



Audio Engineering Society

Convention Paper

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Auditory Component Analysis

Jon Boley^{1*}

¹ Music Engineering Department, University of Miami, Coral Gables, FL, 33124, USA
**now with Shure Incorporated, Niles, IL 60714, USA*
Boley_Jon@shure.com

ABSTRACT

Two of the principle research areas currently being evaluated for the so-called sound source separation problem are Auditory Scene Analysis and a class of statistical analysis techniques known as Independent Component Analysis. This paper presents a methodology for combining these two techniques. It suggests a framework that first separates sounds by analyzing the incoming audio for patterns and synthesizing or filtering them accordingly. It then measures features of the resulting tracks and separates the sounds statistically by matching feature sets and attempting to make the output streams statistically independent. The proposed system is found to successfully separate artificial and acoustic mixes of sounds. As expected, the amount of separation is inversely proportional to the amount of reverberation present, number of sources, and interchannel correlation.

1. INTRODUCTION

The cocktail party effect was first described by Cherry [1] as our ability to listen to a single person in an environment where multiple people are speaking at once. This talent is related to localization, binaural masking level differences, auditory scene analysis, and perhaps several other psychoacoustic phenomena. Our ability to focus on a single sound source is due, in part, to our ability to localize sounds and to recognize patterns. Our ability to localize sounds stems from the fact that we have two ears, and the brain processes these sounds in such a way that sounds emitted from a particular location actually seem to be separated from other sounds within the brain. The goal of the field of

auditory scene analysis is to understand how the human auditory system analyzes our environment, and much of the research in this field directly relates to the problem of sound source separation. Independent component analysis is another active area of research and is being utilized for its ability to separate sounds such that they become statistically independent. Traditionally, the fields of computational auditory scene analysis (CASA) and independent component analysis (ICA) have been used separately to approach the problem of sound source segregation. These approaches appear to work well under very different conditions, thus illustrating the need for a combined system. Perhaps the most profitable application for sound separation is automatic speech recognition in crowded or noisy environments. However, sound separation could potentially also be used to improve spatialization of multichannel audio, to

separate musical instruments, to aid in forensic audio investigations, to improve assisted hearing devices, and many other applications.

2. AUDITORY SCENE ANALYSIS

Auditory Scene Analysis, as described in [2], is the process by which we are able to make sense of the complex auditory environment that we experience every day. At any given moment, our ears may pick up sound from dozens of individual sources. These sounds are mixed acoustically and arrive at our ears in a series of sound waves that are drastically different from the sound waves originating from any of the individual sources. Despite this fact, we are able to focus our attention on sounds originating from an individual source. One of the principal low-level processes that accounts for much of this source discrimination is known as stream segregation.

2.1. Stream Segregation

Much of the research within the field of auditory scene analysis has been related to auditory stream segregation. Stream segregation is the result of applying several grouping rules to organize a complex mixture of sounds into several streams. These streams correspond to the individual sound sources. A few of the cues that we subconsciously use to accomplish this task are harmonicity, common fate, and spectral/temporal proximity.

Perhaps the simplest feature by which sounds are grouped is harmonicity. Natural sounds do not consist of a single tone, but rather a combination of tones that are often harmonically related. A musical instrument playing an A440 note will not only produce a 440 Hz sine wave, but also sine waves at integer multiples of the fundamental frequency. However, we are also able to hear two instruments playing the same note as two separate instruments. In fact, we use several other grouping rules for discriminating between sound sources.

Another grouping rule is that of common fate. Harmonics that are modulated together in amplitude or frequency tend to be grouped together. For instance, if a series of harmonics maintains constant frequency ratios with respect to each other, and then suddenly only two of the harmonics begin oscillating in frequency, the listener will hear a single sound split into two— a steady sound and an oscillating sound. This segregation will also occur if multiple sounds share harmonic frequencies but begin oscillating (in either amplitude or

frequency) differently. In this case, each sound source is identified according to the phase of the modulation.

Sequences of sound events are grouped into perceptual streams when these sounds can be grouped according to various criteria. For example, a series of tone pulses alternating between a high frequency and a low frequency will sound like a single stream when played slowly. However, if this sequence is played fast enough, it will separate into two streams: one stream of low-frequency tones and another of high-frequency tones.

It is not only the temporal spacing that determines the perceptual streams but also the frequency separation. The relative spacing between sounds in the frequency domain determines the grouping of these sounds. In the above example, if the high-frequency tones are only slightly higher in pitch than the low-frequency tones, the separation would be more difficult. But if the relative spectral spacing is great, segregation is much easier.

2.2. The Weft as a Computational Model

The concept of the auditory *weft* was developed by Dan Ellis [3] to describe a new technique for computational ASA. The idea is based on the 3D correlogram, which maps frequency vs. autocorrelation lag and time. In contrast to previous techniques, wefts allow tracking of sounds that share frequency bands.

