NORTHWESTERN
UNIVERSITY

Electrical Engineering and Computer Science Department

Technical Report
NWU-EECS-06-01
February 8, 2006

Active Source Estimation for Improved Source Separation

John Woodruff, Bryan Pardo

Abstract

Recent work in *blind source separation* applied to anechoic mixtures of speech allows for 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.

*An abbreviated version of this paper was submitted on December 1st, 2006, to the *EURASIP Journal on Applied Signal Processing.*

1. INTRODUCTION

Source separation is the process of determining individual source signals, when 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). BSS is an active area of research in many fields, including audio signal processing, telecommunications and medical imaging. This work focuses 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 facilitate post-production of pre-existing recordings, automated music transcription, vocalist and instrument identification, melodic comparison of polyphonic music, sample-based musical composition, multi-channel 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 the *Active Source Estimation* (ASE) approach, designed to isolate multiple simultaneous instruments from an anechoic, stereo mixture of tonal music. ASE 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 ASE to the DUET source separation algorithm on anechoic, stereo mixtures of three and four harmonic instruments, and a discussion of the advantages and limitations of using ASE. Finally, in section 5 we summarize our findings and discuss directions for future research.

2. CURRENT WORK IN SOURCE SEPARATION

Approaches to source separation in audio are numerous, and vary based on factors such as the number of mixture channels available, 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 statistical technique that can be used on the BSS problem when the number of mixtures equals or exceeds the number of source signals (Anemüller 2000, Hyvarinen 2000, Lee 1997, Parra 2001, Stone 2004). ICA assumes source signals are statistically independent, and iteratively determines time-invariant demixing filters to achieve maximal independence between sources. Often ICA is performed independently across frequency sub-bands, and bands are grouped based on cross-channel amplitude and time-shift differences (Lee 1997) or amplitude modulation (Anemüller 2000).

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 (O'Grady 2005). Sparse methods assume that in a time-frequency representation, most time-frequency frames of individual source signals will have magnitude near zero. 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 can be made (Rickard 2002). 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 (Aarabi 2003, Balan 2000, Jourjine 2000, Yilmaz 2004). The DUET algorithm, which we discuss in more detail in a later section, operates in this manner.

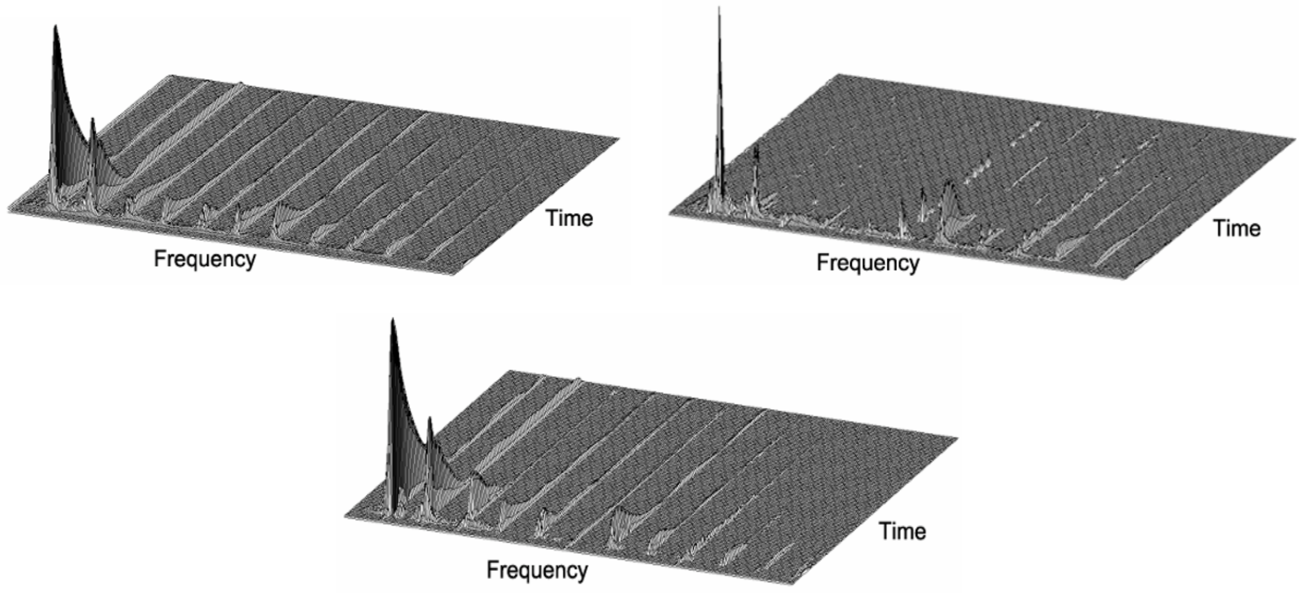

Figure 1: (top left) The spectrogram of a piano playing a C (262Hz). (top right) The spectrogram of the DUET source estimate of the same piano tone when extracted from a mixture with a saxophone playing G and French horn playing C. (bottom) The spectrogram of ASE source estimate of the same piano tone extracted from the same mixture.

Tonal music makes extensive use of multiple simultaneous instruments, playing *consonant intervals*. When two harmonic sources, such as pitched musical instruments, 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. This creates a problem for binary time-frequency masking methods that distribute each mixture frame to only one source signal. Reconstructed music signals can have audible gaps and artifacts. Figure 1 provides an illustration of this.

One approach to overcome the challenge presented by overlapping source signals has been to incorporate prior analysis or models of the source signals (Balan 2005, Ellis 2006, Reyes-Gomez 2004, Srinivasan 2005, Vincent 2004, Vincent 2005). Srinivasan and Wang (Srinivasan 2005) reconstruct a target speech signal when corrupted by interfering signals by using pre-determined phonemic templates and existing techniques for phonemic recognition from the corrupted speech signal. If the correct phonemes can be identified, the corrupted signal is refined based on the phonemic templates, resulting in improved audio quality. Due to the diversity of timbre in different instrument signals, pre-existing timbral templates, which would be the musical counterpart phonemic templates in speech, do not exist. Researchers that do not rely on pre-existing templates or models often analyze isolated source signals to create models that can be used in a similar manner (Balan 2005, Reyes-Gomez 2004, Vincent 2004, Vincent 2005). In this work we are interested in *blind* source separation, and thus avoid prior analysis of the individual signals.

Other researchers have incorporated heuristics commonly used in *Computational Auditory Scene Analysis* (CASA) to deal more effectively with source signal overlap. CASA researchers are interested in the source separation problem as it relates to human auditory perception. Humans are particularly adept at selectively listening to individual sound sources in a complex auditory scene. This provides motivation for computational methods based on the known principles governing the organization of sound by human listeners (Rosenthal 1998). CASA methods typically process a single mixture (one-channel) and model low-level auditory processing with a *correlogram* representation. High-energy components of the mixture are identified and grouped into *auditory objects* by utilizing perceptually important cues, such as pitch, amplitude and frequency modulation, and common onset and offset (Brown 2005, Hu 2004, Ellis 1996). An auditory object can be considered an individual sound event, a musical note or a spoken utterance. *Auditory streams* are formed by sequentially grouping the perceptually significant objects based on the sound source that generated them.

The goal of most CASA research is to create a symbolic representation of a sound scene in terms of individual sources (Rosenthal 1998). CASA heuristics can be used within source separation algorithms however, to both identify mixture regions in which source signals overlap and to guide the reconstruction of source signals in overlap regions (Anemüller 2000, Every 2004, Klapuri 2001, Vincent 2004, Vincent 2006, Virtanen 2001, Virtanen 2002, Viste 2003, Viste 2003). We now describe the implementation of these features in current music separation systems in more detail.

