SPEECH/NOISE SEPARATION USING TWO MICROPHONES AND A VQ MODEL OF SPEECH SIGNALS

Alex Acero, Steven Altschuler[*] and Lani Wu[*]

Speech Technology Group
Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA
http://research.microsoft.com/srg

ABSTRACT

In this paper we address the problem of using two or more microphones to enhance speech corrupted by nonstationary noise, such as that of a competing speaker (cocktail party effect) at very low SNR, by means of linear filtering of two microphone signals. This work is a variant to the probabilistic Independent Component Analysis (ICA) method but using a more accurate probability distribution of the speech signal based on a mixture Autoregressive model. Comparison with other algorithms on published real recordings shows more separation than with previously existing ICA algorithms.

1. Introduction

The problem of separating the desired speech from interfering sources, the *Cocktail Party Effect* [5], has been one of the holy grails in signal processing. In this paper we discuss a technique to extract the speech signal using two microphones. A goal of this work is to improve the robustness of speech recognition systems to reverberation and interfering noises, a major cause for degradation.

Blind Source Separation (BSS) is a set of techniques that assume no information about the mixing process or the sources, hence is termed blind. *Independent Component Analysis* (ICA) is a set of techniques developed in the last few years [8][11] to solve the BSS problem that estimate a set of linear filters to separate the mixed signals under the assumption that the original sources are statistically independent. *Microphone arrays* are another set of techniques that limits the degrees of freedom in the filters to mostly delay and sum.

In this paper we develop a new objective function that is not blind, because it uses a more accurate probabilistic model of speech. While many of the techniques described in the literature have been proven to work on simulated data or recorded in a sound-proof room, the proposed technique has shown good separation on real recordings with office noise, background music, interfering speakers and reverberation. In fact, on a set of widely available two-source two-microphone recordings in an office environment [9], this algorithm outperformed several others published in the literature according to informal listening tests.

In Section 2 we describe ICA for instantaneous mixing, and in Section 3 an ICA convolutional mixing model, more realistic than the instantaneous mixing, for the case of using only two microphones. In Section 4 we then present a general MAP estimate using an accurate model, a speech recognizer, and in Section 5 an approximation using Vector Quantization. Implementation details are included in Section 6 with experimental results in Section 7.

2. Instantaneous Mixing ICA

We outline Independent Component Analysis [11] for instantaneous mixing in this section. Let's assume that the *R* microphone signals $y_i[n]$, $\mathbf{y}[n]=(y_1[n], y_2[n],\cdots, y_R[n])$, are obtained by a linear combination of the *R* unobserved source signals $x_i[n]$, denoted by $\mathbf{x}[n]=(x_1[n], x_2[n],\cdots, x_R[n])$:

$$\mathbf{y}[n]=\mathbf{V}\mathbf{x}[n] \tag{1}$$

for all *n* with **V** being the *R*x*R* mixing matrix. This mixing is termed *instantaneous* since the sensor signals at time *n* depend on the sources at the same, but no earlier, time point. Were the mixing matrix being given, its inverse could have been applied to the sensor signals to recover the sources by $\mathbf{x}[n]=\mathbf{V}^{-1}\mathbf{y}[n]$. In the absence of any information about the mixing, the *blind separation* problem consists of estimating a separating matrix $\mathbf{W}=\mathbf{V}^{-1}$ from the observed microphone signals alone. The source signals can then be recovered by

$$\mathbf{x}[n]=\mathbf{W}\mathbf{y}[n] \tag{2}$$

We'll use here the probabilistic formulation of ICA, though alternate frameworks for ICA have been derived too [6]. Let $p_{\mathbf{x}}(\mathbf{x}[n])$ be the probability density function (pdf) of the source signals, so that the pdf of microphone signals $\mathbf{y}[n]$ is given by

$$p_{\mathbf{y}}(\mathbf{y}[n])=|\mathbf{W}|\,p_{\mathbf{x}}(\mathbf{W}\mathbf{y}[n]) \tag{3}$$

and if we furthermore assume the sources $\mathbf{x}[n]$ are independent from themselves in time, $\mathbf{x}[n+i]$ $i\neq 0$, then the joint probability is given by

$$\begin{aligned} e^{\Psi} &= p_{\mathbf{y}}(\mathbf{y}[0],\mathbf{y}[1],\cdots,\mathbf{y}[N-1]) \\ &= \prod_{n=0}^{N-1} p_{\mathbf{y}}(\mathbf{y}[n]) = |\mathbf{W}|^N \prod_{n=0}^{N-1} p_{\mathbf{x}}(\mathbf{W}\mathbf{y}[n]) \end{aligned} \tag{4}$$

