BLIND SEPARATION OF REAL WORLD AUDIO SIGNALS USING OVERDETERMINED MIXTURES

Alex Westner and V. Michael Bove, Jr.

MIT Media Laboratory
20 Ames Street
Cambridge, MA 02142, USA
westner@media.mit.edu, vmb@media.mit.edu

ABSTRACT

We discuss the advantages of using overdetermined mixtures to improve upon blind source separation algorithms that are designed to extract sound sources from acoustic mixtures. A study of the nature of room impulse responses helps us choose an adaptive filter architecture. We use ideal inverses of acquired room impulse responses to compare the effectiveness of different-sized separating filter configurations of various filter lengths. Using a multi-channel blind least-mean-square algorithm (MBLMS), we show that, by adding additional sensors, we can improve upon the separation of signals mixed with real world filters.

1. INTRODUCTION

Humans have the ability to focus their attention on any one sound in an environment filled with many different sounds. Digital audio systems, as well, would benefit from having this ability (termed by E. Collin Cherry in 1953 as the "cocktail-party effect."[3]); some potential applications include: instrument separation in a multitrack recording studio, speaker separation in a videoconferencing session, and audio stream segregation for a wearable audio computer [14]. In this work, we attempt to improve upon the extraction of acoustic sound signals by studying *overdetermined* mixtures, where we have more microphones than sound sources.

These applications all have in common the task of *source separation.* Furthermore, we do not know beforehand what the sounds are or how they are mixed together, so we must exclusively use the sound mixtures themselves to extract out the original sound sources, a process commonly known as *blind source separation* [7].

When using conventional source separation algorithms, we assume that the original signals are mixed together instantaneously [7]. Two microphones in a room, however, will record an acoustic sound at two different propagation delays. In addition, the microphones will pick up several delayed and modified copies of the original sound source, as it reflects off of walls and objects in the room. The reverberation and absorption characteristics of a room can be modeled as a finite impulse response (FIR) filter and convolved with the original sound source to simulate the signal recorded by a microphone [11].

Several researchers have extended blind source separation algorithms to cope with delayed and convolved sources [15, 8, 9, 4]; most of these algorithms have only implemented NxN configurations, using N sensors to separate out N sources. (Lambert [8] implemented an MxN example with more sensors than sources.)

Researchers often use beamforming microphone arrays when recording sounds in a reverberant environment. Beamforming arrays target their sound capture toward a desired spatial area, improving upon the signal-to-noise ratio (SNR) of the sounds recorded from that region. The delay and sum beamforming algorithm time-aligns the signals recorded by each sensor in the array and then adds them together. Thus, signal components emanating from a desired location combine coherently, while components from other locations combine incoherently. This increases the gain of the desired signal over the undesired noise; the SNR is a monotonically increasing function of the number of sensors [13].

In an effort to take advantage of the SNR gains that microphone arrays can achieve, we propose to extend current blind sound source separation algorithms to that of the overdetermined case, where we have *more* sensors than sources. We begin with a study of the nature of room impulse responses to help us choose an adaptive filter architecture. We then use the ideal inverses of acquired room impulse responses to compare the effectiveness of different-sized separating filter con-

This work has been supported by the Digital Life Consortium at the MIT Media Laboratory.

figurations of various filter lengths. Finally, using a multi-channel blind least-mean-square (MBLMS) algorithm, we show that, by adding additional sensors, we can improve the blind separation of signals mixed with real world filters.

2. ROOM IMPULSE RESPONSES

The most efficient way to experiment with real world signals is to generate them by taking a clean sound source (i.e. close-miked speech in a dry room) and convolving it with a known impulse response of a room. By using artificially generated mixtures, we know what the mixing filters are and we can use them to determine how long our separating filters need to be to achieve good results. In addition, we can more easily perform a quantitative analysis on our results.

Using the system designed by Bill Gardner and Keith Martin [5], we took impulse response measurements of a 3.5m x 7m x 3m conference room. Two and a half walls of the room are covered with whiteboards, one wall is covered with a projection screen and a large table sits in the middle of the room. A large projector and a lighting grid (to which the microphones are attached) hang from the ceiling. See Figure 1 for a photo of the room, and Alex Westner's thesis [16] for a more detailed diagram of the room layout.