In the one-channel (monophonic) case, multiple researchers (Every 2004, Klapuri 2001, Virtanen 2001, Virtanen 2002) assume 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 (Klapuri 2001, Virtanen 2001, Virtanen 2002) use multi-pitch estimation to determine instrument pitches, time-frequency overlap regions are resolved by assuming 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 in 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 non-harmonic energy in pitched instrument signals (Risset 1982), which can cause the resulting signal to sound artificial.

Every and Szymanski (Every 2004) 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 (Virtanen 2001, Virtanen 2002), but the assumption that linear interpolation between the amplitude of known harmonics can be used to determine the amplitude of unknown harmonics is somewhat unrealistic.

In the two-channel case, Viste and Evangelista (Viste 2003b) show 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 and Rodet (Vincent 2004, Vincent 2006) propose 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. ACTIVE SOURCE ESTIMATION

In this section, we present the *Active Source Estimation* (ASE) algorithm. ASE is designed to separate anechoic, two-mixture (stereo) recordings of any number of harmonic musical sources without prior analysis of the sources and without knowledge of the musical score. ASE 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 ASE uses an existing time-frequency masking approach for initial source separation, it requires 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 An Overview of ASE

Assume $\boldsymbol{N}$ sources are recorded using two microphones. If the sound sources are in different locations, the distances that each source travels to the individual microphones will produce a specific 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 algorithm, 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 $\boldsymbol{j}$, denoted $\hat{\boldsymbol{S}}_j$. 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) and (2). The third case is where more than two sources are active; *multi-source 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 ASE. In the first stage, we create initial signal estimates using the *Delay and Scale Subtraction Scoring* (DASSS) method (Master 2003), which identifies time-frequency frames from the mixture that contain energy from only one source. If we assume 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 multi-source. Two-source frames are then distributed, further refining the source estimates.

In the final stage we analyze the amplitude modulation of the partially reconstructed sources to inform the estimation of source energy in multi-source frames. The remainder of this section describes the implementation of the ASE algorithm in greater detail.

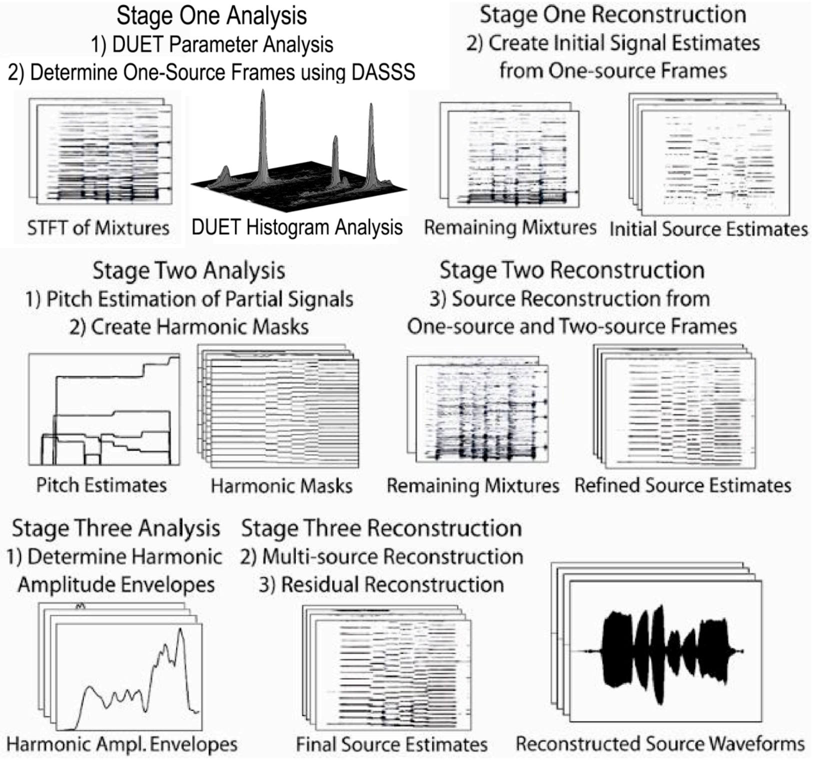
Stage One Analysis
1) DUET Parameter Analysis
2) Determine One-Source Frames using DASSS
Stage One Reconstruction
2) Create Initial Signal Estimates from One-source Frames
STFT of Mixtures
DUET Histogram Analysis
Remaining Mixtures
Initial Source Estimates
Stage Two Analysis
1) Pitch Estimation of Partial Signals
2) Create Harmonic Masks
Stage Two Reconstruction
3) Source Reconstruction from One-source and Two-source Frames
Pitch Estimates
Harmonic Masks
Remaining Mixtures
Refined Source Estimates
Stage Three Analysis
1) Determine Harmonic Amplitude Envelopes
Stage Three Reconstruction
2) Multi-source Reconstruction
3) Residual Reconstruction
Harmonic Ampl. Envelopes
Final Source Estimates
Reconstructed Source Waveforms

Figure 2: An Illustration of the three stages of the Active Source Estimation algorithm.

3.2 Mixing Parameter Estimation

In this section, we give a brief overview of mixing parameter estimation and demixing using DUET. A more thorough discussion of the DUET algorithm is provided in (Yilmaz 2004).

Let $\boldsymbol{X_1(\tau,\omega)}$ and $\boldsymbol{X_2(\tau,\omega)}$ represent the short-time Fourier transforms of two signal mixtures containing $\boldsymbol{N}$ source signals, $\boldsymbol{S_j(\tau,\omega)}$, recorded by two, omni-directional microphones.

$$X_1(\tau,\omega) = \sum_{j=1}^{N} S_j(\tau,\omega) \qquad (1)$$

$$X_2(\tau,\omega) = \sum_{j=1}^{N} a_j e^{-i\omega\delta_j} S_j(\tau,\omega) \qquad (2)$$

Here, $\boldsymbol{a_j}$ is the amplitude scaling coefficient and $\boldsymbol{\delta_j}$ is the time-shift between the two microphones for the *j*th 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 signals are approximately *window-disjoint orthogonal*, meaning most time-frequency frames in the mixture contain energy from no more than one source (Rickard 2002). Any frame that meets this requirement should match the amplitude scaling, $\boldsymbol{a_j}$, 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.

Amplitude scaling and time-shift values between the two mixture signals are first calculated for every time-frequency frame of the mixtures. This is accomplished by calculating the ratio $\boldsymbol{R(\tau,\omega)}$, as defined in (3). The amplitude scaling and time-shift are then calculated as shown in (4) and (5).

$$R(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} \qquad (3)$$

$$a(\tau,\omega) = | R(\tau,\omega) | \qquad (4)$$

$$\delta(\tau,\omega) = \frac{-1}{\omega} \angle R(\tau,\omega) \qquad (5)$$

The notation $|z|$ denotes the magnitude and the notation $\angle z$ denotes the phase angle of a complex number. In the case where either $X_1(\tau,\omega)$ or $X_2(\tau,\omega)$ is 0, $a(\tau,\omega)$ is set to 1 and $\delta(\tau,\omega)$ is set to 0.

The most common values for $a(\tau,\omega)$ and $\delta(\tau,\omega)$ can be found by creating a smoothed (using a rectangular kernel) two-dimensional weighted histogram in the space of amplitude scaling and time-shift values, $H(a,\delta)$. When the number of sources is known, DUET uses a k-means clustering algorithm (Theodoridis 2003) to find the N most prominent peaks in the smoothed histogram. The amplitude scaling and time-shift values associated with each peak in histogram $H(a,\delta)$ are assumed to be the mixing parameters corresponding to a particular source in the mix.