It can be shown [11] that the gradient of Ψ is given by

$$\frac{\partial \Psi}{\partial \mathbf{W}} = \left(\mathbf{W}^T\right)^{-1} + \frac{1}{N}\sum_{n=0}^{N-1}\phi(\mathbf{W}\mathbf{y}[n])\left(\mathbf{y}[n]\right)^T \tag{5}$$

where $\phi(\mathbf{x})$ is given by

$$\phi(\mathbf{x}) = \frac{\partial \ln p_{\mathbf{x}}(\mathbf{x})}{\partial \mathbf{x}} \tag{6}$$

from which a gradient descent solution, the so-called *infomax* rule [4], can be obtained for **W** given $p_{\mathbf{x}}(\mathbf{x})$.

[*] Presently at Rosetta Impharmatics, Kirkland, WA http:///www.rii.com

It can be shown [7] that an exact solution can be found up to a scaling factor and source permutation. If a Gaussian density is assumed for $p_{\mathbf{x}}(\mathbf{x})$, $\phi(\mathbf{x})$ is linear and the solution merely decorrelates the signals and doesn't guarantee separation: it could be a rotation of the original signals. Empirically, it has been found that using a non-linearity for $\phi(\mathbf{x})$ in Eq. (5), such as a sigmoid function [4], provides better separation. The use of other density functions for $p_{\mathbf{x}}(\mathbf{x})$, such as a mixture of Gaussians [3], also result in a nonlinear $\phi(\mathbf{x})$ and has also shown better separation.

3. Convolutional Mixing ICA

The case of instantaneous mixing is not realistic, as we need to consider the transfer functions between the sources and the microphones created by the room acoustics. While the problem can be formulated using *R* sources as shown in the previous section, we will limit ourselves from now on to the case of two sources and two microphones. The results can easily be extended to *R* sources and microphones.

Let $x_1[n]$ and $x_2[n]$ be two point sources, which are captured by two microphones whose signals, $y_1[n]$ and $y_2[n]$ respectively. In the model of Figure 1, the input signals are filtered with filters $g_{ij}[n]$, and then mixed to generate the microphone signals. These filters could be estimated directly through ICA [14].

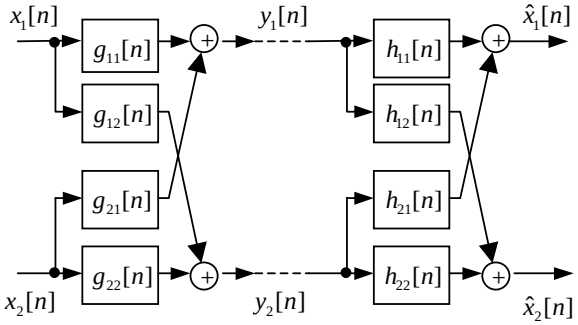

Figure 1. Model of source mixing and reconstruction.

It can be shown that the reconstruction filters $h_{ij}[n]$ in Figure 1 will completely recover the original signals $x_i[n]$ if and only if their *z*-transforms are the inverse of the *z*-transforms of the mixing filters $g_{ij}[n]$:

$$\begin{pmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{pmatrix} = \begin{pmatrix} G_{11}(z) & G_{12}(z) \\ G_{21}(z) & G_{22}(z) \end{pmatrix}^{-1} = \frac{1}{G_{11}(z)G_{22}(z) - G_{12}(z)G_{21}(z)} \begin{pmatrix} G_{11}(z) & G_{12}(z) \\ G_{21}(z) & G_{22}(z) \end{pmatrix} \tag{7}$$

It's reasonable to assume $g_{ij}[n]$ to be FIR filters, whose length will generally depend on the reverberation time, which in turn depends on the room size, microphone position, wall absorbance, etc. In general this means that the reconstruction filters $h_{ij}[n]$ have an infinite impulse response. In practice, it's convenient to assume such filters to be FIR of length *q*, which means that the original signals $x_1[n]$ and $x_2[n]$, will not be recovered exactly. The problem now is to estimate the reconstruction filters $h_{ij}[n]$ directly from the microphone signals $y_1[n]$ and $y_2[n]$ so that the estimated signals $\hat{x}_i[n]$ are as close as possible to the original signals.

