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

The goal in *blind source separation (BSS)* is to determine the original source signals, given mixtures of these sources. Recent work in blind source separation (the DUET and DASSS methods) allows source reconstruction from mixtures when the time-frequency components of the sources rarely overlap in the mixtures. While speech mixtures often approximate this constraint, music mixtures rarely do, requiring new approaches. We introduce a method to assign energy to source estimates from music mixtures that builds on partial signal estimates provided by the existing DUET and DASSS algorithms. We estimate *harmonic masks* from the initial source estimates and assign energy from the mixed signal to sources by estimating the number of active sources in each time-frequency frame and allowing source estimates to share mixture energy from a single frame. This allows dealing with mixtures that contain time-frequency frames in which multiple harmonic sources are active without requiring knowledge of source characteristics.

Keywords: Source separation, scene analysis, music

*A version of this paper was submitted to the *EURASIP Journal on Applied Signal Processing*, on December 1st, 2005.

1. INTRODUCTION

Collections of music recordings (such as .mp3 or .wav files) are currently indexed by such features as title, composer, and performer. People, however, often wish to access and use music documents based on aspects of their musical content. To satisfy these information needs, researchers in music information retrieval must create systems that can find perceptually relevant structure in the audio signal. Often, this requires the ability to separate audio mixtures into source signals. A tool able to perform source separation would be of use to recording engineers, composers, multimedia producers and researchers. Tasks that accurate source separation could facilitate include automated transcription, vocalist and instrument identification, suppression or amplification of an instrument within a mixture, or melodic comparison of different recordings.

Researchers in *computational auditory scene analysis* (CASA) [7,8,9,19,20] are interested in the source separation problem as it relates to human auditory perception. They develop methods to parse audio mixtures as a means of informing and testing perceptual hypotheses [8,9]. Unfortunately, our understanding of human auditory perception is not currently sufficient to allow robust automated source separation based only on perceptual models, and often those seeking source separation for other purposes take approaches that are not perceptually motivated.

In a musical context, Goto developed a system to detect melody and bass lines in polyphonic music by using continuity measures on fundamental frequency estimates in limited frequency bands [11]. This work concentrated on monaural signals, but much information can be gained by looking at localization cues available in multi-channel signals, such as stereo recordings.

In the multi-channel context, a single mixture corresponds to a monaural recording of one or more sources, such as the left channel in a stereo recording, and a source is an individual musical instrument or voice. Much recent work in parsing multi-channel audio mixtures into source signals has focused on *blind source separation* (BSS) [1,2,4,12-16,24,25]. Blind source separation is so named because the methods applied assume little or no knowledge of the properties of the source signals composing the mixture, allowing for a wide range of applications.

Independent component analysis (ICA) [2,15,22] can be used to solve the BSS problem when the number of mixtures $\boldsymbol{m}$, equals or exceeds the number of source signals, $\boldsymbol{n}$. However, when $\boldsymbol{m} < \boldsymbol{n}$ (fewer mixtures than sources are available), the problem is considered the degenerate case of BSS. Since millions of audio recordings exist in a stereo format (two-mixtures), but typically consist of more than two source signals, it should be clear why solving the degenerate case of BSS is of considerable interest to researchers.

If the sources are disjoint in some way, one can solve the degenerate case [17]. For example, if at most one source is active at any given time, the sources are time-disjoint and can be separated successfully [13]. Numerous researchers [1,14,25] have proposed demixing (source separation) solutions for the case where only one of the sources contributes energy to the mixture for any given combination of time and frequency.

The DUET [25] method is a time-frequency masking technique that has been applied with some success to mixtures where the sources are approximately time-frequency disjoint and has been shown to work well on anechoic mixtures of speech signals.

The delay and scale subtraction scoring (DASSS) [14] method extends DUET by estimating which time-frequency frames in a mixture contain energy from a single source. Source reconstruction can then be done from only the single-source frames. This works well when the preponderance of frames in a mixture corresponds to a single source. Reconstruction is increasingly poor as the proportion of single-source frames decreases.

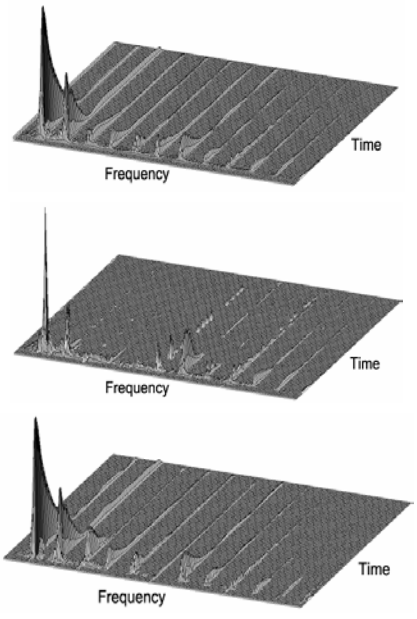

Figure 1: (top) The spectrogram of a piano playing a C (262Hz). (middle) The spectrogram of the DUET source estimate of the same piano tone when extracted from a mixture with a saxophone playing G and French horn playing C. (bottom) The spectrogram of ASE source estimate of the same piano tone extracted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