Once the mixing parameters for each source have been estimated, binary time-frequency masks are created for each source. Yilmaz and Rickard propose a maximum likelihood function to determine which source was most likely to have generated each time-frequency frame. In each frame, the binary mask of the source with the highest likelihood score is given a value of 1, while all other sources are given a 0. The binary masks are then multiplied by the mixture signal, $X_1(\tau,\omega)$, and transformed back to the time domain, resulting in source signals estimates (Yilmaz 2004).

In the rest of this work we assume that the amplitude scaling, a_j, and time-shift, δ_j, can be estimated correctly for each source j using DUET's parameter estimation. Alternate approaches that simulate binaural hearing in humans have been proposed to localize and separate source sounds with significant overlap or in reverberant environments (Roman 2003, Viste 2003a, Viste 2004), however in this work we assume 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 (Master 2003) 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. ASE 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.

Finding One-source Frames

To determine which frames in a stereo mixture correspond to a single source, define a function, ***Y***Y_j, for each pair of mixing parameters, (a_j, δ_j), associated with a source signal ***j***.

$$Y_j(\tau,\omega) = X_1(\tau,\omega) - \frac{1}{a_j} e^{i\omega\delta_j} X_2(\tau,\omega) \qquad (6)$$

If only one source is active in a given time-frequency frame, $Y_j(\tau,\omega)$ takes on one of two values. Equation (7) represents the predicted values of the $Y_j(\tau,\omega)$ functions, under the assumption that a single source, ***g*** (represented by the superscript ^{g}), was active.

$$\hat{Y}_j^g(\tau,\omega) = \begin{cases} 0 & if\ j = g \\ (1 - \frac{a_g}{a_j} e^{i\omega(\delta_j - \delta_g)}) X_1(\tau,\omega) & if\ j \neq g \end{cases} \qquad (7)$$

Equation (8) is a scoring function to compare the predicted values in $\hat{Y}_j^g(\tau,\omega)$ to the calculated $Y_j(\tau,\omega)$.

$$d(g,\tau,\omega) = \frac{\sum_{\forall j} \left| \hat{Y}_j^g(\tau,\omega) - Y_j(\tau,\omega) \right|}{\sum_{\forall j} \left| Y_j(\tau,\omega) \right|} \qquad (8)$$

As the function $d(g,\tau,\omega)$ approaches zero, the likelihood that source ***g*** was the only active source during the time-frequency frame (τ,ω) increases. A threshold value can then be used to determine which frames are one-source. These can be assigned directly to the estimate for source ***g*** (Master 2003).

Initial Source Reconstruction

We distribute the full energy from each one-source frame directly to the appropriate initial signal estimate, $\hat{S}_g$, as shown in equation (9).

$$\hat{S}_g(\tau,\omega) = \begin{cases} X_1(\tau,\omega) & \text{if } \big(d(g,\tau,\omega) < T\big) \wedge \\ & \quad (g = \underset{\forall j}{\arg\min}(d(j,\tau,\omega)) \\ 0 & \text{else} \end{cases} \qquad (9)$$

Here, ***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. Once an initial signal estimate is created for each source, the signals are analyzed and further source reconstruction is accomplished in stage two.

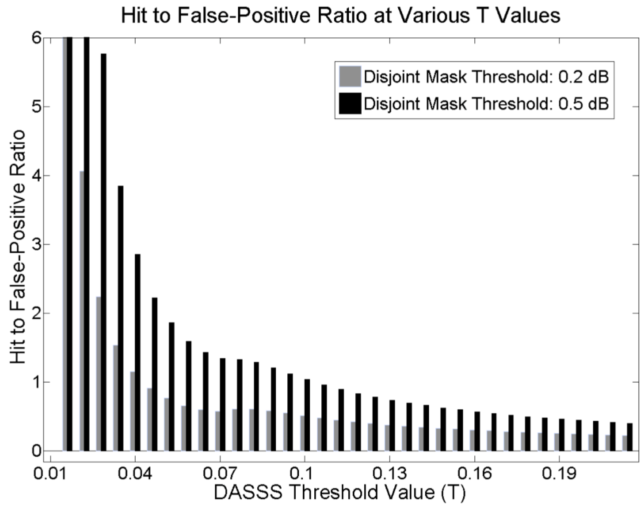

Figure 3: Hit to false-positive ratio in one-source frame identification at various DASSS threshold values, ***T***, over 400 instrument signals. Disjoint masks are determined according to (30) and (31) at two TdB values and (9) is used to identify one-source frames at 35 T values, ranging from 0.015 to 0.21).

When setting ***T***, we must both limit the error in $\boldsymbol{\hat{S}_g}$ and distribute enough frames to each source estimate so fundamental frequency estimation in stage two is possible. In Section 4.1, we discuss the calculation of *disjoint masks* for each source signal in a mix using equations (30) and (31). We determine an acceptable amount of error in the reconstructed source signals and use this value to identify the time-frequency frames of the mixture we will consider one-source frames. In setting the threshold value, ***T***, we attempt to identify as many of the correct one-source frames as possible, while limiting the number of frames that are falsely identified. Figure 3 shows the ratio of correctly identified frames (hits) and misidentified frames (false-positives) at 35 different values of ***T***. In a test on 100 four-instrument mixtures, we found that values below ***T*** = 0.15 provide an acceptable hit to false-positive ratio, while values above $\boldsymbol{T} = 0.1$ identify at least 50% of the true one-source frames. For the testing provided in section 4, we set $\boldsymbol{T} = 0.15$.

Figure 4 shows spectrograms of the initial source estimates, created during stage one of ASE, and the remaining mixture signals. The figure illustrates that many of the source's harmonics are correctly resolved during stage one, although with significant gaps or missing regions. The motivation for this approach comes from the fact that useful features can be estimated from these partial signals, and used to assist with the distribution of energy remaining.

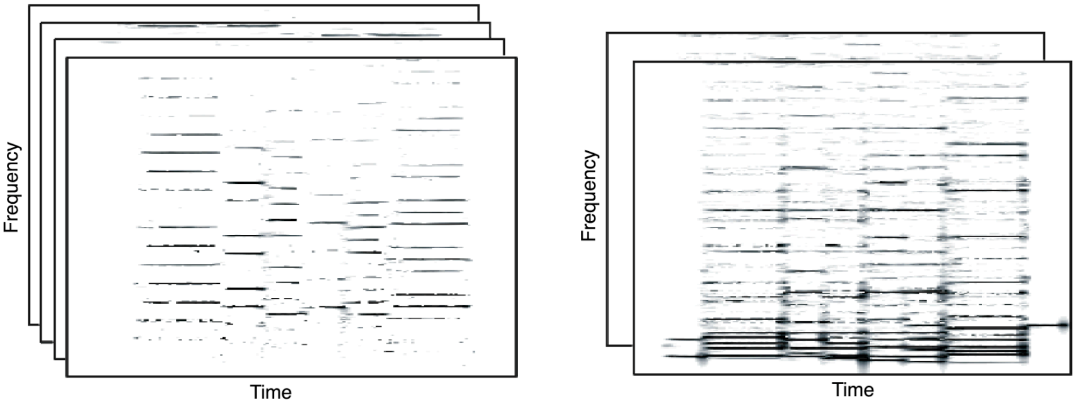

Figure 4: (left) Initial signal estimates of four string instruments, created during stage one of ASE. (right) Remaining mixture signals after stage one.

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*, which allow the determination of 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.

Determining the Active Source Count using Harmonic Masks

We denote the fundamental frequency of signal estimate $\hat{\boldsymbol{S}}_g$ for time window $\boldsymbol{\tau}$ as $\boldsymbol{F_g(\tau)}$ (shown in Figure 5). We determine fundamental frequency and harmonics-to-noise ratio, $\boldsymbol{HNR_g(\tau)}$**,** of each signal estimate using an autocorrelation-based technique described in (Boersma 1993).