Reverberation can be accounted for with this model as well as arbitrary transfer functions. The assumption that the model makes is that there are no nonlinearities and there are only two point sources (no extra noise for example). The assumption of no nonlinearities is deemed quite reasonable, but the assumption that there are only two sources may not be, and the fact that those two sources are point sources may also not be accurate. If the matrix in Eq. (7) is not invertible, separability is impossible. This can happen if both microphones pick up the same signal, which could happen if either the two microphones are too close to each other or the two sources are too close to each other, though in our experiments this did not happen.

The instantaneous mixing ICA algorithm of Section 2 cannot directly be used for convolutive mixing. However, if we operate in the frequency domain, we can apply it to individual frequency components independently [2][10][13][15]. Scaling and permutation still occur on each frequency, which is now more serious because the reconstructed signals could have frequency components belonging to different signals. This frequency based ICA algorithm can separate synthetically mixed signals reasonably well, but it does not do nearly as well on real recordings.

4. MAP Filters Using a Speech Recognizer

The probability distributions in [2][10][13][15] are not very accurate for speech signals because they do not model correlation across time directly. We propose in this section a more accurate probabilistic model of speech signals to guide us in the estimation of the reconstruction filters.

In the scenario we are describing we are only interested in obtaining the target speech signal, as the other source is considered interference noise. Without lack of generality, we will concentrate on an estimate of the desired signal $\hat{x}[n]$:

$$\begin{aligned} \hat{x}[n] &= h_1[n] * y_1[n] + h_2[n] * y_2[n] \\ &= \sum_{l=0}^{q-1} h_1[l] y_1[n-l] + \sum_{l=0}^{q-1} h_2[l] y_2[n-l] \end{aligned} \tag{8}$$

Now let's introduce some vector notation that will help algorithm description. Let's define vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ as

$$\begin{aligned} \mathbf{h}_1 &= \left(h_1[0], h_1[1], \cdots, h_1[q-1]\right)^T \\ \mathbf{h}_2 &= \left(h_2[0], h_2[1], \cdots, h_2[q-1]\right)^T \end{aligned} \tag{9}$$

and the *M* sample microphone signals for *i*=1,2 as

$$\mathbf{y}_i = \left\{y_i[0], y_i[1], \cdots, y_i[M-1]\right\} \tag{10}$$

We would like to find the MAP estimate for the reconstructed signal $\hat{\mathbf{x}} = \left\{\hat{x}[0], \hat{x}[1], \cdots, \hat{x}[M-1]\right\}$ by summing over all possible word strings *W* and all possible filters $\mathbf{h}_1$ and $\mathbf{h}_2$:

$$\begin{aligned} \hat{\mathbf{x}} &= \underset{\hat{\mathbf{x}}}{\operatorname{argmax}}\, p(\hat{\mathbf{x}} \mid \mathbf{y}_1, \mathbf{y}_2) = \underset{\hat{\mathbf{x}}}{\operatorname{argmax}} \sum_{W, \mathbf{h}_1, \mathbf{h}_2} p(\hat{\mathbf{x}}, W, \mathbf{h}_1, \mathbf{h}_2 \mid \mathbf{y}_1, \mathbf{y}_2) \\ &\approx \arg \max_{\hat{\mathbf{x}}} \max_{W} \max_{\mathbf{h}_1, \mathbf{h}_2} p(\mathbf{y}_1, \mathbf{y}_2 \mid \hat{\mathbf{x}}, \mathbf{h}_1, \mathbf{h}_2)\, p(W \mid \hat{\mathbf{x}})\, p(\mathbf{h}_1, \mathbf{h}_2) \end{aligned} \tag{11}$$

where we and used the standard Viterbi approximation, assuming the sum is dominated by the most likely word string *W* and most likely filters. If we further assume there is no additive noise, as in Figure 1, then $p(\mathbf{y}_1, \mathbf{y}_2 \mid \hat{\mathbf{x}}, \mathbf{h}_1, \mathbf{h}_2)$ is a delta

function. Furthermore, and in the absence of prior information for the filters, the approximate MAP filter estimates are

$$(\hat{\mathbf{h}}_1,\hat{\mathbf{h}}_2)=\arg\max_{\mathbf{h}_1,\mathbf{h}_2}\left\{\arg\max_{W} p(W\mid\hat{\mathbf{x}})\right\} \tag{12}$$