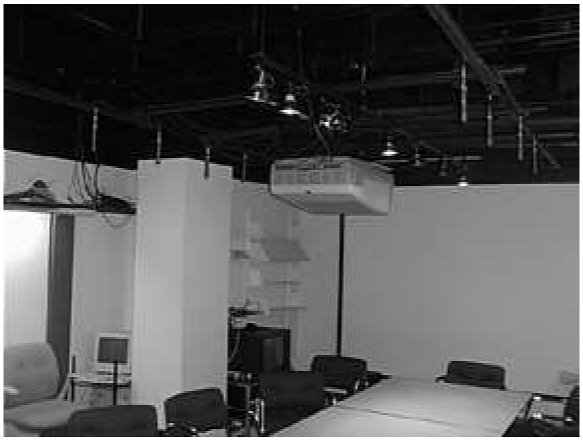

Figure 1: *There are eight microphones hanging from the lighting grid in the conference room.*

Based upon the orientation of the lighting grid, we constructed two linear microphone arrays, each with four elements. The microphones within each array are spaced about a half-meter apart from one another, as suggested by Dan Rabinkin *et. al.* [13] in optimum sensor placement.

We collected 8 impulse responses from 24 different locations around the room. To ensure that we would capture the full response of the room, we set the acquisition software to compute responses of approximately 750ms. After downsampling the data to a sampling rate of 11.025kHz, this equates to a 8,192-point response (See Figure 2).

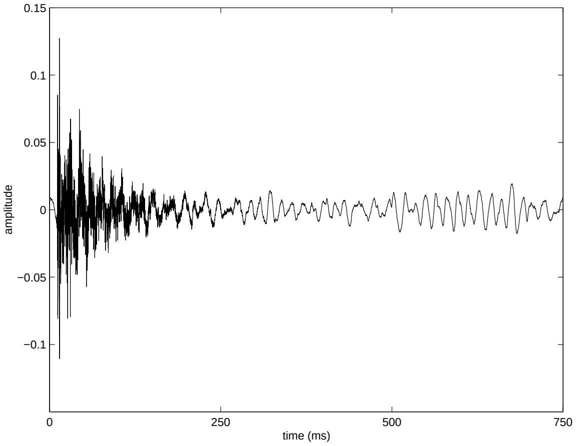

Figure 2: *A typical room impulse response.*

A strong characteristic low-frequency murmur, an artifact of the room configuration, dominates the impulse response. Following the example of Rabinkin *et. al.* [12], we applied a 200Hz high-pass filter to the impulse responses to remove this "room mode noise." The resulting impulse response is both aurally and visually cleaner (See Figure 3). It is perfectly acceptable to filter the impulse response before convolving it with the source; it has the same effect as filtering a signal recorded directly from the room itself.

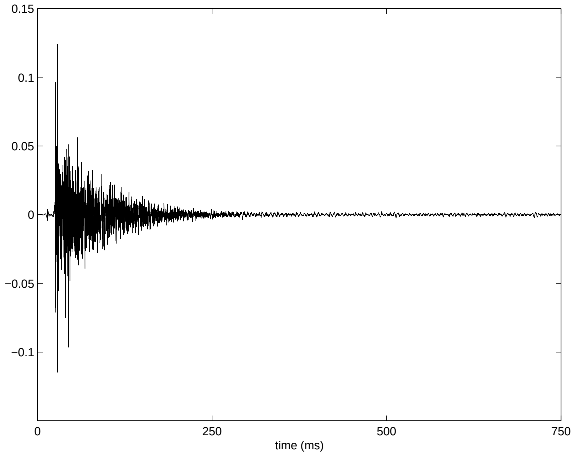

Figure 3: *A high-pass filtered impulse response.*

Upon visual inspection of the room impulse response, we can see that it is non-minimum phase. In general, for a filter to be minimum phase, the first sample should be larger than all other samples, and the response should decay rapidly [10]. In terms of blind separation algorithms, this means that we will be unable to use a feedback filter configuration, like the one suggested by Kari Torkkola [15], since they are only capable of inverting minimum phase filters [6].

3. INVERTING ROOM IMPULSE RESPONSES

We can model the problem of blind separation of real world audio signals by forming an FIR polynomial matrix, $A(t)$, whose elements are the room impulse responses that, when convolved with a vector of sound sources $s(t)$, will generate a vector of mixed signals, $x(t)$:

$$x(t) = A(t) * s(t)$$

The goal is to determine $W(t)$, the inverse of $A(t)$, which we can use to convolve with $x(t)$ to yield estimates, $u(t)$, of the original sources:

$$u(t) = W(t) * x(t)$$

As described in Russell Lambert's thesis [8], we can apply standard scalar matrix algorithms to invert FIR polynomial matrices. The following shows how to invert a 2x2 FIR matrix A.