To smooth spurious, short-lived variation in the $\boldsymbol{F_g}$ estimates, any change in $\boldsymbol{F_g}$ over 6% (roughly a *semitone*) that lasts less than 60ms is changed to match the fundamental frequency estimate in the frame prior to the transition. 60ms was chosen because it is nearly a sixteenth note at 120bpm (beats per minute) and is the shortest event we expect to process. This parameter can be altered for processing music in which more rapid note transitions are present.

We have low confidence in $\boldsymbol{F_g}$ estimates for times $\boldsymbol{\tau}$ with low harmonics-to-noise ratio ($\boldsymbol{HNR_g(\tau)} < \boldsymbol{H_{min}}$). For these times, we set the fundamental frequency estimate to be equal to that of the most correlated neighbor estimate. Let $\boldsymbol{\hat{S}_g(\tau_n,\omega)}$ indicate the vector of values for signal estimate $\boldsymbol{\hat{S}_g}$ at all frequencies of analysis at time $\boldsymbol{\tau_n}$. For each low-confidence estimate, we measure cross-correlation between $\boldsymbol{\hat{S}_g(\tau_n,\omega)}$ and the immediately preceding step, $\boldsymbol{\hat{S}_g(\tau_{n-1},\omega)}$, and between $\boldsymbol{\hat{S}_g(\tau_n,\omega)}$ and the next time step with a confident fundamental frequency estimate, $\boldsymbol{\hat{S}_g(\tau_{n+d},\omega)}$. We replace $\boldsymbol{F_g(\tau_n)}$ with the value from the time-step (either $\boldsymbol{F_g(\tau_{n-1})}$ or $\boldsymbol{F_g(\tau_{n+d})}$) with the greatest cross-correlation.

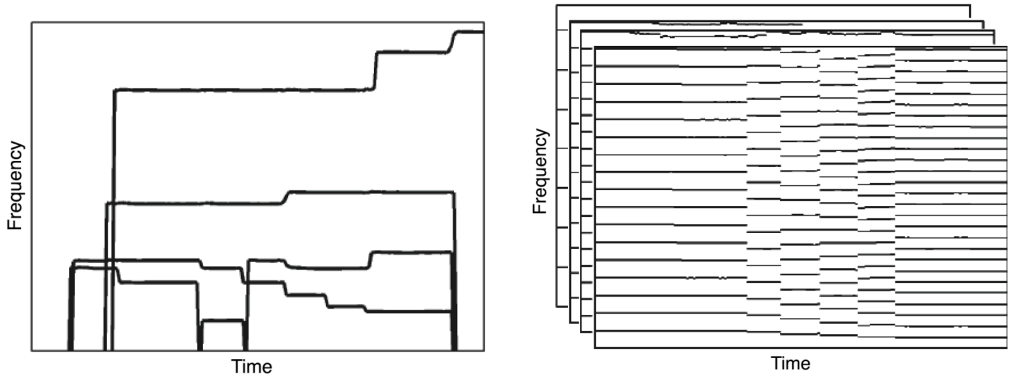

Figure 5: (left) Fundamental frequency estimates of all four instruments, created during stage two. (right) Harmonic mask of each instrument, created during stage two.

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*, $\boldsymbol{M_g(\tau,\omega)}$, or binary time-frequency mask for each source (shown in Figure 5). 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.

$$M_g(\tau,\omega) = \begin{cases} 1 \text{ if } (\exists k \text{ such that } |kF_g(\tau) - \omega| < \Delta_\omega) \\ 0 \text{ else} \end{cases} \tag{10}$$

Here, $\boldsymbol{k}$ is an integer and Δ_ω is the maximal allowed difference in frequency from the k^{th} harmonic and is set 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 multi-source frames. We do this by summing the harmonic masks for all the sources to create the *active source count* for each frame, ***C*(τ,ω).**

$$C(\tau,\omega) = \sum_{\forall g} M_g(\tau,\omega) \qquad (11)$$

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 (9) to accept enough time-frequency frames to estimate ***F*$_g$(τ)** for each source. We remove energy that might have been mistakenly given to each source by taking,

$$\hat{S}_g^{two}(\tau,\omega) = \hat{S}_g^{one}(\tau,\omega) M_g(\tau,\omega) \qquad (12)$$

In (12) and (13) we add superscripts to the source estimate notation to clarify which stage of source reconstruction is specified. Thus, equation (12) eliminates time-frequency frames from the initial source estimates that are not near the predicted harmonics of that source. We then add energy to the estimates in any one-source frames identified by the active source count that were not identified by DASSS.

$$\begin{aligned}\hat{S}_g^{two}(\tau,\omega) = X_1(\tau,\omega)\ \textbf{if}\ &(C(\tau,\omega) = M_g(\tau,\omega) = 1) \\ &\wedge (\hat{S}_g^{one}(\tau,\omega) = 0)\end{aligned} \qquad (13)$$

In time-frequency frames where the source count ***C*(τ,ω)** = 2, we presume the frame has two active sources and use the system of equations in (14) and (15) to solve for the source values.

$$X_1(\tau,\omega) = \hat{S}_g(\tau,\omega) + \hat{S}_h(\tau,\omega) \qquad (14)$$

$$X_2(\tau,\omega) = a_g e^{-i\omega\delta_g} \hat{S}_g(\tau,\omega) + a_h e^{-i\omega\delta_h} \hat{S}_h(\tau,\omega) \qquad (15)$$

We can solve for source ***g*** as in (16) and use (14) to solve for source ***h***.

$$\hat{S}_g(\tau,\omega) = \frac{X_2(\tau,\omega) - a_h e^{-i\omega\delta_h} X_1(\tau,\omega)}{a_g e^{-i\omega\delta_g} - a_h e^{-i\omega\delta_h}} \qquad (16)$$

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 ***C*(τ,ω)** > 2 are distributed in stage three. Figure 6 shows the refined source estimates and the mixture signals still remaining. Notice the changes in the remaining mixtures between Figure 4 and Figure 6. The mixtures remaining after stage one still contained many instrument harmonics, especially in low frequency regions. The mixtures after stage two, although still containing some harmonic energy, are primarily made up of *attack noise*, or non-harmonic energy. Stage three will first distribute any remaining harmonic energy before distributing the non-harmonic energy.

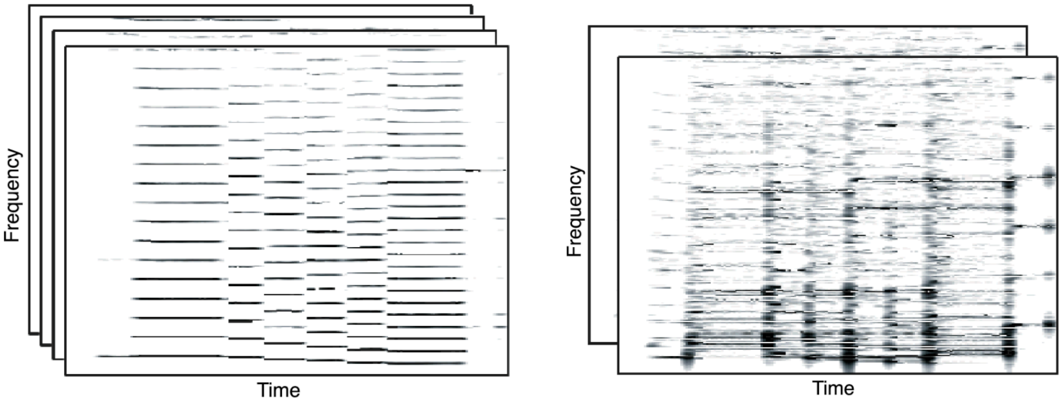

Figure 6: (left) Refined signal estimates of each instrument, created during stage two. (right) Remaining mixtures after stage two.