To use a standard HMM-based speech recognition system, we typically decompose the input signal $\hat{\mathbf{x}}$ into T frames $\hat{\mathbf{x}}^t$ of length N samples each:

$$\hat{x}^t[n]=\hat{x}[tN+n] \tag{13}$$

so that the inner term in Eq. (12) can be expressed as

$$\arg\max_{W} p(W\mid\hat{\mathbf{x}})=\prod_{t=0}^{T-1}\sum_{k=0}^{K-1}\gamma_t[k]\,p\left(k\mid\hat{\mathbf{x}}^t\right) \tag{14}$$

where $\gamma_t[k]$ is the *a posteriori* probability of frame t belonging to Gaussian k, one of K Gaussians in the HMM. Large vocabulary systems can often use on the order of 100,000 Gaussians.

The term $p(k\mid\hat{\mathbf{x}}^t)$ in Eq. (14), as used in most HMM systems, includes cepstral vectors and results in a nonlinear equation. In the next section we present an approximation that results in a mathematically tractable solution.

5. Using a VQ Codebook of LPC Vectors

The approximation used here consists in using an autoregressive (AR) model instead of cepstral model. Let's define $e_t^k[n]$ as the linear prediction (LPC) error of class k for signal $\hat{x}^t[n]$ as

$$e_t^k[n]=\sum_{i=0}^{p}a_i^k\hat{x}^t[n-i] \tag{15}$$

where $i=0,1,2,\cdots,p$ and a_i^k represents the LPC coefficients of that class k, with $a_0^k=1$. Furthermore, let's define E_t^k as the average energy of that prediction error for that frame t:

$$E_t^k=\frac{1}{N}\sum_{n=0}^{N-1}\left|\mathbf{e}_t^k\right|^2 \tag{16}$$

The probability we propose here for each class is an exponential density function of the energy of the linear prediction error:

$$p(\hat{\mathbf{x}}_t\mid k)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{E_t^k}{2\sigma^2}\right\} \tag{17}$$

In continuous density HMM systems Viterbi search is done, so that most $\gamma_t[k]$ in Eq. (14) are zero and the rest would correspond to the mixture weights of the current state. To speed up computation and avoid the search process altogether, we can approximate the summation in Eq. (14) with the maximum:

$$\begin{aligned}\sum_{k=0}^{K-1}\gamma_t[k]\,p(k\mid\hat{\mathbf{x}}^t)&\approx\arg\max_k\frac{p(\hat{\mathbf{x}}^t\mid k)\,p[k]}{p(\hat{\mathbf{x}}^t)}\\&=\arg\max_k p(\hat{\mathbf{x}}^t\mid k)\end{aligned} \tag{18}$$

where we have assumed all classes are equally likely:

$$p[k]=1/K\qquad k=1,2,\cdots,K \tag{19}$$

Inserting Eq. (18) into (14) and (12), the reconstruction filters can be obtained as

$$(\hat{\mathbf{h}}_1,\hat{\mathbf{h}}_2)=\arg\min_{\mathbf{h}_1,\mathbf{h}_2}\frac{1}{T}\sum_{t=0}^{T-1}\left\{\min_k E_t^k\right\} \tag{20}$$

where we have replaced maximization of a negative quantity by its minimization and ignored the constant terms. Normalization by T is done for ease of comparisons over different frame sizes.

The optimal filters in Eq. (20) minimize the accumulated prediction error with the closest codeword per frame.

6. Implementation Details

In this section we derive the formulae to solve Eq. (20). The autocorrelation of $\hat{x}^t[n]$ can be obtained after doing some algebraic manipulation on Eq. (8):

$$\begin{aligned}R_{\hat{x}\hat{x}}^t[i,j]&=\frac{1}{N}\sum_{n=0}^{N-1}\hat{x}^t[n-i]\hat{x}^t[n-j]\\&=\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_1[u]h_1[v]R_{11}^t[i+u,j+v]\\&+\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_1[u]h_2[v]\left(R_{12}^t[i+u,j+v]+R_{12}^t[j+u,i+v]\right)\\&+\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_2[u]h_2[v]R_{22}^t[i+u,j+v]\end{aligned} \tag{21}$$