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

The inverse to A is:

$$W = A^{-1} = \frac{1.}{a_{11} * a_{22} - a_{12} * a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

In the overdetermined case, however, $A(t)$ is not a square matrix. Therefore, we need to do a pseudoinverse to find $W(t)$. The pseudoinverse of a matrix is simply $\mathrm{inv}(A^H A) * A^H$, where A^H denotes the Hermitian transpose.

Figure 4 shows a block diagram of how to obtain $W(t)$ from $A(t)$. To speed computation, we transform $A(t)$ into the frequency domain by applying an FFT to each filter in the matrix. This allows us to multiply filters together instead of having to convolve them in the time domain. After computing the pseudoinverse, we move back into the time domain by applying an IFFT to each filter in the pseudoinverse matrix. Since $A(t)$ contains non-minimum phase filters, its inverse will be anti-causal. Therefore, we then need to rotate the leading weights of the time-domain inverse to the middle of the filters. Finally, to "clean up" the edges of the filters, we apply a Hanning window to the shifted, time-domain inverse, $W(t)$.

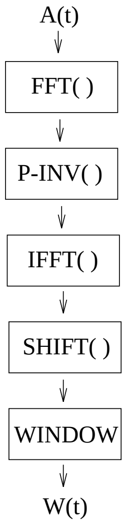

Figure 4: *A block diagram of how to invert an overdetermined room impulse response matrix.*

4. IDEAL UNMIXING FILTERS

We experimented with four different filter configurations for the blind separation and deconvolution of two sound sources, and for each configuration, we used unmixing filters of various lengths. The experiment proceeded as follows.

First, we generated acoustic sound mixtures by convolving clean sound sources (downloaded from Dominic Chan's web site [2]) with a matrix of room impulse responses. Using the appropriate impulse responses and sources, we created two sets of mixtures: for one set, we put a source in channel 1 and nothing in channel 2, and for the other set, we put a source in channel 2 and nothing in channel 1. By processing the two mixtures separately, it was easier to determine the resultant SNR's.

Using the procedure described in the previous section, we determined the separating matrix by inverting the mixing matrix. To vary the filter lengths, we applied an L-point Hanning window (centered around the peak of each filter) to the 8,192 tap unmixing filters, where L is the desired filter length. We then convolved the separating matrix with the mixture vectors to get an estimate of the original sources. We obtained separation SNR measurements by computing how much the channels with the sources bled into the channels without the sources.

Figure 5 shows one unmixing filter from each of the four configurations that we tested. We can make two important observations by visually comparing these unmixing filters. Most importantly, notice how dense the

2x2 unmixing filter is compared to the other three. Each 2x2 unmixing filter clearly requires more information to separate the mixtures than the other three configurations.

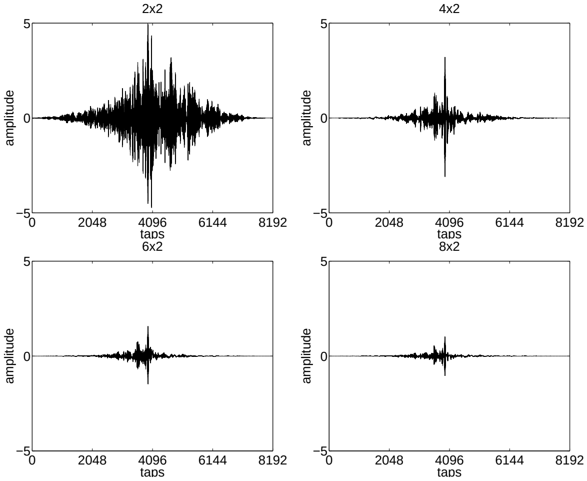

Figure 5: *Unmixing filters for a 2x2, 4x2, 6x2, and 8x2 configuration. Filter lengths of 8,192 taps were used to generate these filters.*

Secondly, the range in amplitude of the unmixing filters decrease as the number of sensors in the configuration increase. This can be explained by the fact that each unmixing filter in an MxN configuration adds M modified copies of a mixed signal to produce the output. Therefore, the more copies that are added together, the lower the amplitude for each copy. An important corollary of this observation follows: when using a blind deconvolution algorithm, (most of) the weights are initialized to zero. It is, therefore, beneficial if the slowly-adapting filters do not need to reach such high amplitudes to converge upon a solution.