The correlogram is generated by first passing the sound through a cochlear filterbank, half-wave rectifying the resulting signals, and smoothing over a 1ms window. For each subband, the short-time autocorrelation is calculated such that the samples are spaced logarithmically to approximate human pitch perception. The resulting three-dimensional correlogram then displays the autocorrelation of each subband, with various lag periods, at a number of time frames.

A periodogram is then computed from the autocorrelation by normalizing the autocorrelation values (relative to zero-lag) and summing across subbands, resulting in a two-dimensional representation of autocorrelation lag vs. time. A threshold may then be set to identify strong modulation periods and thus possible weft elements. A search algorithm starts at the shortest period and looks for peaks. Each time it finds a peak, any peaks occurring at integer multiples of that period are subtracted to get rid of the subharmonic aliases created by the autocorrelation process. A hysteresis may be used to allow an existing period track to continue existing for a short period even if the peaks are slightly below the threshold. The spectra for each of the period tracks are then extracted from the 3D

correlogram at points corresponding to the period in question. Subharmonic aliases are also then subtracted from the correlogram so that the remaining value of the peak is roughly proportional to the energy of the modulation.

3. BLIND SOURCE SEPARATION

Blind source separation (BSS) is the technique used to separate independent signals from a set of mixed signals without any prior knowledge of the signals. The source signals $\mathbf{s}(t)=[s_1(t),s_2(t),\dots,s_m(t)]^T$ are to be estimated from the observed signals $\mathbf{x}(t)=[x_1(t),x_2(t),\dots,x_m(t)]^T$. The system is modeled as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \quad (1)$$

where the mixing matrix, \mathbf{A} , is unknown. It is assumed that the original source signals are independent and the mixing matrix is nonsingular; thus it is possible to compute a demixing matrix \mathbf{W} as follows:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (2)$$

where \mathbf{y} is the output signal vector. The signals represented by \mathbf{y} are similar to \mathbf{s} but they may be scaled and the order may be different.

Separation of component signals may be easily performed using blind source separation if the number of observed signals is equal to the number of original source signals. Notice that this implies that the mixing matrix, \mathbf{A} , is square.

3.1. Independent Component Analysis

Independent component analysis (ICA) is a method of blind source separation that is gaining tremendous popularity. Originally developed to tackle the cocktail party effect, ICA has found great use in brain imaging, telecommunications, financial analysis and various other fields.

The first principle of ICA estimation is that of nonlinear decorrelation. Not only should the components be uncorrelated, but the transformed components should also be uncorrelated. The second principle is that of maximum non-Gaussianity. By keeping the variance of \mathbf{y} constant, and finding the local maxima of the non-Gaussianity of $\mathbf{y}=\sum_i w_i x_i$, the independent component may be found. Each local maximum of the non-Gaussianity corresponds to one independent component.

The entropy of a vector \mathbf{y} is defined as

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \quad (3)$$

where $p(\mathbf{y})$ is the probability density function of \mathbf{y} . Because a Gaussian distribution, by definition, has the greatest entropy for a given variance, the negentropy J can be defined as

$$J(\mathbf{y}) = H(\mathbf{y}_{\text{gauss}}) - H(\mathbf{y}) \quad (4)$$

where $\mathbf{y}_{\text{gauss}}$ is a Gaussian random vector with the same correlation matrix as \mathbf{y} . To curtail the complexity of the negentropy calculations, several approximations have been introduced. One of the best approaches is to use expectations of higher-order functions. By replacing the polynomials y^3 and y^4 with other functions G_i , we can approximate the negentropy with the following equation:

$$J(\mathbf{y}) \approx k_1(E\{G_1(\mathbf{y})\})^2 + k_2(E\{G_2(\mathbf{y})\} - E\{G_2(v)\})^2 \quad (5)$$

where k_1 and k_2 are positive constants and v is a Gaussian variable with zero mean and unit variance. By carefully choosing the functions G_1 and G_2 , robust calculations can be obtained. The following functions have been shown in [4] to work well:

$$G_1(y) = \frac{1}{a_1} \log \cosh a_1 y \quad (6)$$

$$G_2(y) = -\exp\left(\frac{-y^2}{2}\right)$$

Where $1 \leq a_1 \leq 2$ is a constant, usually chosen to be one.

The Fast ICA algorithm introduced in [5] uses the above equations to implement ICA as shown in Figure 1.

