A BLIND APPROACH TO JOINT NOISE AND ACOUSTIC ECHO CANCELLATION

Siow Yong Low and Sven Nordholm

Western Australian Telecommunications Research Institute (WATRI) *,
Crawley, WA 6009, Australia

ABSTRACT

This paper introduces a new scheme which combines the popular blind signal separation (BSS) and a post-processor to jointly suppress noise and acoustic echo. The new L element structure uses the BSS as a front-end processor to spatially extract the target signal from the interference (noise and echo). Statistical measures are then employed to select the target signal dominant signal from the BSS outputs. The remaining $L-1$ BSS outputs (noise and echo dominant) and the existing far-end line echo are then used as the reference signals in an adaptive noise canceller (ANC) to temporally enhance the target signal. The novel structure bypasses the need for any a priori information whilst compensating the separation quality of the BSS temporally. Real room evaluations demonstrate the efficacy of the scheme in both noisy double-talk and non double-talk situation.

1. INTRODUCTION

Fundamentally, there are three important tasks to fulfill in hands-free communications systems, namely, noise suppression, room reverberation suppression and acoustic echo cancellation of the hands-free loudspeaker. Indeed, the challenge to achieve all of the mentioned tasks is evident from the fact that each of the criteria is an intensive research area itself. For instance, noise suppression techniques have been widely studied over the years, ranging from the single channel method to the more popular multichannel solutions [1]. The popularity of the multichannel systems is attributed to the additional dimension called spatial diversity which can be steered by electronic means [1]. In other words, given the location of the target signal, a small number of microphones can be arranged in space such that it spatially allows the target signal to be passed whilst rejecting sources from other directions.

Nevertheless, beamforming based methods require a priori knowledge about the array geometry and the source location. A promising alternative to beamforming is blind signal separation (BSS) [2]. With BSS, all the a priori information needed by conventional beamforming is not required at all. A direct consequence of that is the uncoupling of the disastrous steering vector errors (parametric to non-parametric). Here, the BSS attempts to recover the unobserved sources from several observed mixtures by using independence as the adaptation criteria. In speech enhancement, however, there is usually only one source of interest in a noisy (or multiple noise sources) environment. Under such underdetermined (more sources than sensors) situation, standard BSS may not perform satisfactorily and there is no information as to which BSS outputs is the desired signal.

This paper targets the mentioned problems by introducing a new scheme which incorporates the BSS and the suppression capability of an adaptive noise canceller (ANC) into an efficient speech enhancement scheme. Unlike standard BSS techniques, this structure recovers/enhances a specific speech signal (even under the influence of the hands-free loudspeaker) that is spatially closest to the array. In an effort to address the problems of acoustic feedback in hands-free communication systems, an acoustic echo canceller is also embedded in the novel structure. Since the overall structure is "*blind*" in nature, the proposed scheme can handle double-talk situation just like the BSS. In summary, the new structure has the following features,

- no array geometry and source localisation,
- no voice activity detector (VAD),
- no assumptions about the cumulative densities of the signals,
- handles double-talk situation and performs joint noise and acoustic echo cancellation.

Evaluations in a real room hands-free situation show that the structure is capable in both noisy non double-talk and double talk scenarios with noise and echo suppressions up to 20 dB.

2. THE PROPOSED STRUCTURE

2.1. Overview

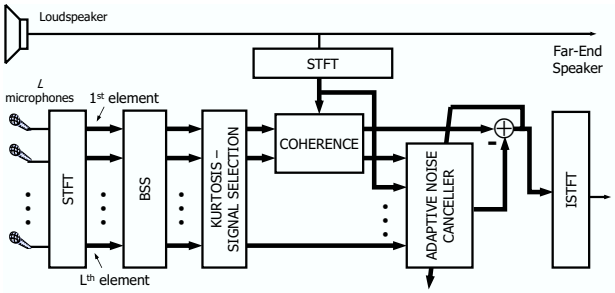

Fig. 1. *The proposed joint noise and acoustic echo cancellation processor with L microphones.*

Figure 1 shows the block diagram of the proposed structure. Essentially, the BSS acts as a front-end processor to separate the target signal from the interference (e.g. acoustic echo, ambient noise or babble) using the L observations. There is, however, a fundamental limitation in the separation quality of the BSS. This

*WATRI is a joint venture between Curtin University of Technology and the University of Western Australia. The work has also been sponsored by the Australian Research Council (ARC) under grant no. DP0451111.

is due to the multipath/reverberant environment [3] and the underdetermined situation in the real world. A straightforward way to overcome this limitation is to employ post-processing [4, 5]. In this paper, we use an ANC to refine the desired output and extend it to jointly perform acoustic echo cancellation. Also, a statistical measure is incorporated in the system to provide additional information for the BSS to distinguish the target signal from its L outputs.

Consider a hands-free scenario whereby the target signal is under the influence of both the noise and acoustic echo. Assuming that the BSS algorithm converges, the separation process will yield two speech dominant outputs i.e. target signal and far-end feedback (the remaining $L-2$ are noise dominant output(s), assuming $L \geq 3$). Following that, the BSS outputs are then ranked according to their respective kurtosis values (speech signals have higher kurtosis value than noise). With this in mind, the top two ranked outputs will therefore be the target dominant and the echo dominant signals. The coherence of both the signals are then computed against the far-end line echo to ascertain which is the target signal dominant output. Naturally, the echo dominant signal will be more coherent to the line echo compared to the target signal dominant. Thus, the signal which yields lower coherence will be the speech dominant signal. Finally, all of the other $L-1$ BSS outputs and the far-end line echo itself will serve as the references for the ANC (see Figure 1).