For music mixtures, the assumption that (approximately) one source is active for the majority of time-frequency regions is often invalid. Tonal music styles, such as Jazz, Rock, Pop, and Classical music, make extensive use of *consonant intervals* (such as unisons, octaves and perfect fifths) between pairs of harmonic sound sources. When two harmonic sources form a consonant interval, their fundamental frequencies are related by a rational ratio that results in significant overlap between the high-energy frequency bands (the *harmonics* or *partials*) of one source and those of another. Consonant intervals result in large numbers of time-frequency frames with energy from multiple sources. Thus, source reconstructions based on binary time-frequency masking methods, such as DASSS and DUET, are often incomplete or inaccurate.

Figure 1 provides an example of this. The top image shows the spectrogram of a piano playing middle C (262 Hz). The middle image shows the spectrogram of the DUET source estimate of the same piano tone when extracted from a mixture containing the piano, a saxophone playing G, and a French horn in unison with the piano on C. Note the significant portions of the signal that are missing in the estimate. The lower image is the source estimate of the piano when extracted from the mixture using a new method to separate mixtures of harmonic sources (such as music recordings) with significant time-frequency overlap. This method is called Active Source Estimation, or ASE.

In the following sections we describe ASE and present a review of those portions of the DUET [25], and DASSS [14] algorithms used in ASE. We then provide a comparison of ASE to the performance of DUET and DASSS on stereo mixtures of musical instruments.

2. DEMIXING WITH BSS ALGORITHMS

The first step in separating multiple sources from a stereo mixture is mixing parameter estimation. Assume N sources recorded using two microphones. The sound from each source will travel a different distance to reach each microphone, and therefore the signal picked up by one microphone will have a different amount of attenuation and time delay from the signal picked up at another microphone.

Mixing parameter estimation is simply associating a particular attenuation and delay time value with each source. Let $x_1(t)$ and $x_2(t)$ represent the two signal mixtures of N sources recorded by the two microphones. These mixtures are defined in Equations 1 and 2. Here, a_j is the attenuation coefficient and δ_j is the delay time from the jth source to the second microphone.

$$x_1(t) = \sum_{j=1}^{N} s_j(t) \tag{1}$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t - \delta_j) \tag{2}$$

We do not lose generality by assuming microphone one picks up the source with zero delay and zero attenuation. If a source is closer to microphone two, mixture $x_2(t)$ will simply have attenuation greater than unity, and a negative time delay.

To estimate the mixing parameters for each mixture signal, we first calculate the windowed short term Fourier transform (STFT) for each mixture. We refer to the STFT of signal j as $X_j(\tau,\omega)$ where τ represents the center of a time window used for the STFT and ω represents a frequency of analysis used in the STFT. A time-frequency frame is a particular pair of values (τ,ω).

If numerous time-frequency frames in a mixture contain energy from no more than one source, one may estimate the amplitude, a_j, and delay, δ_j, mixing parameters for the jth source signal from the amplitude and delay for each frame using the DUET algorithm (Section 2.1). Once the mixing parameters for source j are identified, a time-frequency mask can be created for each source that accepts only time-frequency frames whose mixing parameters fall no farther than some epsilon from (a_j,δ_j). When the masked signal is transformed back to the time domain, the result will be the isolated source j [25].

2.1 The DUET Algorithm

The DUET [25] algorithm assumes two signal mixtures. To calculate the attenuation and delay parameters for the sources, we first calculate the ratio between the two signals $R(\tau,\omega)$ for each time and frequency, as shown in Equation 3.

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

The attenuation, a, and delay, δ, are then calculated as shown in Equations 4 and 5. Here, 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.

$$a(\tau,\omega) = |R(\tau,\omega)| \tag{4}$$

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

Once $a(\tau,\omega)$ and $\delta(\tau,\omega)$ have been calculated for the set of time-frequency frames of interest, 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 and delay values, $H(a,\delta)$.

The amplitude and delay 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. When the number of sources is known, we use a k-means clustering algorithm [23] to find the N most prominent peaks in the smoothed histogram.

Once the mixing parameters for each source have been estimated, the signals can be separated. To determine which time-frequency frames are associated with only one source, we use Aaron Master's modification of the demixing stage of the DUET algorithm, called delay and scale subtraction scoring (DASSS).

2.2 The DASSS Algorithm

The DUET algorithm allows for successful demixing when sources do not simultaneously produce energy at the same frequency and time. The delay and scale subtraction scoring (DASSS) [14] method allows estimation of which time-frequency frames in a mixture satisfy this condition. One may then reconstruct a source from the time-frequency frames in which only that source was active.

To determine which frames in a stereo mixture correspond to a single source, define a function, Y_j, for each pair of mixing parameters, (a_j,δ_j), associated with a source signal j. This is shown in Equation 6.