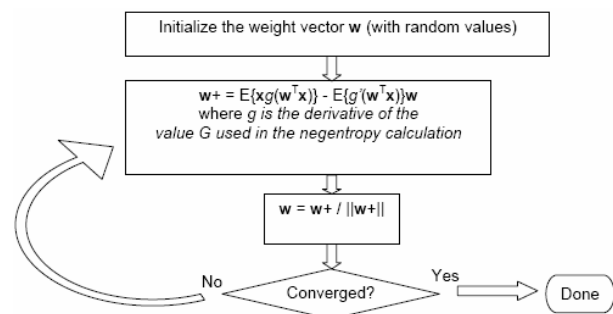


Fig 1. Fast ICA Algorithm

3.2. Intelligent ICA

“Intelligent ICA” was presented by Mitianoudis and Davies as a conference paper at the May 2002 AES Convention [6]. They began by extracting several feature vectors from solo recordings of instruments that they desired to recognize, and creating a database of Gaussian Mixture Models (GMM). The Intelligent ICA algorithm then used the information in the GMM of the desired instrument to iteratively locate this instrument and separate it from the mix. Combined with Fast ICA, the resulting algorithm is diagrammed in Figure 2.

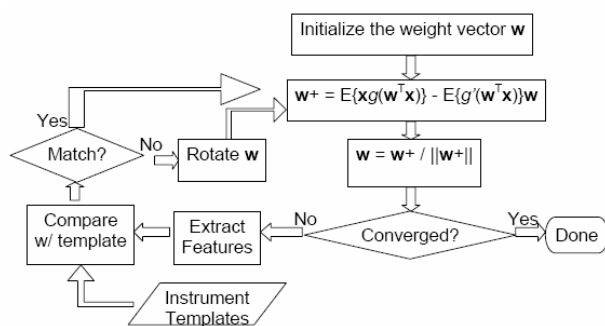


Fig 2. Intelligent ICA Algorithm

In his work on musical timbre, John Grey [7] pointed out that the three greatest contributing factors to timbre are the spectral envelope, how the spectrum changes over time, and the attack characteristics of the sound. Since then, several algorithms have been designed to extract these features (and more) from a sound. The frequency envelope can be captured using linear prediction coefficients (LPC), mel-frequency cepstral coefficients (MFCC), or perceptual linear predictive coefficients (PLPC). The MFCC, as described in [8], has proven to be an effective measure for various sound sources. To calculate this, the Fourier transform is first performed on the frame of audio data, the amplitudes are converted to a logarithmic scale, the frequencies are converted to the perceptual mel-frequency scale, and finally the discrete cosine transform is performed.

4. AUDITORY COMPONENT ANALYSIS

4.1. Analysis of the Auditory Scene

The proposed algorithm uses this correlogram to group sounds based on phase coherence and spectral/temporal proximity. The tracking of these sounds is reminiscent of the sinusoidal tracking by MacAuley and Quatieri [9], but is accomplished in a

slightly different manner. Points in the 3D correlogram are grouped together if they exceed a predetermined threshold and fall within a given distance from another point. For simplicity, the number of resulting tracks may be limited to the largest N tracks. An example of this track creation is shown below in Figure 3.

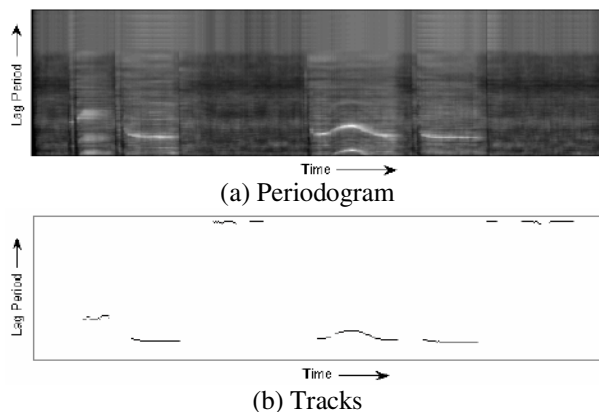


Fig 3. Creation of Period Tracks

Each track may then be either resynthesized or filtered out from the original track. To synthesize a track, the fundamental period at a particular moment is used to create a pulse train (like that shown in Figure 4) that may be filtered to resemble the spectrum of the original track at that time (as shown in Figure 5). By adding together several overlapping frames that have been synthesized in this manner, the entire track is synthesized.

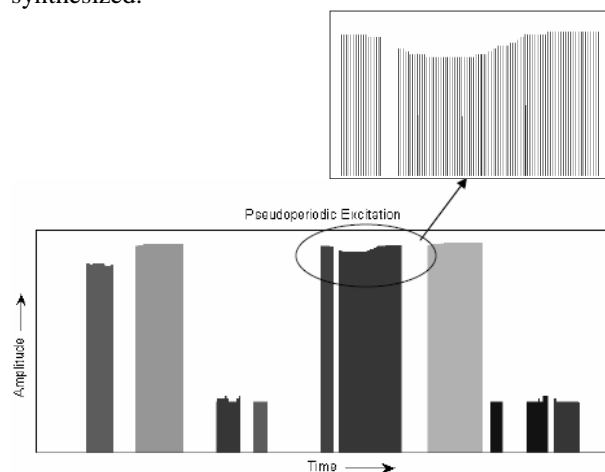


Fig 4. Pulse trains corresponding to extracted periods

Alternatively, the track may be filtered out of the mixture by setting bandpass filters at the fundamental and harmonic frequencies within each frame. These filters may correspond to masking curves appropriate

for the amplitude and frequency of the source. For each synthesized or filtered track, feature vectors (such as Mel-frequency cepstral coefficients) are calculated and a Gaussian mixture model (GMM) is created.