The SNR measurements, listed in Table 1, clearly show the benefits of using overdetermined mixtures to separate acoustic sound mixtures. To obtain the data, we ran several trials for each filter configuration and filter length, using different source locations and different combinations of sound sources. We observed no bias for any particular source location or type of sound used, so we averaged our results based on the filter configuration and filter length.

As expected, longer unmixing filter lengths yield better separation. As we shortened the filter lengths to 256 and 128 taps, the separating filters began to severely distort the signals, therefore making these SNR measurements invalid.

	8192	4096	2048	1024	512	256	128
8x2	**36.9**	**33.4**	**24.6**	**16.4**	8.6	3.3	*5.1*
6x2	**33.0**	**28.4**	**21.8**	**13.0**	3.3	-3.0	*1.2*
4x2	**28.9**	**24.9**	**13.4**	4.2	2.0	*-2.4*	*6.8*
2x2	**15.8**	**13.8**	8.2	4.0	0.6	*-1.2*	*-3.4*

Table 1: *SNR measurements: boldface numbers show good, consistent separation (based on listening to the outputs); italicized SNR values are invalid due to signal distortion. All values are in dB.*

More significant to this work, however, is that using more microphones yields better separation. With filter lengths of 1,024 points, for example, using 8 microphones instead of 2 or 4 provides an additional 12dB of separation. Since it is generally more difficult for a blind separation and deconvolution algorithm to adapt to longer filters, these results encourage us to use overdetermined mixtures whenever possible.

5. OVERDETERMINED BLIND SEPARATION

Some of the more commonly used blind separation and deconvolution adaptation rules are constrained to only handling square matrices of filters [1]. For our experiment, we used the multichannel blind least-mean-square algorithm (MBLMS) described by Lambert [8].

The MBLMS algorithm attempts to minimize the cost function J where u is the estimated output and g is the Bussgang nonlinearity that uses prior knowledge of the probability density function (pdf) of the sources.

$$J = tr\ E\{(u-g)(u-g)^H\}$$

The weight update equation is determined from the cost function, where x is the mixture of sources:

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial u}\frac{\partial u}{\partial W} = \frac{\partial J}{\partial u}x^*$$

$$\frac{\partial J}{\partial u} = (u-g)$$

$$W = W + \mu(u-g)x^*$$

We ran the algorithm on the four different filter configurations used in the previous section, using truncated and windowed room impulse responses as our mixing filters. Figure 6 shows one of these responses. We set the algorithm to learn 512-tap filters, and we used 200,000 gamma-distributed (speech-like) random samples.

We used a Multichannel Intersymbol Interference (ISI) performance metric to determine how close the learned unmixing filters were to a scaled and/or permuted identity FIR matrix [8]:

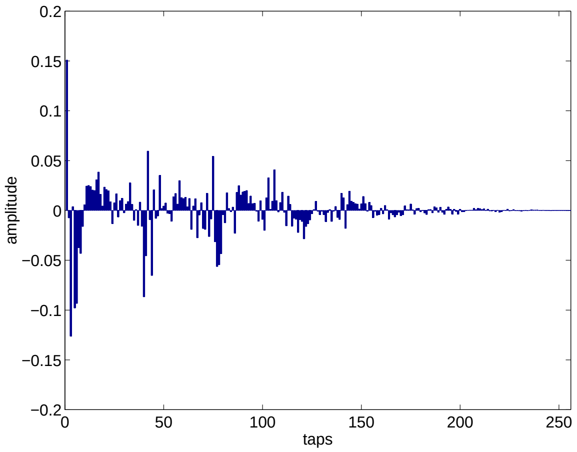

Figure 6: *A shortened room impulse response used in MBLMS algorithm. The sampling rate was 11.025kHz.*

$$ISI_i = \frac{\sum_j \sum_k |s_{ij}(k)|^2 - max_{j,k}|s_{ij}(k)|^2}{max_{j,k}|s_{ij}(k)|^2}$$

where s_{ij} are the filter elements of the mixing matrix, W convolved with the separating matrix, A. The ISI converges to zero for a perfectly learned unmixing matrix. Figure 7 shows a comparison plot of the ISI measurements as the algorithm ran through all the sample points for each configuration.