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

If only one source is active in a given time-frequency frame, $Y_j(\tau,\omega)$ takes on one of two values, as shown in Equation 7.

$$Y_j(\tau,\omega) = \begin{cases} 0 & \text{if } j = g \\ \alpha(j,g) X_1(\tau,\omega) & \text{if } j \neq g \end{cases} \tag{7}$$

In Equation 7, g is the index of the single active source. If $j = g$, then $Y_j(\tau,\omega)$ is equal to 0. If j is not the single active source, then $Y_j(\tau,\omega)$ takes the second value shown in Equation 7. This value depends on Equation 8. Here, e is the base of the natural logarithm and i is the imaginary number.

$$\alpha(j,g) = \frac{a_g}{a_j} e^{i\omega(\delta_j - \delta_g)} \tag{8}$$

Equation 9 compares the expected values (given the single-source assumption) for $Y_j(\tau,\omega)$ to the observed values for source g at time $\mathbf{t}$ and frequency $\boldsymbol{\omega}$.

$$d(g,\tau,\omega) = \frac{Y_g(\tau,\omega) + \sum_{\forall j \neq g} |\alpha(j,g) X_1(\tau,\omega) - Y_j(\tau,\omega)|}{\sum_{\forall j} |Y_j(\tau,\omega)|} \tag{9}$$

Here, as the function $d(g,\tau,\omega)$ approaches zero, the likelihood that source g was the only active one during the frame increases. The source that minimizes Equation 9 is taken to be the most active source in the time-frequency frame. A threshold value can then be used to determine which frames can be assigned directly to a single source estimate. [14].

3. ACTIVE SOURCE ESTIMATION (ASE)

DUET separates signals from a mixture when most time-frequency frames in the mixture contain energy from a single source. DASSS estimates which time-frequency frames are single-source, allowing source reconstruction from only single-source frames. As stated previously, reconstruction from only the single-source frames results in increasingly poor reconstruction as the proportion of single-source frames decreases.

The *Active Source Estimation* (ASE) algorithm for reconstruction of harmonic sources in musical mixtures improves reconstruction by using the energy in multi-source frames. ASE estimates *how many* sources are active in a particular time-frequency frame. It also estimates *which* sources are active. Finally, where two or more sources are active, it estimates *how much* energy to assign to each source.

We begin by observing that, in assigning energy from a time-frequency frame in a pair of mixtures to a set of sources, there are three cases of interest. The first case is where at most one source is active. In this case, the full energy from mixture X_1 may be assigned directly to an estimate of the source j, denoted $\hat{S}_j$.

The second case is where exactly two sources are active. In this case, we explicitly solve for the correct energy distribution by using the frequency domain equivalent of the system of equations described by Equations 1 and 2. This system is described by Equations 10 and 11.

$$X_1(\tau,\omega) = S_j(\tau,\omega) + S_k(\tau,\omega) \tag{10}$$

$$X_2(\tau,\omega) = a_j e^{i\omega\delta_j} S_j(\tau,\omega) + a_k e^{i\omega\delta_k} S_k(\tau,\omega) \tag{11}$$

Here, we assume the attenuation a, and delay δ, parameters are known for both source j and source k. In practice, attenuation and delay are estimated values. Solving for source j results in Equation 12.

$$S_j(\tau,\omega) = \frac{X_2(\tau,\omega) - a_k e^{i\omega\delta_k} X_1(\tau,\omega)}{a_j e^{i\omega\delta_j} - a_k e^{i\omega\delta_k}} \tag{12}$$

Substituting the value for source j into Equation 10 gives the value for source k. Once we have calculated the energy for both sources in the frame, we can add this energy to the source signal estimates.

The third case is where more than two sources are active. Since we have at least three unknown complex values, we cannot explicitly solve for the appropriate source energy as in the two sources case. In this case, we estimate the energy of each source by making assumptions about the class of input signals and learning from the signal energy distributed in the one and two active source cases.

Since we are primarily interested in demixing tonal music and tonal music depends on sources that have identifiable pitch, we assume the input signals are pitched sounds. If a sound has a pitch it is called *harmonic*, and it typically contains strong energy at integer multiples of a fundamental frequency called *harmonics*.

Constraining the class of input signals in this manner helps us estimate the energy in regions where sources are not disjoint in the time-frequency representation by estimating each source's fundamental frequency and making *harmonic masks* that represent the expected high-energy time-frequency frames for each source. We use these masks to estimate the number of active sources in each time-frequency frame. For frames with more than two active sources, we estimate each source's amplitude and phase to distribute energy to all active sources. The remainder of this section describes the ASE algorithm in detail.

3.1 An overview of ASE

In this section, we provide a step-by-step outline of the ASE source separation method. For further detail on any particular step, see the indicated subsection of the paper.