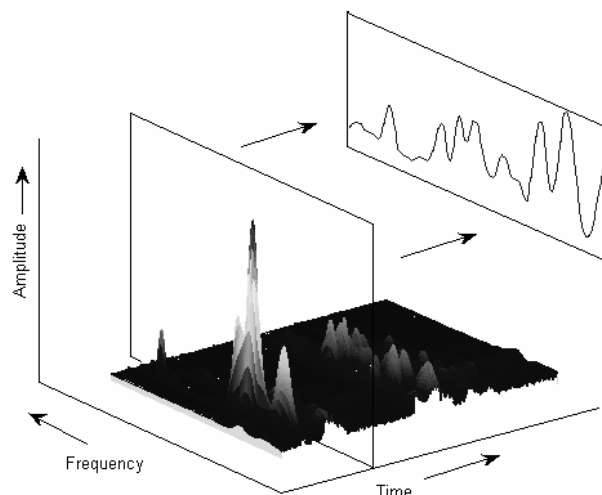


Fig 5. Scaling Function used to shape pulse train

4.2. Source Identification and Separation

The proposed algorithm bypasses the need for a database by dynamically creating a model for each track. After track analysis, a user may select the desired track. After each iteration of ICA, a GMM for the separated component is compared to the GMM of the desired track. If they do not match, the algorithm attempts to separate another component. Once a match is found, the algorithm continues to converge until the desired sound has been separated from the mix.

In contrast to the Intelligent ICA algorithm described in [6], this algorithm does not require any training period. As in Intelligent ICA, the GMM is adapted according to the Expectation Maximization algorithm, but instead of a prolonged training phase, the learning takes place during the separation process. Because the training is derived from the same input signal, the probability density function described by the associated GMM models the desired source very closely. The GMM is initialized according to the harmonic signal extracted by means of the ASA processing block. This information is used to identify the independent component and extract it from the mix using the following algorithm. After each iteration of the FastICA algorithm, feature vectors are calculated for the separated component and compared with the GMM corresponding to the desired source. While the independent component matches the desired source, the mixing matrix continues to adapt until it converges.

However, if the independent component does not match the desired source, the estimated mixing matrix is rotated such that an orthogonal signal becomes the next estimate for the output. (This is done just as it is accomplished in Intelligent ICA.) The process is repeated until the mixing matrix converges to a point such that the output independent component matches the desired GMM, or until a maximum number of iterations have been performed. Twenty iterations appear to be more than enough for convergence. The resulting two independent components correspond to the desired source and the remaining mixture of all other sources. This process may be used iteratively on the residual mixture(s) to extract multiple components, as shown in Figure 6.

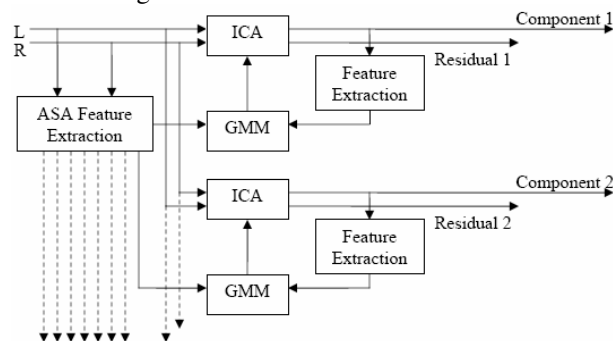


Fig 6. Auditory Component Analysis Algorithm

5. EXPERIMENTS

5.1. Measurements

Evaluation criteria for separation algorithms, both CASA and BSS systems, have often been subjective and nearly always inconsistent. Many have evaluated their systems based on speech recognition rate, while others have used signal-to-noise ratios or simple subjective listening tests. A standardized set of tests was proposed in [10] which defines measures of distortion and separation.

Distortion is defined as follows:

$$D_j = 10 \log \left(\frac{E\{(x_{j,s_j} - \alpha_j y_j)^2\}}{E\{x_{j,s_j}^2\}} \right) \quad (7)$$

where

$$\alpha_j = \frac{E\{x_{j,s_j}^2\}}{E\{y_j^2\}} \quad (8)$$

The indices j are chosen such that output y_j corresponds to an original audio source signal s_j . Notice that this distortion measure is minimal when the output y_j is equal to x_{j,s_j} (the contribution of the source signal s_j to the observed signal x_j). The distortion may also be calculated as a function of frequency by applying the short-time Fourier transform:

$$D_j(\omega) = 10 \log \left(\frac{E \left\{ \left| \text{STFT} \{ x_{j,s_j} - \alpha_j y_j \} \right|^2 \right\}}{E \left\{ \left| x_{j,s_j} \right|^2 \right\}} \right) \quad (9)$$

The quality of separation is defined as

$$S_j = 10 \log \left(\frac{E \left\{ \left| y_{j,s_j} \right|^2 \right\}}{E \left\{ \left| \sum_{i \neq j} y_{j,s_i} \right|^2 \right\}} \right) \quad (10)$$

where y_{j,s_i} is the output when only source s_i is active.