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 multi-source mixture frame, using the reconstructed source signals created during stages one and two as a guide.

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

Let $[\tau_s,\tau_{s+n}]$ be a sequence of multi-source frames at frequency of analysis ω**.** 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 (Risset 1982). In source separation, (Anemüller 2000, Viste 2003b) make this assumption and CASA algorithms commonly use correlated amplitude modulation as a grouping mechanism (Brown 2005, Hu 2005, Ellis 1996).

If *harmonic amplitude envelopes*, or the amplitude modulation trend of each source's harmonics, can be determined, we can use them as a guide for the amplitude modulation of each source during the sequence of multi-source frames. If we also assume that each source's phase progresses linearly over the sequence, we have a means of estimating how each source's energy changes during the sequence $[\tau_s,\tau_{s+n}]$.

If we can then estimate the value of $\hat{S}_g(\tau_s,\omega)$ for each active source, the linear phase change assumption and harmonic amplitude envelopes can be used to determine $\hat{S}_g(\tau_{s+1},\omega)$ through $\hat{S}_g(\tau_{s+n},\omega)$.

We first show the method used for determining *harmonic amplitude envelopes*, and then proceed with a discussion of how to estimate $\hat{S}_g(\tau_s,\omega)$, the first complex value of each active source in the sequence of multi-source frames.

Determining Harmonic Amplitude Envelopes

To calculate the overall harmonic amplitude envelope for source ***g***, we first find the amplitude envelope of each harmonic in the signal estimate for ***g***, using (17). Here, ***k*** denotes the harmonic number. We include time-frequency frames in the estimate of $A_g(\tau,k)$ if the center frequency of the frame is both within Δ_ω (as defined in (10)) of the harmonic frequency, and the source signal estimate from stage two contains energy in that frame.

$$A_g(\tau,k) = \underset{\forall\omega\in\Gamma(k)}{\text{mean}}(|\hat{S}_g(\tau,\omega)|) \qquad (17)$$

$$\begin{gathered}\omega\in\Gamma(k) \text{ if } (|\omega - kF_g(\tau)| < \Delta_\omega \wedge \\ \hat{S}_g(\tau,\omega) > 0)\end{gathered} \qquad (18)$$

Equation (19) normalizes each amplitude envelope so that each harmonic contributes equally to the overall amplitude envelope.

$$\widetilde{A}_g(\tau,k) = \frac{A_g(\tau,k)}{\max\limits_{\forall\tau}(A_g(\tau,k))} \qquad (19)$$

Equation (20) is used to determine the overall harmonic amplitude envelope, which we denote, $\boldsymbol{H_g(\tau)}$. 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 Equation (21). Here, $\boldsymbol{L}$ specifies a time window over which the signal energy is calculated. We include the amplitude scaling in (20) so the relative strength of each source's harmonic amplitude envelope corresponds to the overall loudness of each source during the time window $\boldsymbol{L}$.

$$H_g(\tau) = \operatorname*{mean}_{\forall k}(\widetilde{A}_g(\tau,k))E_g(\tau) \qquad (20)$$

$$E_g(\tau) = \sum_{\lambda=-L/2}^{L/2} \sum_{\forall\omega} |\hat{S}_g(\tau+\lambda,\omega)|^2 \qquad (21)$$

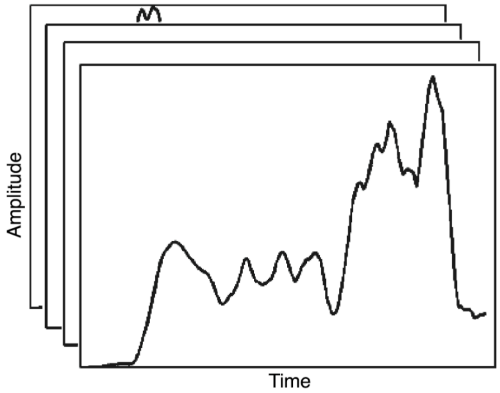

Figure 7: Harmonic amplitude envelopes for each instrument, created during stage three.

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

If, for each source $\boldsymbol{g}$, the first value in the sequence, $\boldsymbol{\hat{S}_g(\tau_s,\omega)}$, can be estimated, then (22) and (23) can be used to estimate the values of the sources in the remaining multi-source frames, $[\boldsymbol{\tau_{s+1}},\boldsymbol{\tau_{s+n}}]$. Here, we set $\boldsymbol{\tau_a} = \boldsymbol{\tau_s}$ and $\boldsymbol{\tau_b} \in [\boldsymbol{\tau_{s+1}},\boldsymbol{\tau_{s+n}}]$.

$$\left|\hat{S}_g(\tau_1,\omega)\right| = \frac{H_g(\tau_1)}{H_g(\tau_0)}\left|\hat{S}_g(\tau_0,\omega)\right| \qquad (22)$$

$$\angle\hat{S}_g(\tau_b,\omega) = \mathrm{mod}\left(\angle\hat{S}_g(\tau_a,\omega) + (\tau_b - \tau_a)\omega, 2\pi\right) \qquad (23)$$

Estimation from a prior example

The frame immediately before the start of the sequence of multi-source frames in question is (τ_{s-1},ω). If a source estimate was already given energy in this frame during stage one or two (if $|\hat{S}_g(\tau_{s-1},\omega)| > 0$), we can use $\hat{S}_g(\tau_{s-1},\omega)$ to estimate $\hat{S}_g(\tau_s,\omega)$ using (22) and (23) by setting $\tau_a = \tau_{s-1}$ and $\tau_b = \tau_s$.

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

We denote a source whose amplitude was estimated using (22) as h, and now estimate the amplitude of any remaining active source in frame τ_s.

$$\left|\hat{S}_g(\tau_s,\omega)\right| = \frac{H_g(\tau_s)}{H_h(\tau_s)}\left|\hat{S}_h(\tau_s,\omega)\right| \qquad (24)$$

We set the phase of sources whose amplitudes are derived using (24) to a value of 0.

Estimation without a prior example

If after stage two, $|\hat{S}_g(\tau_{s-1},\omega)| = 0$ for all sources, we must use an alternate method of estimating $\hat{S}_g(\tau_s,\omega)$. 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 $[\tau_s,\tau_{s+n}]$ 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 τ_{MaxInt}. With this assumption, equation (12), altered for the N active source case in frame (τ_{MaxInt},ω), yields (25), where Φ is the set of active sources in the multi-source sequence, $[\tau_s,\tau_{s+n}]$.

$$\left|X_1(\tau_{MaxInt},\omega)\right| = \sum_{\forall g \in \Phi}\left|\hat{S}_g(\tau_{MaxInt},\omega)\right| \qquad (25)$$

The amplitude of any active source ***g*** can then be determined using (26).

$$\left|\hat{S}_g(\tau_{MaxInt},\omega)\right| = \left|X_1(\tau_{MaxInt},\omega)\right| \frac{H_g(\tau_{MaxInt})}{\sum_{\forall h \in \Phi} H_h(\tau_{MaxInt})} \qquad (26)$$

To find $|\hat{S}_g(\tau_s,\omega)|$ from $|\hat{S}_g(\tau_{MaxInt},\omega)|$ we apply (22) with $\tau_a = \tau_{MaxInt}$ and $\tau_b = \tau_s$. We set the phase values of each active source during the first frame, $\angle\hat{S}_g(\tau_s,\omega)$, to a default value of 0.