1. Determine the attenuation and delay parameters, (a_g, δ_g) for each source g. (DUET, in Section 2.1)
2. Find those time-frequency frames from mixture X_1 that correspond to a single source. (DASSS, in Section 2.2)
3. For each source g, create an initial estimate, $\hat{S}_g$ from the single-source frames corresponding to that source. (Section 3.2)
4. Create the *interference signals*, $I_1(\tau,\omega)$ and $I_2(\tau,\omega)$, from all frames not assigned to a single source. (Section 3.2)
5. For each source estimate $\hat{S}_g$, estimate the fundamental frequency at each time step τ. (Section 3.3)
6. Use the estimates of the source fundamental frequencies to create a *harmonic mask* that predicts the locations of harmonics for each source estimate $\hat{S}_g$. (Section 3.4)
7. Create the *source count*, $C(\tau,\omega)$ for each frame in interference signal $I_1(\tau,\omega)$ by counting the number of predicted source harmonics in each frame. (Section 3.4)
8. For every frame where the source count $C(\tau,\omega) = 2$, assign energy from $I_1(\tau,\omega)$ and $I_2(\tau,\omega)$ to the appropriate source estimates. (Section 3.4)
9. For every frame where $C(\tau,\omega) > 2$, estimate the magnitude and phase of each source and distribute the energy accordingly. (Section 3.5)
10. Convert each source estimate $\hat{S}_g$ into a time domain signal, $\hat{s}_g$, using overlap and add synthesis [22].

Figure 2 provides a diagram of the algorithm for clarity.

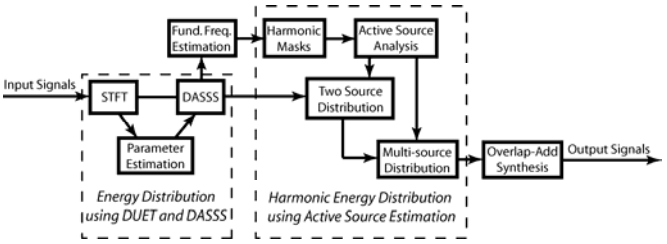

Figure 2: Signal flow diagram of the ASE algorithm.

The remainder of Section 3 describes the steps in the ASE algorithm in greater detail.

3.2 Creating Initial Source Signal Estimates

We use DASSS data and Equation 9 to distribute the time-frequency frames we are confident contain significant energy from only one source. Equation 13 calculates $\hat{S}_g$, the initial signal estimate for each source g.

$$\hat{S}_g(\tau,\omega) = \begin{cases} X_1(\tau,\omega) & \text{if } \left(d(g,\tau,\omega) < T\right) \wedge \\ & \left(g = \arg\min_{\forall j}(d(j,\tau,\omega))\right) \\ 0 & \text{else} \end{cases} \quad (13)$$

Equation 13 only assigns energy to a source estimate if it is the single active source in a particular frame. Here, T is a threshold value that determines how much energy from multiple sources a frame may contain and still be considered a "single source" frame.

Equation 14 defines the *interference signals*. Here, the subscript c indicates either the left or right mixture in a stereo recording. Each interference signal is simply the original mixture with all of the frames distributed to individual sources by Equation 13 set to 0.

$$I_c(\tau,\omega) = \begin{cases} X_c(\tau,\omega) \; if \; \forall g, d(g,\tau,\omega) > T \\ 0 \quad else \end{cases} \quad (14)$$

We now improve our source signal estimates by assigning energy from the interference signals to individual sources.

3.3 Fundamental Frequency Estimation

We denote the fundamental frequency of signal estimate $\hat{S}_g$ at time τ as $F_g(\tau)$. We determine fundamental frequency and harmonics-to-noise ratio, $HNR_g(\tau)$, of each signal estimate using an autocorrelation-based technique described in [6].

To smooth spurious, short-lived variation in the F_g estimates, any change in F_g over 6% (roughly half a critical band) that lasts less than 60 milliseconds results in the F_g estimates within that 60 millisecond window being changed to match the estimate in the frame prior to the transition. The duration of 60 milliseconds was chosen to roughly equal a sixteenth note at 120 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 F_g estimates for times τ with low harmonics-to-noise ratio ($HNR_g(\tau) < H_{min}$). For these times, we set the fundamental frequency estimate to be equal to that of the most correlated neighbor estimate. Let $\hat{S}_g(\tau_n,:)$ indicate the vector of values for signal estimate $\hat{S}_g$ at all frequencies of analysis at time τ_n. We begin with time frame one and move forward through the time frames. For each low-confidence estimate encountered, we measure cross-correlation between $\hat{S}_g(\tau_n,:)$ and the immediately preceding step, $\hat{S}_g(\tau_{n-1},:)$, and between $\hat{S}_g(\tau_n,:)$ and the next time step with a confident fundamental frequency estimate, $\hat{S}_g(\tau_{n+d},:)$. We replace $F_g(\tau)$ with the value from the time-step (either τ_{n-1} or τ_{n+d}) with the greatest cross-correlation.