5.2. Test Data

Several standard test cases (also proposed in [10]) were used to evaluate the performance of this algorithm, ranging from a time-varying tone mixed with noise to complex mixtures involving various amounts of reverberation.

The first set of audio test files is derived from a recording of two speakers, one male and one female, in the presence of background noise. The room measured 3.1 meters high, 4.2 meters wide, and 5.5 meters deep. The recordings were made with both speakers in the presence of the two omni-directional microphones, but with only one person speaking at a time. The contributions of the two speakers to each microphone may be added together to create the observed signals x_1 and x_2 . The data set consists of the four audio files described in Table 1.

File Name	Description
spc1s1m1.wav	Contribution of Source 1 to Microphone 1
spc1s1m2.wav	Contribution of Source 1 to Microphone 2
spc1s2m1.wav	Contribution of Source 2 to Microphone 1
spc1s2m2.wav	Contribution of Source 2 to Microphone 2

Table 1. Audio Test Data Set #1

The remaining sets of test data have been created synthetically and include the sounds described in Table 2. Each signal is accompanied by a corresponding gated signal which turns the signal on and off at various points.

File Name	Description
sine.wav	Sine wave oscillating in frequency
sawtooth.wav	Sawtooth wave oscillating in frequency
square.wav	Square wave oscillating in frequency
gaussn.wav	Gaussian noise
cauchyn.wav	Cauchy noise
gsine.wav	Gated sine wave oscillating in frequency
gsawtooth.wav	Gated sawtooth wave oscillating in frequency
gsquare.wav	Gated square wave oscillating in frequency
ggaussn.wav	Gated Gaussian noise
gcauchyn.wav	Gated Cauchy noise

Table 2. Synthetic Audio Test Data

The gated signals were defined to overlap with each other as shown in Figure 7. These sounds may be mixed in a number of ways. The simplest option is to add signals together to create composite mixes. The sources may also be filtered by a simple FIR or Head Related Transfer Function (HRTF) filter before mixing. Yet another possibility is to apply a simulated or measured room response, specifying the location of sources and sensors

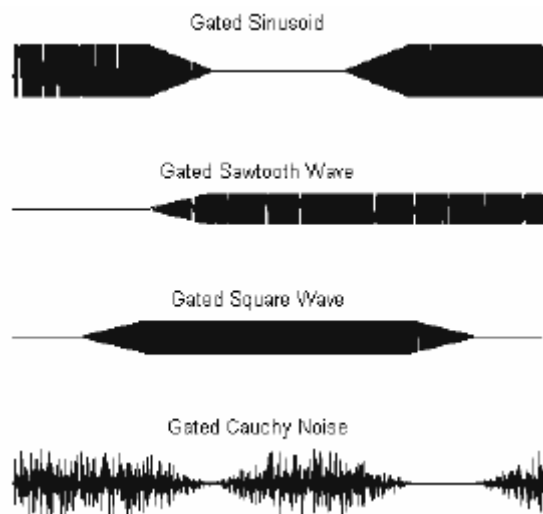


Figure 7. Gated test signals

5.3. Test Procedure

Distortion and separation measures are calculated for several different variations of sources as well as assorted source and sensor locations. Five test cases are described below.

Test Case #1: An oscillating sine tone mixed with Cauchy noise.

Test Case #2: The stereo recording of the male and female speakers.

Test Case #3: Gated sawtooth wave, gated sine wave, gated Cauchy noise filtered with Head Related Impulse Responses (HRIR) at 30° left of center, center, and 30° right of center, respectively. (These are simply used to act as low-order mixing filters, not because they are related to spatial listening.) As shown in Figure 27, there are two HRIRs for each angle. Both HRIRs are applied to the signal. The signals corresponding to the left ear are added together and the signals corresponding to the right ear are added together. The center channel is not actually filtered with an HRIR.

Test Case #4: Sawtooth wave, gated sine wave, and Cauchy noise, filtered by a dense mixing FIR filter, resembling the response of a room. The FIR filters are designed such that none of the coefficients are zero. The coefficients are generated from an exponentially decaying Cauchy noise.

Test Case #5: Gated square wave, sine wave, sawtooth wave, and gated Cauchy noise, each placed in corners of a virtual room (1 meter from each wall), with a pair of virtual sensors in the center of the room 1 meter apart.

