May 2002.

[19] H. Viste and G. Evangelista, "Separation of harmonic instruments with overlapping partials in multi-channel mixtures," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 25–28, New Paltz, NY, USA, October 2003.

[20] J. C. Risset and D. Wessel, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, pp. 26–58, Academic Press, New York, NY, USA, 1982.

[21] A. S. Master, "Sound source separation of n sources from stereo signals via fitting to n models each lacking one source," Tech. Rep., CCRMA, Stanford University, Stanford, Calif, USA, 2003.

[22] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[23] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx '03)*, London, UK, September 2003.

[24] H. Viste and G. Evangelista, "Binaural source localization," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx '04)*, pp. 145–150, Naples, Italy, October 2004.

[25] J. Woodruff and B. Pardo, "Active source estimation for improved source separation," Tech. Rep. NWU-EECS-06-01, EECS Department, Northwestern University, Evanston, Ill, USA, 2006.

[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, vol. 17, pp. 97–110, Amsterdam, The Netherlands, 1993.

[27] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 1989.

[28] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR '06)*, Victoria, British Columbia, Canada, October 2006.

[29] L. Fritts, University of Iowa Musical Instrument Samples, http://theremin.music.uiowa.edu.

[30] R. Gribonval, L. Benaroya, E. Vincent, and C. Fevotte, "Proposals for performance measurement in source separation," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA '03)*, Nara, Japan, April 2003.

John Woodruff is a doctoral student and Teaching Assistant in the Ohio State University, Department of Computer Science and Engineering. He received a B.F.A. degree in performing arts and technology in 2002 and a B.S. degree in mathematics in 2004 from the University of Michigan. He received an M.Mus. degree in music technology in 2006 from Northwestern University. At Michigan, he was a Laboratory Instructor for the School of Music and both Manager and instructor for the sound recording facilities at the Duderstadt Center. While at Northwestern, he was a Research Assistant in the Department of Electrical Engineering and Computer Science and a Teaching Assistant in the School of Music. His current research interests include music source separation, music signal modeling, and computational auditory scene analysis. He is also an active Recording Engineer, Electroacoustic Composer, and Songwriter, and performs on both guitar and laptop. His music is available on the 482-music recording label.

Bryan Pardo is an Assistant Professor in the Northwestern University, Department of Electrical Engineering and Computer Science with a courtesy appointment in Northwestern University's School of Music. His academic career began at the Ohio State University, where he received both a B.Mus. degree in Jazz Composition and an M.S. degree in Computer Science. After graduation, he spent several years working as a Jazz Musician and Software Developer. As a Software Developer he worked for the Speech & Hearing Science Department of Ohio State and for the statistical software company SPSS. He then attended the University of Michigan, where he received an M.Mus. degree in Jazz and Improvisation, followed by a Ph.D. degree in Computer Science. Over the years, he has also been featured on five albums, taught for two years as an Adjunct Professor in the Music Department of Madonna University, and worked as a researcher for general dynamics on machine learning tasks. When he is not programming, writing, or teaching, he performs on saxophone and clarinet throughout the Midwest.