Given F_g estimates during each time frame, we now create harmonic masks that represent the expected high energy regions of each source signal.

3.4 Creating Harmonic Masks, Determining Active Sources and Adding to Source Estimates

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 binary time-frequency masks that have a value of 1 for frames near integer multiples of the fundamental frequency and a value of 0 for all other time-frequency frames. This operation is defined by Equation 15.

$$H_g(\tau,\omega) = \begin{cases} 1 \text{ if } (\exists k \text{ such that } |kF_g(\tau) - \omega| < \Delta_\omega) \\ 0 \text{ else} \end{cases} \quad (15)$$

Here, Δ_ω, is the distance from the frequency of a harmonic that we tolerate when adding a frame to the mask. The value for Δ_ω is determined based on the expected accuracy of the fundamental frequency estimates and the frequency resolution of the time-frequency frames.

We use the harmonic masks to divide the time-frequency frames in the interference signal into two categories: frames where we believe two sources are active, and frames where we believe three or more sources are active. We do this by summing the harmonic masks for all the sources to create the active source count for each frame, $C(\tau,\omega)$.

$$C(\tau,\omega) = \sum_{\forall j} H_j(\tau,\omega) \quad (16)$$

If $C(\tau,\omega) = 2$, we presume the frame has two active sources and we use Equations 10 through 12 to solve for the source values, substituting the interference signals I_1 and I_2 for the mixture signals X_1 and X_2. If $C(\tau,\omega) > 2$, we estimate the relative strengths of the active sources using the method in Section 3.5.

3.5 Source Amplitude and Phase Estimation for Multi-source Frames

We call a time-frequency frame with three or more active sources a *multi-source* frame. Estimating the assignment of energy to the active sources for a multi-source frame is not a trivial task. Most separation algorithms for the two-mixture, multi-source case avoid estimation by either assuming disjoint representations of the sources exist and constraining the class of inputs accordingly [13,14,17,25], or by dealing only with binary time-frequency masks in which sources may share a frame. In the second case, no attempt is made to estimate the true source energies during that frame [5,9]. In this and the following sections, we present a method that allows us to create an estimate of the source amplitude and phase values from the source signal estimates created thus far.

When harmonics from multiple sources overlap, the result is often a sequence of multi-source time-frequency frames at frequency ω. With this in mind, we designed our method to deal with sequences of such frames.

Let τ_s be the starting time-step and τ_{s+n} be the ending time-step in a sequence of multi-source frames of frequency ω. We wish to estimate the values for $\hat{S}_g(\tau_s, \omega)$ through $\hat{S}_g(\tau_{s+n}, \omega)$, based on the energy in the multi-source frames $I_1(\tau_s, \omega)$ through $I_1(\tau_{s+n}, \omega)$ in the interference signal. To begin, we need an estimate of the energy in the first frame of the sequence.

3.5.1 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 the initial mixture was single or double source in the frame and source g was one of the active sources, then $\hat{S}_g(\tau_{s-1}, \omega)$ had a value assigned to it in a previous step.

If the fundamental frequency estimate for time-step $F_g(\tau_{s-1})$ falls within 6% of the current estimate $F_g(\tau_s)$, we presume the frames are drawn from the same event (the

same note) and that it is reasonable to infer a value for $\hat{S}_g(\tau_s, \omega)$ from $\hat{S}_g(\tau_{s-1}, \omega)$. In this case, we use Equation 17 to assign an amplitude to frame $\hat{S}_g(\tau_s, \omega)$.

$$\left|\hat{S}_g(\tau_s,\omega)\right| = \frac{A_g(\tau_s)}{A_g(\tau_{s-1})}\left|\hat{S}_g(\tau_{s-1},\omega)\right| \quad (17)$$

$$A_g(\tau) = \underset{\forall\omega \text{ s.t. } \hat{S}_g(\tau,\omega)>0}{mean}\left(\hat{S}_g(\tau,\omega)\right) \quad (18)$$

For the amplitude estimation we assume that a corrupted (interfered with by at least two other signals) harmonic is correlated in amplitude modulation with other (uncorrupted) harmonics from the same source signal. Although harmonics of a source are not always precisely correlated in amplitude modulation [18], they are typically correlated to some degree, making this a reasonable assumption. Equation 18 calculates $A_g(\tau)$, the average amplitude of all non-zero time-frequency frames within a time-step.

In Equation 19, we simply assume the current phase is the same as the phase of frame $\hat{S}_g(\tau_{s-1}, \omega)$. While this estimate could be improved in the future by measuring the change in phase over time, our current approach is straightforward and depends on only a single reliable time-frequency estimate.

$$\angle\hat{S}_g(\tau_s,\omega) = \angle\hat{S}_g(\tau_{s-1},\omega) \quad (19)$$

Equation 18 provides us with an amplitude estimate of at least one of the active sources for frame (τ_s, ω). Call this source *k*. We now estimate the other source amplitudes using Equation 20.

$$\left|\hat{S}_g(\tau_s,\omega)\right| = \frac{A_g(\tau_s)}{A_k(\tau_s)}\left|\hat{S}_k(\tau_s,\omega)\right| \quad (20)$$

The idea behind Equation 20 is that the relationship between the amplitude of source *k*'s harmonic and source *g*'s harmonic at frequency ω will be proportional to the relationship between source *k*'s average harmonic amplitude and source *g*'s average harmonic amplitude.