6. RESULTS

Performance ranged from 116dB of separation in the most simple case (tone+noise) to only 0.2dB in the most difficult case (2 simultaneous speakers). However, additional tests showed that recordings of 3 similar instruments playing the same note simultaneously could be separated by at least 16dB in some scenarios.

6.1. Data

Test Case #1 (An oscillating sine tone instantaneously mixed with Cauchy noise) The oscillating sine tone was extracted from the mixture by choosing the GMM that corresponds to the appropriate period track, and the separation was measured to be 116dB according to (10).

Figure 8 shows the distortion of each channel for the first test case. It can be seen that the separated tone (BSS output 2) has very little distortion from the noise (source 2). Instead, nearly all of the energy is due to the tone itself (source 1).

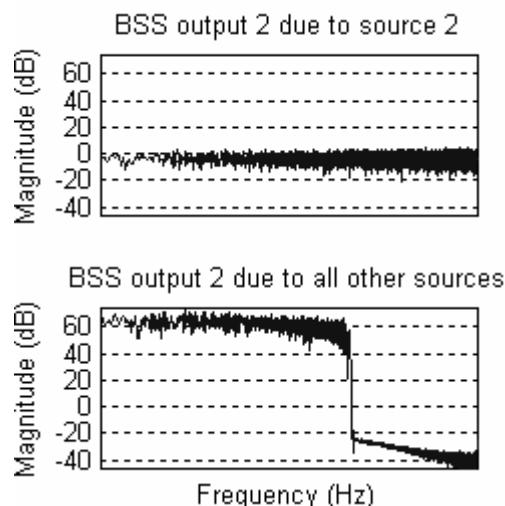


Fig 8. Distortion Measures for Test Case #1

Test Case #2 (Stereo recording of the male and female speakers) The male speaker was extracted from the mixture, and the separation was measured to be 0.2dB.

Figure 9 shows the distortion of each channel for the second test case. Notice that each output has about an equal amount of energy from both sources.

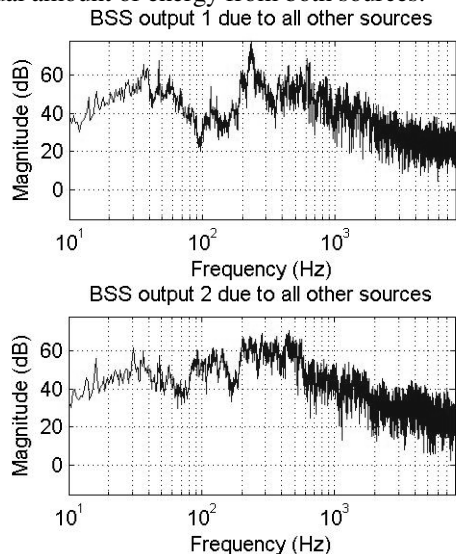


Figure 9. Distortion Measures for Test Case #2

Test Case #3 (Gated sawtooth wave, gated sine wave, gated Cauchy noise filtered with Head Related Impulse Responses at 30° left of center, center, and 30° right of center, respectively.) The sawtooth wave was extracted from the mixture, and the separation was measured to be 43.7dB.

Figure 10 shows the distortion of each channel for the third test case. Note that the energy in BSS output 1 due to the combination of source 1 and source 2 is lowest, signifying that source 3 (the sawtooth wave) is most prevalent.

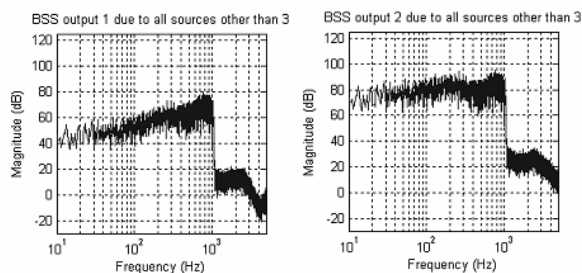


Figure 10. Distortion Measures for Test Case #3

Test Case #4 (Sawtooth wave, gated sine wave, and Cauchy noise, filtered by a dense mixing FIR filter, resembling the response of a room.) The sawtooth waveform was extracted from the mixture, and the separation was measured to be 15.5dB.

Figure 11 shows the distortion of each channel for the fourth test case. Note that the energy in BSS output

2 due to the combination of source 1 and source 2 is lowest, indicating that source 3 (the sawtooth wave) has been separated.

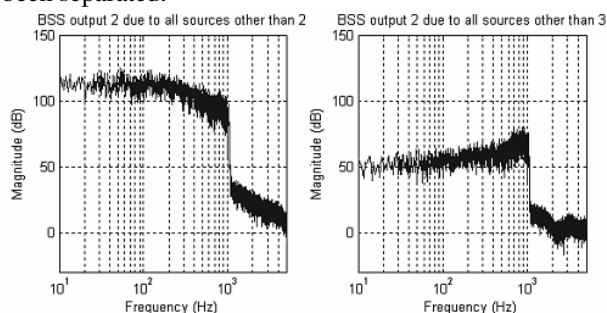


Figure 11. Distortion Measures for Test Case #4

Test Case #5: (Gated square wave, sine wave, sawtooth wave, and gated Cauchy noise, each placed in corners of a virtual room (1 meter from each wall), with a pair of virtual sensors in the center of the room 1m apart.) The square wave was extracted from the mixture, and the separation was measured to be approximately 1.6dB.