where we have defined the cross-correlation functions as

$$R_{ij}^t[u,v]=\frac{1}{N}\sum_{n=0}^{N-1}y_i^t[n-u]y_j^t[n-v] \tag{22}$$

Note that the autocorrelation in Eq. (21) has the following symmetry properties

$$R_{ij}^t[u,v]=R_{ji}^t[v,u] \tag{23}$$

Inserting (15) into (16) and using (21) we can express E_t^k as

$$\begin{aligned}E_t^k&=\frac{1}{N}\sum_{n=0}^{N-1}\left(\sum_{i=0}^{p}a_i^k\hat{x}^t[n-i]\right)\left(\sum_{j=0}^{p}a_j^k\hat{x}^t[n-j]\right)\\&=\sum_{i=0}^{p}\sum_{j=0}^{p}a_i^k a_j^k R_{\hat{x}\hat{x}}^t[i,j]\\&=\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_1[u]h_1[v]\left\{\sum_{i=0}^{p}\sum_{j=0}^{p}a_i^k a_j^k R_{11}^t[i+u,j+v]\right\}\\&+2\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_1[u]h_2[v]\left\{\sum_{i=0}^{p}\sum_{j=0}^{p}a_i^k a_j^k R_{12}^t[i+u,j+v]\right\}\\&+\sum_{u=0}^{q-1}\sum_{v=0}^{q-1}h_2[u]h_2[v]\left\{\sum_{i=0}^{p}\sum_{j=0}^{p}a_i^k a_j^k R_{22}^t[i+u,j+v]\right\}\end{aligned} \tag{24}$$

Thus inserting Eq. (24) into (20) yields our reconstructions filters. To achieve such minimization we need an iterative algorithm (EM algorithm) that iterates between finding the best codebook indices $\hat{k}_t$ and the best $(\hat{h}_1[n],\hat{h}_2[n])$:

1. *Initialization*. Start with initial $h_1[n]$ and $h_2[n]$
2. *E-Step*. For $t=0,1,\cdots,T-1$, find the best codeword.

$$\hat{k}_t=\arg\min_k E_t^k \tag{25}$$

3. *M-step*. Find $h_1[n]$ and $h_2[n]$ that minimize the overall error energy

$$(\hat{h}_1[n],\hat{h}_2[n])=\arg\min_{h_1[n],h_2[n]}\frac{1}{T}\sum_{t=0}^{T-1}E_t^{\hat{k}_t} \tag{26}$$

4. *Convergence*. If converged stop, otherwise go to 2.

Since Eq. (24) given E_t^k is quadratic in $h_1[n]$ and $h_2[n]$, the optimal filters can be obtained by taking the derivative and equating to 0. If all parameters are free, the trivial solution is $h_1[n] = h_2[n] = 0 \quad \forall n$, because we did not use σ^2 in Eq. (17). To avoid that, we set $h_1[0] = 1$ and solved for the remaining coefficients. It results in the following set of 2q-1 linear equations:

$$\sum_{u=0}^{q-1} h_1[u] b_{11}[u,v] + \sum_{u=0}^{q-1} h_2[u] b_{21}[v,u] = 0 \qquad v = 1, 2, \cdots, q-1 \quad (27)$$

$$\sum_{u=0}^{q-1} h_1[u] b_{21}[u,v] + \sum_{u=0}^{q-1} h_2[u] b_{22}[u,v] = 0 \qquad v = 0, 1, \cdots, q-1 \quad (28)$$

where

$$\begin{aligned} b_{11}[u,v] &= \sum_{t=t_0}^{T-1} \sum_{i=0}^{p} \sum_{j=0}^{p} a_i^k a_j^k R_{11}^t[i+u, j+v] \\ b_{21}[u,v] &= \sum_{t=t_0}^{T-1} \sum_{i=0}^{p} \sum_{j=0}^{p} a_i^k a_j^k R_{12}^t[i+u, j+v] \\ b_{22}[u,v] &= \sum_{t=t_0}^{T-1} \sum_{i=0}^{p} \sum_{j=0}^{p} a_i^k a_j^k R_{22}^t[i+u, j+v] \end{aligned} \quad (29)$$

Eq. (27) and (28) can be solved using any linear algebra package. Notice that the time index does not start at 0, but rather at t_0, because we do not have samples of $y_1[n]$ and $y_2[n]$ available for $n < 0$.