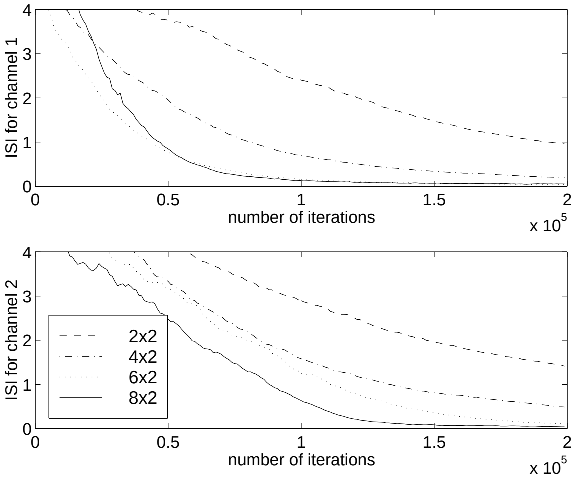

Figure 7: *ISI measurements obtained from the MBLMS algorithm for four different filter configurations.*

The plots show that the MBLMS algorithm performs significantly better when the number of sensors are increased.

6. CONCLUSION

Much of the work done so far in the blind separation of real world audio signals has been centered around square mixtures, where N sensors are used to separate N sources. We have examined room impulse responses (and their inverses) and compared blind separation and deconvolution results from different filter configurations (using an MBLMS algorithm). The results reported in this work encourage us to extend these blind separation algorithms to handle the overdetermined case.

7. ACKNOWLEDGMENTS

We would like to thank Russell Lambert, Te-Won Lee, and Paris Smaragdis for their insightful comments and assistance on this work. We are additionally grateful to Bill Gardner for sharing his knowledge of room impulse responses and to Freedom Baird for her support and encouragement.

8. REFERENCES

[1] Amari, S., A. Cichocki and H. Yang. A New Learning Algorithm for Blind Signal Separation. *Advances in Neural Information Processing Systems, 8*, pp. 757-763, 1996.

[2] Chan, Dominic. Blind signal separation audio demonstrations WWW page. www2.eng.cam.ac.uk/~dcbc1/research/demo.html

[3] Cherry, E. Collin. Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustical Society of America, 24*, pp. 975-979, 1953.

[4] Ehlers, F. and Schuster, H. G. Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. *IEEE Transactions on Signal Processing, 45(1) no.10*, pp. 2608-2612, 1997.

[5] Gardner, Bill and Keith Martin. HRTF Measurements of a KEMAR Dummy-Head Microphone. *MIT Media Lab Perceptual Computing - Technical Report #280*, 1994.

[6] Haykin, Simon. *Adaptive Filter Theory*, Third Edition, Upper Saddle River, New Jersey: Prentice Hall, 1996.

[7] Herault, Jeanny and Christian Jutten. Space or time adaptive signal processing by neural network models. *AIP Conference Proceedings, 151*, pp. 206-211, 1986.

[8] Lambert, Russell H. Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures. PhD Thesis, University of Southern California, 1996.

[9] Lee, Te-Won, Anthony J. Bell and Russell H. Lambert. Blind separation of delayed and convolved sources. *Advances in Neural Information Processing Systems, 9*, pp. 758-764, 1996.

[10] Neely, Stephen T. and Jont B. Allen. Invertibility of a room impulse response. *Journal of the Acoustical Society of America, 66(1)*, pp. 165-169, 1979.

[11] Orfanidis, Sophocles J. *Introduction to Signal processing*, Upper Saddle River, New Jersey: Prentice Hall, 1996.

[12] Rabinkin, Daniel V., Richard J. Renomeron, Arthur Dahl, Joseph C. French, James L. Flanagan and Michael H. Bianchi. A DSP Implementation of Source Location Using Microphone Arrays, *Proceedings of the SPIE, 2846*, pp. 88-99, 1996.

[13] Rabinkin, Daniel V., Richard J. Renomeron, Joseph C. French, and James L. Flanagan. Optimum microphone placement for array sound capture. *Proceedings of the SPIE, vol.3162*, pp. 227-239, 1997.

[14] Roy, Deb K., Nitin Sawhney, Chris Schmandt and Alex Pentland. Wearable Audio Computing: A Survey of Interaction Techniques. *Technical Report*, MIT Media Lab, 1997.

[15] Torkkola, Kari. Blind Separation of Convolved Sources Based on Information Maximization. *Proceedings of the 1996 IEEE Workshop on Neural Networks for Signal Processing*, pp. 423-432, 1996.

[16] Westner, Alexander G. Object-Based Audio Capture: Separating Acoustic Sounds. M.S. Thesis, Massachusetts Institute of Technology Media Laboratory, 1998.