We now apply (22) and (23) to determine $\hat{S}_g(\tau_{s+1},\omega)$ through $\hat{S}_g(\tau_{s+n},\omega)$ from $\hat{S}_g(\tau_s,\omega)$, and complete this process for each sequence of multi-source frames determined by the source count, $C(\tau,\omega)$.

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 non-harmonic frequencies due to factors such as excitation noise (Risset 1982). The non-harmonic 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 (9),

$$\hat{S}_g(\tau,\omega) = \begin{cases} X_1(\tau,\omega) \text{ if } g = \arg\min\limits_{\forall j}(d(j,\tau,\omega)) \\ 0 \quad \text{else} \end{cases} \qquad (27)$$

Once the residual energy has been distributed, each source estimate, $\hat{S}_g(\tau,\omega)$, is transformed back into the time domain using the overlap-add technique (Openheim 1989). The result is a time domain waveform of each reconstructed source signal. Figure 8 shows the final estimate of the source shown in Figures 4 and 6 in both the time-frequency and time domains.

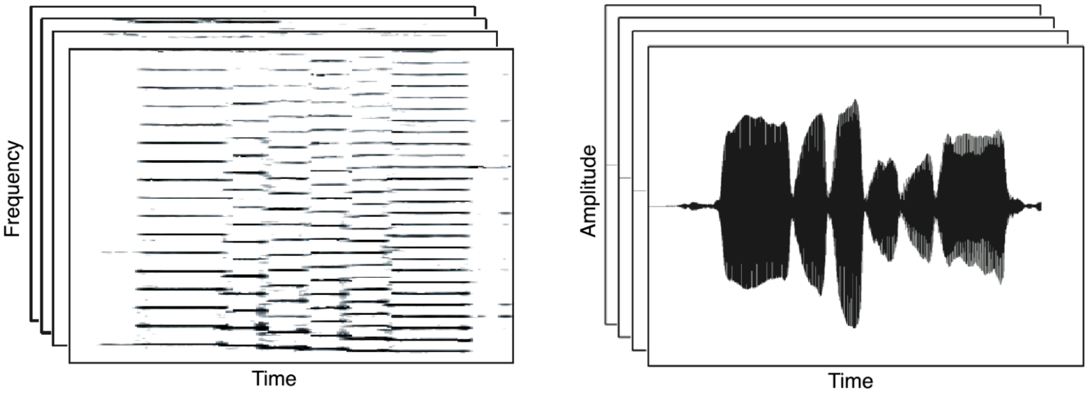

Figure 8: Final source signal estimates for each instrument in the time-frequency (left) and time (right) domains.

4. EXPERIMENTAL RESULTS

In this section we compare the performance of the ASE and DUET algorithms on mixtures of three and four harmonic instruments. We use two data sets, one consisting of short mixtures containing a single long-tone from each instrument, and the other consisting of short excerpts from J.S. Bach chorale harmonizations. We chose to compare the performance of ASE to DUET because ASE is designed with the same mixture models and constraints, making it a natural extension of time-frequency masking techniques such as DUET. We now describe the methods used in the creation of signal mixtures and analysis of algorithm performance.

4.1 Three and Four Instrument Mixtures

The instrument recordings used in the testing mixtures are individual long-tones played by alto flute, alto and soprano saxophone, bassoon, B-flat and E-flat clarinet, French horn, oboe, trombone and trumpet, all taken from the University of Iowa musical instrument database (Fritts).

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 (Yilmaz 2004). Instruments were placed in a semi-circle 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. Each mixture was separated into reconstructed source signals by the ASE and DUET algorithms, using a window length of 46ms and step size of 6ms for STFT processing. Extracted sources were then compared to the original sources using the *signal-to-distortion ratio* (SDR) described in (Gribonval 2003).

$$SDR = 10\log_{10}\left(\frac{\left|\langle \hat{s}, s\rangle\right|^2}{\left|\langle \hat{s}, \hat{s}\rangle\right|^2 - \left|\langle \hat{s}, s\rangle\right|^2}\right) \qquad (28)$$

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

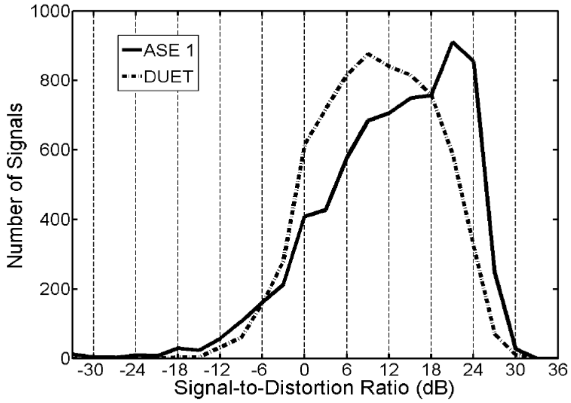

Figure 9: Histogram of ASE 1 and DUET SDR performance over all signals.

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 (see Figure 9), so we performed a nonparametric sign test over all mixtures and found the median performance to be significantly different between ASE 1, ASE 2 and DUET, with $\boldsymbol{p} < 10^{-50}$ in all three comparisons.

	ASE 1	ASE 2	DUET
All Mixtures	13.77 dB	12.26 dB	10.22 dB
3-Instrument Mixtures	18.63 dB	17.57 dB	14.12 dB
4-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

Table 1: Median Signal-to-Distortion Ratio of the ASE and DUET algorithms on 1000 3-instrument mixtures and 1000 4-instrument mixtures (7000 signals). Also shows median performance on 3 and 4-instrument mixtures containing specific musical intervals: unison (2383 signals), octave (366 signals), perfect fifth (1395 signals) and perfect fourth (1812 signals).

A primary goal of the ASE system was to reduce the reliance of time-frequency masking techniques on nearly disjoint source signals. Since ASE relies on fundamental frequency estimation of partial signals, created from only the disjoint (non-overlapping) time-frequency frames of each signal, we expect source reconstruction to deteriorate as the amount of interference from other source signals increases.

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 a threshold value, $\boldsymbol{T_{dB}}$, 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 $\boldsymbol{T_{dB}}$ due to the other sources in the mixture, while a value of 1 reflects that interference from other sources is restricted to less than $\boldsymbol{T_{dB}}$ in all time-frequency frames.

We define the *disjoint energy*, $\boldsymbol{DE_g}$, of each source signal as,

$$DE_g = \frac{\sum_{\forall\tau}\sum_{\forall\omega} DM_g(\tau,\omega)\,|\,S_g(\tau,\omega)\,|}{\sum_{\forall\tau}\sum_{\forall\omega}|\,S_g(\tau,\omega)\,|} \tag{29}$$

The calculation of disjoint energy relies on (30) and (31).

$$MSR_g(\tau,\omega) = 20\log_{10}\left(\frac{\sum_{\forall j}|\,S_j(\tau,\omega)\,|}{|\,S_g(\tau,\omega)\,|}\right) \tag{30}$$

$$DM_g(\tau,\omega) = \begin{cases} 1 \text{ if } MSR_g(\tau,\omega) < T_{dB} \\ 0 \text{ else} \end{cases} \tag{31}$$

Equation (30) defines the *mixture-to-signal ratio*, $\boldsymbol{MSR_g(\tau,\omega)}$, and (31) defines *disjoint masks*, $\boldsymbol{DM_g(\tau,\omega)}$, for each source. The ***MSR*** represents the amount of possible amplitude difference we could see in the mixture, which gives us a guide of how much the other source signals interfere with a particular source. By defining an allowable decibel error, $\boldsymbol{T_{dB}}$, we create disjoint masks that isolate the time-frequency frames of each source signal that are relatively unaffected by energy from other sources. For the data presented here, we set $\boldsymbol{T_{dB}}$ to 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.