7. Experimental Results

The codebook of LPC vectors was derived from 35000 utterances of the Wall Street Journal corpus [12] recorded with a close-talking microphone, sampled at 16kHz, for 150 male and female speakers of North American English. LPC analysis was performed using the autocorrelation method every 10 ms. The LPC coefficients are transformed to Line Spectral Frequencies (LSF) from which a codebook is computed using the Lloyd algorithm. The size of the dictionary used was 256 entries, though our experience has been that reasonable results may be obtained for codebooks of as little as16 entries.

The algorithm was evaluated with real recordings, not digitally mixed, made by Te-Won Lee [9]. Two sets of recordings with two far field microphones in a normal office room at a sampling rate 16kHz were used:

- *Speech in Music*. A speaker has been recorded with two far field microphones in a normal office room with loud music in the background. The distance between the speaker, cassette player and the microphones is about 60cm in a square ordering.
- *Cocktail Party Effect*. Two Speakers have been recorded speaking simultaneously. Speaker 1 says the digits from one to ten in English and speaker 2 counts at the same time the digits in Spanish. The distance between the speakers and the microphones is about 60cm in a square ordering.

For these recordings, it has been observed that starting with initial estimates $h_1[n] = \delta[n]$ and $h_2[n] = 0$, it takes 2 or 3 iterations to reach convergence. The reconstructed waveforms [1] have been compared with the reconstructed waveforms using other four ICA variants [2][10][13][15] through informal listening tests done with 5 subjects. All listeners agreed the proposed algorithm sounded better than the separated signals published by other researchers. Listeners noted that the proposed algorithm had less interference, less reverberation and less of the *whitening* effect of other techniques that cannot distinguish between frequency response of the filter or sources.

8. Conclusions and Future Work

We have introduced an algorithm to separate speech from interfering sources using two microphones and linear filtering. The standard model used in ICA is extended to use a mixture autoregressive model for speech. Results on real recordings show more separation than with previously existing ICA algorithms.

Future work includes modifying the model to account for additive noise or inaccuracies in the modeling due to the use of FIR filter instead of IIR, using prior knowledge about the filters and using a mixture Gaussian model instead of a VQ model. We'll also explore adaptive filtering implementations instead of block processing for a causal real-time implementation, and to relax the assumption that the sources must be spatially stationary. We'll explore the use of a speech recognition system as outlined in Section 4.

REFERENCES

[1] A. Acero. Audio separation examples. Aug 2000. http://research.microsoft.com/~alexac/audio/.

[2] Jörn Anemüller, "Correlated modulation: a criterion for blind source separation", *Joint meeting of the Acoustical Society of America and the European Acoustics Association*, Berlin, Germany, March 14-19, 1999

[3] H. Attias. "Independent Factor Analysis". *Neural Computation* 11, 803-851, 1998.

[4] A. J. Bell and T. J. Sejnowski. "An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6), 1129-1159.

[5] A. A. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[6] J. F. Cardoso. "Infomax and Maximum Likelihood for Blind Source Separation". *IEEE Signal Processing Letters* 4, 112-114, 1997.

[7] J. Cardoso. "Blind Signal Separation: Statistical Principles". *Proc. of the IEEE*, Oct, 1998.

[8] P. Comon. "Independent Component Analysis: A new Concept". *Signal Processing*, 36, 287-314, 1994.

[9] T. W. Lee. "Examples of Blind Source Separation of Recorded Speech and Music Signals". http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html

[10] T. W. Lee, A. Ziehe, R. Orglmeister and T.J. Sejnowski. "Combining Time-Delayed Decorrelation and ICA Towards Solving the Cocktail Party Effect". *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. May 1998, Seattle, Vol 2, pp. 1249-1252.

[11] T. W. Lee. *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, 1998.

[12] Linguistic Data Consortium. Wall Street Journal 1 Corpus August 1993.

[13] L. Parra and C. Spence. **"**Blind Source Separation based on Multiple Decorrelations". *IEEE Trans. on Speech and Audio Processing* pp. 320-327, May 2000.

[14] J. Platt and F. Faggin. "Networks for the Separation of Sources that are Superimposed and Delayed" in *Advances in Neural Information Processing Systems* 4, pp. 730-737., 1992.

[15] H. C. Wu and J. Principe. "Simultaneous Diagonalization in the Frequency Domain for Source Separation", *ICA '99*. France.