Figure 12 shows the distortion of each channel for the fifth test case. The energy distribution of the various sources is roughly equivalent, telling us that separation has not occurred.

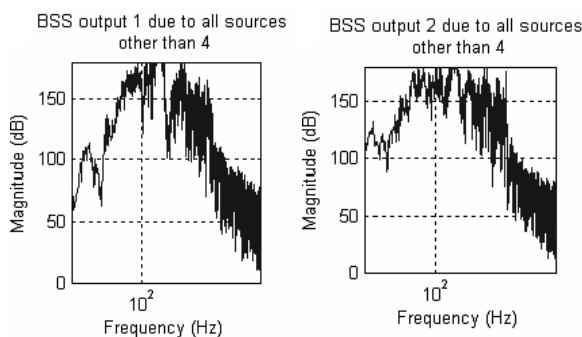


Figure 12. Distortion Measures for Test Case #5

6.2. Additional Data

A few additional tests were performed to obtain more data. The first such test was a simple experiment to see if anechoic recordings of musical instruments could be separated. The recordings were randomly panned. Table 3 shows the various combinations of mixtures tested and the resulting separation.

A bass clarinet (originally panned 18% right) was separated from mixture of a flute panned 38% left, a bassoon panned 28% left, and an alto sax panned 20% right. The separation here measured 14.8dB. However, a trumpet (originally panned 40% right) was only

separated 3dB out of a mixture containing a bass clarinet (18% right) and a violin (46% left). Two additional cases are shown below.

Separated Instrument	Alto Sax (Right 20%)	Alto Sax (Right 20%)	Bass Clarinet (Right 18%)	Trumpet (Right 40%)
Flute (Left 38%)	12.5dB	6.8dB	14.8dB	X
Bassoon (Left 28%)	X			X
Alto Sax (Right 20%)	X	X		X
Bass Clarinet (Right 18%)	X	X	X	3dB
Violin (Left 46%)	X	X	X	

Table 3. Instrument Separation

Because panning seemed to have an effect on the performance of the separation algorithm, another test was performed to ascertain just how panning affects the amount of separation. Four cello recordings were panned to varying degrees in the stereo soundstage. The cellos were placed at -15° , -5° , 5° , and 15° . The separation results are shown in Table 4. As seen here the worst case was a collection of four cellos (playing the same note) placed across 30° of the horizontal plane. The best case, two cellos spaced 30° apart, achieved 79dB of separation.

-15°	-5°	5°	15°	Separation
	X	X		6 dB
X			X	79 dB
X	X	X		16 dB
X	X	X	X	3 dB

Table 4. Separation of panned cellos

Finally, a third additional test was added to show that the separation algorithm would work with speech. Using three anechoic speech samples panned far left, center, and far right, 15.6dB of separation was achieved when separating the center voice.

6.3. Remarks

By far the best performing of the five original test scenarios was the simplest- the separation of the sine wave from the instantaneous mixture with noise was 116dB. This should be expected because the ICA algorithm attempts to maximize non-Gaussianity. Separation of a pure tone (minimum entropy) from a random sequence (much higher entropy) should result in more separation than two signals with similar entropies.

The FIR room response method did fairly well with 15.5dB of separation, but the mixture that simulated a virtual room response did not do so well, with only 1.6dB of separation. This system is a little more complex than the HRTFs, and this may account for the decreased amount of separation.

The scenario which used the simulated room performed rather poorly. This could be due in part to the fact that this system had four sources, whereas the HRTF and FIR mixtures were composed of only three sources. However, it also more closely modeled an actual stereo recording, with longer reverberation tails and a phase difference between the microphone signals. This more complex system presents a significant challenge for sound-separation tasks, and the proposed system will need more work before it handles this situation well.

The most difficult scenario for this system was the stereo recording of two speakers talking simultaneously. With only 0.2dB of enhancement, it did not provide any significant separation. This is simply a more complex situation than the simulated room because it was recorded in an actual room (in the presence of background noise and a significant amount of reverberation).

In the additional tests, instruments were separated from several different mixtures of a variety of instruments. The first test case, an alto saxophone was separated from a mixture of the alto sax and a flute. The resulting output was separated 12.5dB. When a bassoon was added to the mix, the separation of the alto sax was reduced to only 6.8dB. This would suggest that as the number of sources increases, the amount of separation decreases. However, when the bass clarinet was added to the mix (for a total of four instruments) then removed, the separation improved to become 14.8dB. This drastic change suggests that something about the bassoon is significantly different. The bassoon was panned right 18%, while the alto sax, flute, and bassoon were panned right 20%, left 38%, and left 28% respectively. In the previous test, the alto sax was spatially separated from the flute and bassoon, but in this test the bassoon and the alto sax are very close.