Figure 10 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 $\boldsymbol{DE} \in (0, 0.2), (0.2, 0.4), (0.4, 0.6), (0.6, 0.8)$ and $(0.8, 1)$. We show box-plots 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.

If we examine the five cases in Figure 10, we can see that the performance improvement provided by ASE is moderate for signals with ***DE*** greater than 0.8. This is not surprising considering DUET reconstruction is quite good for these signals (median SDR is over 20 dB). As the disjoint energy in a source signal decreases, the improvement provided by ASE becomes more substantial, 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 0.8 (over 4 dB), or signals with roughly half of their energy obstructed. As a source signal's ***DE*** falls below 0.2, the performance by both algorithms is poor.

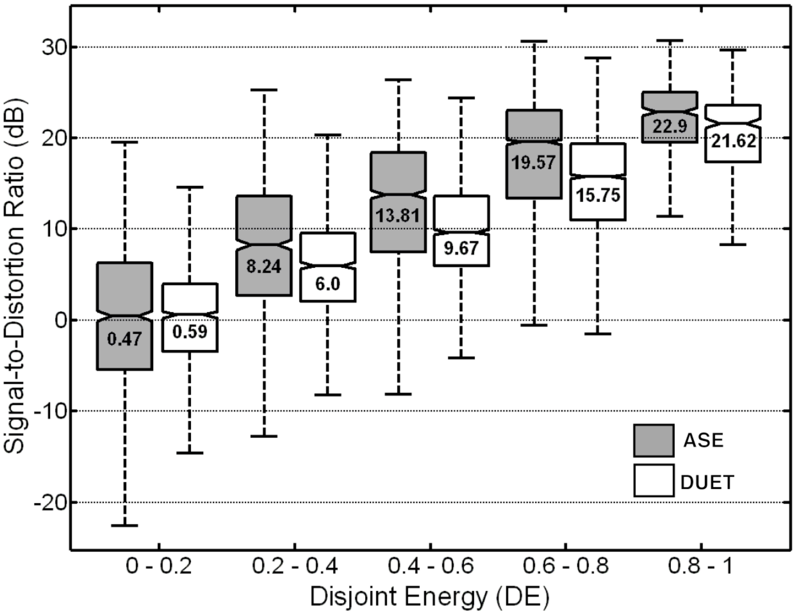

Figure 10: ASE 1 and DUET SDR performance over five groups of signals. Signals are grouped according to disjoint energy, ***DE***. Median performance is shown in the lower half of each box.

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 (non-overlapping) 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.

Figure 11 represents the difference in ***SDR*** performance between ASE and DUET of as a function of ***DE***, without quantizing the data into five groups. For each signal, the performance of DUET was subtracted from ASE, so a positive value in Figure 11 shows better performance by ASE. The heavy dashed line represents the trend of the difference in performance, which was calculated as the line that minimizes that least square error.

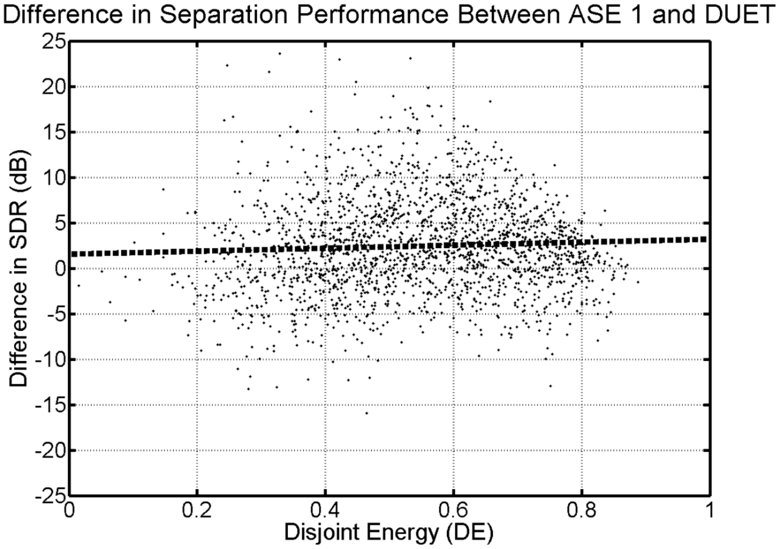

Figure 11: Difference in SDR performance between ASE 1 and DUET (ASE 1 – DUET) over all 3 and 4-instrument mixtures as a function of Disjoint Energy.

4.2 Incorporating a Time-Aligned Musical Score

ASE was also tested on 100 four-part chorale harmonizations by J.S. Bach in order to determine the utility of incorporating pitch information from a time-aligned musical score. For each harmonization, we randomly chose a four second segment, typically equating to about one or two measures in the music. Choosing small sections of the harmonizations allows us to better understand the relationship between the amount of source signal overlap in the mixtures and source separation performance. For each segment of the harmonization chosen, we created three MIDI versions. The first version was an unaltered representation of the selected segment of the harmonization. We call this the *original score*.

From each original score, we created the second MIDI version by randomly altering the tempo of each piece between 71% and 140% of the original tempo, with the average deviation being roughly 20%. This version was used to generate the audio mixture, and we call this the *ideal score*. Although a typical interpretive performance of a piece of music would likely include tempo variation throughout the duration of the piece, our scored segments were only a measure or two long, so we felt that a simple tempo scaling was a reasonable simulation of a performance of the harmonization segment.

For each notated instrument part in the ideal score we created an audio file using recorded samples of violin (soprano and alto part), viola (tenor) and cello (bass). The samples used were from a commercial instrument sample library, *Xsample Professional Sound Libraries, Volume 41: Solo Strings*. These individual audio recordings (one for each instrument part in the score) were then combined to create a stereo audio mixture of each chorale harmonization. We created mixtures in this way in order to measure the difference between the ideal (the pre-mix individual signals) and the source estimates extracted from each mixture.

We then performed score following on each audio mixture, aligning the original score to the audio mixture, as in (Hu 2003). The output of the score follower was a MIDI file that had been time-altered to match the timing of the audio mixture. This is the *aligned score*.

4.3 Performance Results when Incorporating a Musical Score

For each audio mixture we performed source separation four times: once with no score (the standard ASE algorithm), once with the ideal score, once with the aligned score, and once with the original score. For this experiment, we used a window length of 186 ms and a 163 ms overlap between time frames in the time-frequency analysis of the mixture.

We found that using knowledge of the score greatly improved the performance of the source separation algorithm. Without score knowledge, the fundamental frequency estimation in stage 2 of ASE was accurate (within half a semitone) in an average of 69.4% of a source signal's time frames. Using the aligned and refined scores increased this accuracy to 92.9%. The increased accuracy of the fundamental frequency estimates resulted in improved separation performance in 78.25% of the separated signals. The SDR improvement between the median *blind* and median *aligned score* performance was 1.7 dB.

Figure 12 shows notched box-plots of the SDR over all trials for the four score knowledge scenarios. Each box represents the performance on 400 signals, four for each chorale harmonization. The notches in each box show the 95% confidence interval around the median. Since the notches in the box-plot for the blind case and the aligned score do not overlap, we conclude, with 95% confidence, that use of the aligned score provides significant performance improvement.