Since we have no good starting point to estimate phase for the sources whose amplitudes are derived using Equation 20, their initial phase is set to a default value of 0.

3.5.2 Estimation without a prior example

If the conditions required for Equations 18 and 19 are not met, then the prior frame is either from a different event (note) or has not been estimated. In this case, we set the phase estimate to a default value of 0 and use another method of amplitude estimation.

We rely on the fact that most multi-source activity takes place over successive time frames. Within this time span the interference signal will exhibit amplitude *beating*, or amplitude modulation that results from constructive and destructive interference between sources. The point of maximal amplitude in the interference signal between times τ_s and τ_{s+n} corresponds to the frame in which the most constructive interference between active sources takes place. To make our initial source amplitude estimation, we assume that this point of maximal constructive interference results from all active sources having the same phase and call this frame, τ_{MaxInt}. With this assumption, Equation 10, altered for the ***N*** active source case in frame (τ_{MaxInt}, ω), yields Equation 21.

$$\left|X_1(\tau_{MaxInt},\omega)\right| = \sum_{\forall g \text{ s.t. } H_g(\tau_{MaxInt},\omega)=1}\left|\hat{S}_g(\tau_{MaxInt},\omega)\right| \quad (21)$$

Equation 22 defines the relative amplitudes of each source based on a reference source ***k***.

$$RA_g = \frac{A_g(\tau_{MaxInt})}{A_k(\tau_{MaxInt})} \quad (22)$$

Combining Equations 21 and 22 to solve for source *k*'s amplitude value results in Equation 23.

$$\left|\hat{S}_k(\tau_{MaxInt},\omega)\right| = \frac{\left|X_1(\tau_{MaxInt},\omega)\right|}{1+\sum_{\forall g\neq k} RA_g} \quad (23)$$

To find $|\hat{S}_k(\tau_s,\omega)|$ from $|\hat{S}_k(\tau_{MaxInt},\omega)|$ we apply Equation 24.

$$\left|\hat{S}_g(\tau_s,\omega)\right| = \frac{A_k(\tau_s)}{A_k(\tau_{MaxInt})}\left|\hat{S}_k(\tau_{MaxInt},\omega)\right| \quad (24)$$

We then estimate the amplitudes of the other active sources using Equation 19. Phases for the frames in time-step τ_s are set to a default value of 0.

Once an amplitude and phase value has been estimated for each active source in the beginning of our multi-source sequence, we impose the phase and amplitude modulation trends of each source on the remaining frames (τ_1,ω) through (τ_{M-1},ω) by applying Equations 17 through 19.

We complete this process for each sequence of multi-source frames encountered in the interference signals.

4. EXPERIMENTAL RESULTS

In this section we compare the relative abilities of the ASE, DUET, and DASSS algorithms to separate sources from difficult (containing many consonant intervals) three instrument signal mixtures and one four-part piece of music. It should be noted that the results shown for the DASSS algorithm are based on a threshold ***T*** (from Equation 13) that accepts only frames that are highly likely to be single source, thus all frames in which more than two or more sources are active are ignored. Letting ***T***

increase to infinity results in the DASSS algorithm duplicating the performance of the DUET algorithm.

We now describe the methods used in the creation of signal mixtures and analysis of algorithm performance.

4.1 Mixtures with Three Instruments

The instrument recordings used in the testing mixtures are individual long-tones played by French horn, saxophone and oboe, all taken from the University of Iowa musical instrument database [10]. Mixtures of these recordings were created to simulate the stereo microphone pickup of spaced source sounds in an anechoic environment. We fix one instrument on a single pitch, middle C (262 Hz), and then vary the other instruments by half steps to examine the intervallic relationships that are most difficult for the algorithms to handle. This was done as follows.

First, we set all instruments to C. Next, while instrument one and two remain on C, we ascend the chromatic scale with the third instrument up to the C an octave above. We then move instrument two up one half step and again move instrument three in half steps from middle C to one octave up. We repeat this process until instrument two has also traversed the entire chromatic scale. The result is 169 mixtures of the three instrument tones.

Each variation was then separated into three signals by the DASSS, DUET, and ASE algorithms. Extracted sources were then compared for similarity to the original sources and the distance measured. The error measure used is a time-domain correlation measure. This is defined in Equation 25 as the logarithm of the inner product between the original source and the source estimate divided by the inner product of the source with itself.

$$E_g = \left| 20 \log \left(\frac{\left\langle \hat{s}_g, s_g \right\rangle}{\left\| s_g \right\|^2} \right) \right| \tag{25}$$

To provide a single score for each mixture dealt with by the algorithms, we calculate E_g for each of the three source estimates and take the mean error over all three estimates.