Therefore, it appears that the spatial location was not the determining factor in this test, so further testing is needed to determine the specific cause. Most likely, the reason for the increased separation in this test is due to some statistical properties of the bassoon. If the bassoon signal is less Gaussian than the other signals, the separation should be far simpler.

To evaluate the importance of spatial separation, various combinations of stereo panned cellos were mixed together and a single instrument separated from each mixture. When two cellos were panned just slightly (5° left and right), the separation was only 6dB, but when the stereo separation was increased to 15° left/right, the separation increased to 79dB. Apparently, a slight increase in spatial separation can have a significant effect on the amount of separation possible. When this is extended to three and four cellos, the separation diminishes to 16dB then 3dB respectively, showing that the number of sources also has an effect on the amount of separation that may be achieved. As the number of sources increases, the separation task becomes more difficult.

7. CONCLUSION

7.1. Analysis of Results

Over the past several years, great advances have been made in the field of audio processing. New techniques have been created for engineers to analyze and process audio. Recent research into computational auditory scene analysis has given us tools to better understand the perception of our auditory environment. Advances in statistical signal processing, such as independent component analysis, have enabled us to separate conglomerations of measurements into their respective components. However, computational ASA has not given us a framework for effectively separating the many sounds that surround us and ICA generally requires a large number of sensors to allow separation of the independent components.

The proposed algorithm combines these two areas by first examining the perceived audio streams then, using information gathered from these streams, separating the audio such that the output audio files are statistically independent. It was shown that separation of a single sound source from a mixture of many is possible with the proposed algorithm. Separation ranged from 116dB to no separation, depending on the mixing conditions. The results presented in the previous chapter show that as the complexity of the listening environment increases, the sound separation problem

becomes more difficult. However, simple mixtures, such as an instantaneous mix or a low-order filtering, can be separated considerably. It was also shown that separation of stereo panned sounds can be accomplished with relative ease. However, the degree of separation depends not only on the reverberation time, but also the spatial separation and the number of sources present. As spatial separation increases, more separation is possible. As with all blind source separation problems, separation of many sources is much more difficult than two or three.

7.2. Future Work

Given more time, some improvements may be done to improve the performance of this separation algorithm. One such improvement would be to extract more features. A set of cepstral coefficients could be calculated for the attack of each sound in addition to a set of coefficients for the steady-state portion. This would better model the significant effect of attack characteristics on the timbre of a sound. In addition to calculating the cepstral coefficients, features such as the spectral centroid, crest factor, onset asynchrony, and amplitude envelope may allow even greater precision in the separation process. A perceptual model that encompasses and prioritizes these features could be very useful for future sound separation endeavors.

It may even be possible to improve separation by considering sound source location. Incorporating localization information into the auditory scene analysis algorithm would be a challenging project that might produce significant improvements. After all, sound localization is a very strong cue for the human auditory scene analysis problem, especially in noisy, crowded, and reverberant environments. Including such information in a computational model might alleviate some of the difficulties that arise from having multiple sources and long reverberation tails.

8. ACKNOWLEDGEMENTS

This work was supported in part by a grant from Dolby Laboratories.

9. REFERENCES

- [1] Cherry, E. C., "Some experiments on the recognition of speech, with one and with two ears," *Journal of Acoustic Society of America*, vol. 25, pp. 975-979, 1953.

- [2] Bregman, Albert S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990.
- [3] Ellis, Dan and David Rosenthal. "Mid-level Representation for computational auditory scene analysis". In *Proc. Of the Computational Auditory Scene Analysis Workshop*, 1995.
- [4] Hyvärinen, Aapo and Erkki Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, 13(4-5):411-430, 2000.
- [5] Hyvärinen, A. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis." *IEEE Transactions on Neural Networks* 10(3):626-634, 1999.
- [6] Mitianoudis, Nikolaos and Mike Davies, "Intelligent Audio Source Separation Using Independent Component Analysis," In *Proceedings of the 112th Convention of the Audio Engineering Society*. Munich, Germany, May 10-13 2002.
- [7] Grey J. M., "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1270-1277, 1977.
- [8] Mermelstein, Paul, and Steven B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28(4), pp. 357-366, 1980.
- [9] McAulay, Robert J. and Thomas F. Quatieri, "Speech analysis /synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744-754, Aug. 1986.
- [10] Schobben, D., K. Torkkola and P. Smaragdis, "Evaluation of Blind Signal Separation Methods", in *Proceedings Int. Workshop Independent Component Analysis and Blind Signal Separation*, Aussois, France, January 11-15 1999.