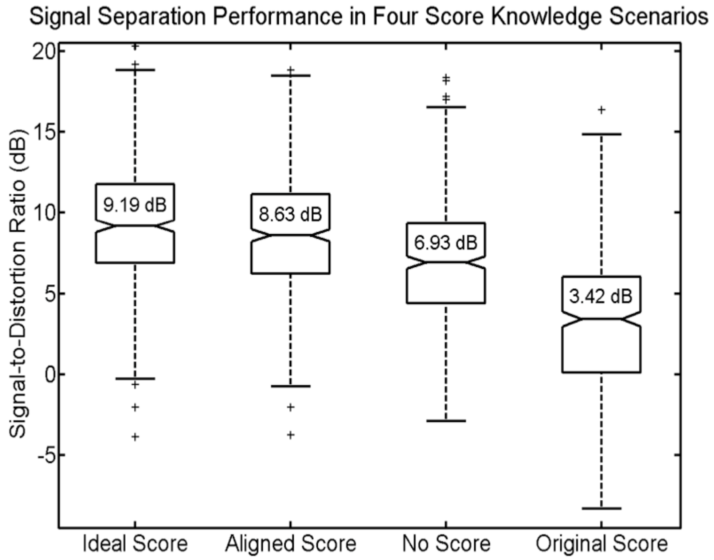

Figure 12: Performance results over all mixtures, compared between score knowledge conditions.

While knowledge of the score can improve the algorithm's performance, a misaligned score can actually degrade separation. In comparing the blind algorithm performance to the performance with the original score (the non-aligned score), the median SDR decreased by 3.51 dB with 79.25% of the cases performing worse when the algorithm had knowledge of the misaligned score. This result emphasizes the necessity of score alignment if one is to incorporate score knowledge into a signal separation algorithm.

5. CONCLUSIONS AND FUTURE WORK

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

ASE improves 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 don't 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 tine-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 numerous obstacles to overcome before robust, blind separation of real-world musical mixtures is a reality, we believe the performance of our approach on anechoic mixtures provides promising evidence that we are nearing a tool that can deal with situations encountered in real recordings.

6. REFERENCES

Aarabi, P., Shi, G., Jahromi, O. "Robust speech separation using time-frequency masking," *Proceedings of the 2003 IEEE Conference on Multimedia and Expo*, Baltimore, Maryland, July 2003, pp. 741-744.

Anemüller, J., Kollmeier, B. "Amplitude modulation decorrelation for convolutive blind source separation," *Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, 2000, pp. 215-220.

Balan, R., Rosca, J. "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," *Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, 2000.

Balan R., Rosca, J. "Source separation using sparse discrete prior models," *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations*, Rennes, France, November 16-18 2005.

Boersma, P. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, Vol. 17, 1993, pp. 97-110.

Bregman, A. *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Massachusetts, 1990.

Brown, G.J., Wang, D. "Separation of Speech by Computational Auditory Scene Analysis", *Speech Enhancement*, J. Benesty, S. Makino and J. Chen (Eds.), Springer, NY, 2005, pp. 371-402.

Ellis, D. "Prediction-driven computational auditory scene analysis", PhD Dissertation, Massachusetts Institute of Technology, Media Laboratory, 1996.

Ellis, D., Weiss, R. "Model-Based Monaural Source Separation Using a Vector-Quantized Phase-Vocoder Representation," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* Toulouse, France, May 2006, pp. 957-960.

Every, M., Szymanski, J. "A spectral-filtering approach to music signal separation", *Proceedings of the 7th International Conference on Digital Audio Effects*, Naples, Italy, Oct. 5-8 2004, pp. 197-200.

Fritts, L. University of Iowa Musical Instrument Samples. Available at http://theremin.music.uiowa.edu.

Gribonval, R., Benaroya, L., Vincent, E., Fevotte, C. "Proposals for Performance Measurement in Source Separation", *Proceedings of the 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 2003.

Hu, G., Wang, D. "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, Vol. 15, No. 5, September 2004, pp. 1135-1150.

Hu, N., Dannenberg, R., Tzanetakis, G. "Polyphonic Audio Matching and Alignment for Music Retrieval," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 19-22 2003.

Hyvarinen, A., Oja, E. "Independent Component Analysis: Algorithms and Applications," *Neural Networks,* 13(4-5): 411-430, 2000.

Jourjine, A., Rickard, S., Yilmaz, O. "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

Klapuri, Anssi P. "Multipitch estimation and sound separation by the spectral smoothness principle," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, 2001.

Lee, T.W., Bell, A.J., Orglmeister, R. "Blind source separation of real world signals," *Proceedings of the IEEE International Conference on Neural Networks*, Houston, Texas, 1997.

Lee, T.W., Bell, A.J., Lambert, R.H. "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems*, 1997.

Master, A.S. "Sound source separation of n sources from stereo signals via fitting to n models each lacking one source," Technical Report, CCRMA, Stanford University, 2003.

O'Grady, P.D., Pearlmutter, B.A., Rickard, S.T. "Survey of Sparse and Non-Sparse Methods in Source Separation," *International Journal of Imaging Systems and Technology*, Vol. 15, (1), 2005, pp. 18-33.

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

Parra, L.C., Spence, C. D. "Separation of non-stationary natural signals", *Independent Component Analysis, Principles and Practice*. Cambridge University Press, 2001, pp. 135-157.

Reyes-Gomez, M.J., Ellis, D., Jojic, N. "Multiband Audio Modeling for Single-Channel Acoustic Source Separation," *Proceedings of the International Conference on Audio, Speech and Signal processing*, Montreal, Canada, May, 2004.

Rickard, S., Yilmaz, O. “On the Approximate W-Disjoint Orthogonality of Speech”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, May 2002, pp. 529-532.

Risset, J.C., Wessel, D. “Exploration of timbre by analysis and synthesis”, *The Psychology of Music*, Academic Press, NY, 1982, pp. 26-58.

Roman, N., Wang, D., Brown, G.J. “Speech segregation based on sound localization*”, Journal of the Acoustical Society of America,* vol. 114, 2003, pp. 2236-2252.

Rosenthal, D.F., Okuno, H.G. *Computational Auditory Scene Analysis.* Lawrence Erinbaum Associates, 1998.

Srinivasan, S., Wang, D. “A schema-based model for phonemic restoration,” *Speech Communication*, 45, 2005, pp. 63-87.

Stone, J.V. *Independent Component Analysis: A Tutorial Introduction.* MIT Press, Cambridge, Massachussets, 2004.

Theodoridis, S., Koutroubas, K. *Pattern Recognition*, Academic Press, San Diego, 2003.

Vincent, E., Rodet, X. “Underdetermined Source Separation with Structured Source Priors,” *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation* , Granada, Spain, Sep. 22-24 2004.

Vincent, E. “Musical Source Separation Using Time-Frequency Priors,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, (1), 2006, pp. 91-98.

Virtanen, T., Klapuri, A. “Separation of harmonic sounds using multipitch analysis and iterative parameter estimation,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Platz, New York, 2001, pp. 83-86.

Virtanen, T., Klapuri, A. “Separation of Harmonic Sounds using Linear Models for the Overtone Series”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, May 13-17 2002, pp. 1757-1760.

Viste, H., Evangelista, G. “On the use of spatial cues to improve binaural source separation,” *Proceedings of the 6th International Conference on Digital Audio Effects*, London, UK, September 8-11 2003.

Viste, H. and Evangelista, G. “Separation of harmonic instruments with overlapping partials in multi-channel mixtures,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Platz, New York, Oct. 19-22, 2003, pp. 25-28.

Viste, H. and Evangelista, G. “Binaural Source Localization”, Proc. of the 7th Int. Conf. on Dig. Audio Effects, Oct. 5-8, 2004, pp. 145-150.

G.H. Wakefield. “Mathematical Representation of Joint Time-Chroma Distributions,” *The International Symposium on Optical Sci., Eng., and Instr.*, Denver, Colorado, 1999.

Yilmaz, O., Rickard, S. (2004). “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Transactions on Signal Processing*, Vol. 52, (7), 2004, pp. 1830-1847.