Figure 3 shows source estimation error generated by ASE, DASSS, and DUET for the 169 mixtures. Here, the darker the square, the greater the difference between the original sources and the estimates extracted from the mixture. The vertical axis is the pitch distance, in half steps, between instrument one (always middle C) and instrument two.

The horizontal axis corresponds to the distance in half steps from the pitch of instrument one to the pitch of instrument three. Thus, the square up four and right seven places from the lower left corner shows the error for a major triad with root C.

As expected, the worst performance in the DUET and DASSS algorithm correlates with the mixtures that have many time-frequency frames with energy from multiple sources.

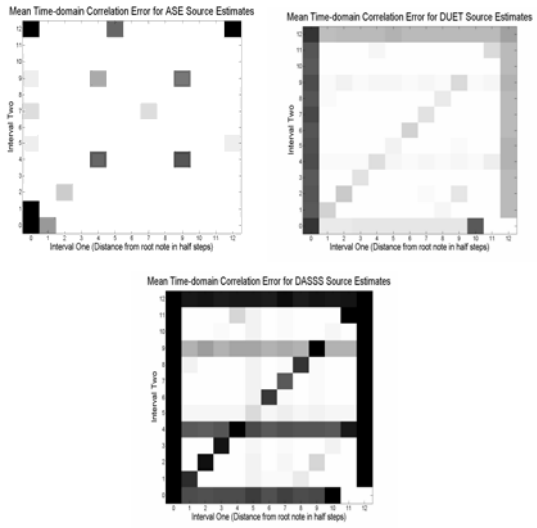

Figure 3: (top left) Mean time-domain correlation error between ASE source estimates and true source signals. (top right) DUET performance error using the same measure. (bottom) DASSS performance error using the same measure.

These problematic mixtures are seen even more clearly when looking at the DASSS performance results. The dark column on the far left side of each square corresponds to the mixtures in which a unison was played between instrument three and one. The far right side represents on octave between three and one, the top row is an octave between two and one and the bottom row is the unison between two and one. We also see poor performance along the diagonal, representing mixtures in which instruments two and three were playing in unison.

Note that the DASSS error is far greater for these consonant intervals because DASSS is designed to only distribute time-frequency frames that *confidently* came from one source.

Mixtures containing many time-frequency frames from multiple sources are precisely what ASE is designed to handle. As shown in Figure 3, the ASE dramatically improves demixing performance on these cases. The mixtures in which all three instruments are playing either in unison or in octaves (the corners of the graph) are still problematic, but the resultant error when two instruments are in unison or active rather than three is vastly improved.

ASE performs more poorly than DUET when it could not estimate the fundamental frequencies of the sources

accurately and therefore did not correctly categorize the multi-source time-frequency frames.

4.2 A Four Part Musical Example

As a final example, we compare ASE and DUET on a Bach four part chorale harmonization of *To God on High All Glory Be*, as played by a clarinet choir. The score of the excerpt used is provided in Figure 4.

Panel (a) of Figure 5 shows the spectrogram of an excerpt from the mixture of *To God on High All Glory Be*. Panel (b) shows the spectrogram of the original bass part. Panel (c) shows the ASE estimate of the bass from the mixture, and panel (d) shows the DUET estimates of the bass part. Note the improved resolution of the source estimate's harmonics when comparing the ASE source estimates to the DUET estimates. Also notice that while ASE is able to match the harmonic activity of the source more closely, it loses some of the low energy signal between harmonics and during note onsets. However, because the first demixing stage of the algorithm uses the DASSS data with threshold $\boldsymbol{T}$, we are able to manipulate $\boldsymbol{T}$ to create output signals that vary between the graph shown in (c) and that shown by DUET in (d).

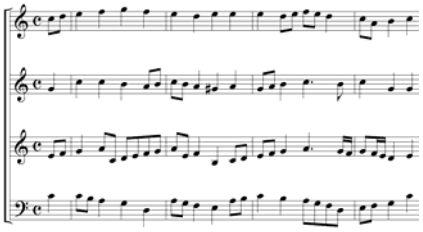

Figure 4: The first four measures of *To God on High All Glory Be.*

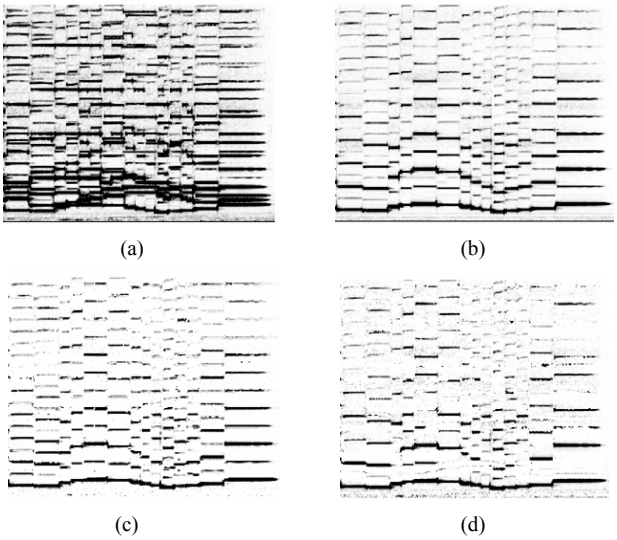

Figure 5. (a) Spectrogram of the four part mixture represented in the score in Figure 3. (b) Spectrogram of the original bass part, prior to mixture. (c) Spectrogram of ASE base estimate. (d) Spectrogram of the DUET bass estimate.

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. Such cases occur with great frequency in music recordings, making this approach especially useful in musical contexts.

We showed the ASE algorithm's improvement over the DASSS and DUET methods at separating individual musical instruments from contexts which contain problematic intervals such as unisons and octaves.

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 on-the-fly source models that don't assume harmonic sources from portions of the output that we are "confident" about. Put another way: if three instruments play a two note phrase and only the second chord results in interference, can we design an accurate source model from the output during the first chord?

In this paper, we introduced an analytic approach to assign energy from two-source time-frequency frames. That said, our methods of assigning energy from frames with more than two sources are heuristic and make somewhat unrealistic phase assumptions. 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 the 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 learning about the sources in order to deconstruct 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

[1] Aarabi, P., Shi, G., Jahromi, O. (2003). *Robust speech separation using time-frequency masking.* IEEE Int. Conference on Multimedia and Expo, Baltimore, Maryland.

[2] Anemüller, J., Kollmeier, B. (2000). *Amplitude modulation decorrelation for convolutive blind source separation.* Proceedings of the second international workshop on independent

component analysis and blind signal separation, Helsinki, Finland, pages 215-220.

[3] Avendano, C. (2003). *Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications.* IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

[4] Balan, R., Rosca, J. (2000). *Statistical Properties of STFT ratios for two channel systems and applications to blind source separation.* Proceedings ICA 2000, Helsinki.

[5] Bartsch, M. (2004). *Automatic Singer Identification in Polyphonic Music.* PhD Dissertation, University of Michigan department of Electrical Engineering and Computer Science.

[6] Boersma, P. (1993). *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, pp. 97-110.

[7] Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* The MIT Press, Cambridge, Massachusetts.

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

[9] Ellis, D. (1996). *Prediction-driven computational auditory scene analysis.* PhD Dissertation, Massachusetts Institute of Technology, Media Laboratory.

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

[11] Goto, M. (2004). *A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals*, Speech Communications 43, pp 311-329

[12] Klapuri, Anssi P. (2001). *Multipitch estimation and sound separation by the spectral smoothness principle.* ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City.

[13] Lee, T.W., Bell, A.J., Orglmeister, R. (1997). *Blind source separation of real world signals.* IEEE International Conference on Neural Networks, Houston.

[14] Master, A.S. (2003). *Sound source separation of n sources from stereo signals via fitting to n models each lacking one source.* Technical Report, CCRMA, Stanford University, 2003. Available from http://www-ccrma.stanford.edu/~amster/.

[15] Parra, L.C., Spence, C. D. (2001). *Separation of non-stationary natural signals.* Independent Component Analysis, Principles and Practice. Cambridge University Press, Pages 135-157.

[16] Reyes-Gomez, M.J., Ellis, D., Jojic, N. (2004) *Multiband Audio Modeling for Single-Channel Acoustic Source Separation.* ICASSP, Montreal.

[17] Rickard, S., Yilmaz, O. (2002) *On the Approximate W-Disjoint Orthogonality of Speech.* ICASSP 2002, Orlando, Florida, May 2002.

[18] Risset, J.C., Wessel, D. (1982). *Exploration of timbre by analysis and synthesis.* The Psychology of Music, Academic Press, NY.

[19] Rosenthal, D. Okuno, H.G. (1995). *Working Notes of the IJCAI-95.* Workshop on Computational Auditory Scene Analysis.

[20] Sheirer, E. (2000). *Music-Listening Systems.* PhD Dissertation, Massachusetts Institute of Technology, Media Laboratory.

[21] Smith, J.O., Serra, X. (1987). *PARSHL*: *An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation.* Proceedings of the International Computer Music Conference, Tokyo.

[22] Stone, J.V. (2004). *Independent Component Analysis: A Tutorial Introduction.* MIT Press, Cambridge, Massachussets.

[23] Theodoridis, S. and Koutroubas, K. (1999) *Pattern Recognition*, Academic Press, San Diego.

[24] Virtanen, T., Klapur, A. (2001). *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.

[25] Yilmaz, O., Rickard, S. (2003) *Blind Separation of Speech Mixtures via Time-Frequency Masking*. IEEE Transactions on Signal Processin