The motivation for the ANC stage comes from the fact that temporal diversity is not fully exploited by the BSS [5]. Having said so, the ANC will further enhance the speech dominant output by cancelling components that are temporally correlated with its references. Further, the additional reference provided by the far-end line echo will provide extra temporal diversity for the ANC to efficiently cancel the remaining far-end echo. In other words, the post-processing stage (ANC) effectively compensates for the residue noise as well as the acoustic echo in the BSS target signal dominant output. The new structure solves the BSS outputs indeterminacy (given L BSS outputs, which is the desired one?) and effectively transforms BSS into "hands-free systems compliant" by selectively enhancing only the desired signal through the ANC post-processing.

2.2. Blind Signal Separation (BSS)

Let us consider a convolutive mixture of N sources (where $L \geq N$), the observed signal vector $\mathbf{x}(t) = [x_1(t), \cdots, x_L(t)]^T$, at each of the sensors is

$$\mathbf{x}(t) = \sum_{p=0}^{P-1} \mathbf{G}(p)\mathbf{s}(t-p) \tag{1}$$

where $\mathbf{s}(t) = [s_1(t), \cdots, s_N(t)]^T$ is the N-source vector, $\mathbf{G}(p)$ is a $L \times N$ mixing matrix, P is the length of the impulse response from the nth source to the lth sensor and $(\cdot)^T$ denotes transposition. The task at hand is to find an unmixing matrix $\mathbf{W}(p)$ $(N \times L)$ of length P to recover the sources (up to an arbitrary scaling and permutation) using only the observed L mixtures.

One way to solve the problem is to perform the separation in the frequency domain [3]. By doing so, the problem becomes an instantaneous mixture for each of the frequency bin. The time domain received data $\mathbf{x}(t)$ can be transformed into the frequency domain by using a Ω-point windowed DFT and assuming that $\Omega \gg P$, a linear convolution can be approximated by a circular convolution [6]. Therefore Eqn. (1) can be rewritten in the frequency domain as

$$\mathbf{x}(\omega, k) = \mathbf{H}(\omega)\mathbf{s}(\omega, k), \tag{2}$$

where $\mathbf{x}(\omega, k)$, $\mathbf{H}(\omega)$ and $\mathbf{s}(\omega, k)$ are the transformed representations of the observations, mixing matrix and source signals respectively. The unmixing model can be written as

$$\mathbf{y}(\omega, k) = \mathbf{W}(\omega)\mathbf{x}(\omega, k), \tag{3}$$

where $\mathbf{y}(\omega, k)$ is the frequency representations of the estimated source signals vector up to a scaling and permutation ambiguities. Here, the unmixing matrix $\mathbf{W}(\omega)$ is determined such that the elements in the estimated sources $\mathbf{y}(\omega, k)$ are as statistically independent from each other as possible. There are largely two types of BSS approaches namely, second order based BSS [6] and higher order [2]. In this paper, we choose to employ the second order decorrelation by exploiting the non-stationarity of the signals.

As explained in [6], diagonalization of single time cross correlation is insufficient to solve for $\mathbf{W}(\omega)$. However, with non-stationarity, additional information can be obtained at separated time intervals. To achieve that, the covariance matrix $\mathbf{R}_x(\omega, m)$ of the received data can be estimated for the M number of intervals as

$$\mathbf{R}_x(\omega, m) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{x}(\omega, mK+k)\mathbf{x}^H(\omega, mK+k), \tag{4}$$

where $m = 0, \cdots, M-1$, k is the index for the intervals to estimate the cross covariance matrix and $(\cdot)^H$ denotes Hermitian transposition. The achieve separation, the M number of covariance matrices in Eqn. (4) are diagonalized as,

$$\mathbf{\Lambda}_s(\omega, m) = \mathbf{W}(\omega)[\mathbf{R}_x(\omega, m)]\mathbf{W}^H(\omega). \tag{5}$$

Following the approach in [6], the solution to Eqn. (5) can be obtained by using a least squares estimate as

$$\hat{\mathbf{W}}(\omega) = \arg\min_{\mathbf{W}(\omega)} \sum_{m=0}^{M-1} \parallel \mathbf{E}(\omega, m) \parallel_F^2, \tag{6}$$

where $\parallel \cdot \parallel_F^2$ is the squared Frobenius norm and the error function is $\mathbf{E}(\omega, m) = \mathbf{W}(\omega)[\mathbf{R}_x(\omega, m)]\mathbf{W}^H(\omega) - \mathbf{\Lambda}_s(\omega, m)$.

The least squares solution in Eqn. (6) can be found by using the gradient descent algorithm as follows, $\mathbf{W}^{(n+1)}(\omega)$

$$= \mathbf{W}^{(n)}(\omega) - \mu(\omega) \frac{\partial}{\partial \mathbf{W}^{(n)*}(\omega)} \left\{ \sum_{m=0}^{M-1} \parallel \mathbf{E}^{(n)}(\omega, m) \parallel_F^2 \right\} \tag{7}$$

where $(\cdot)^*$ is the conjugation operator and $\mu(\omega)$ is the step size. However, the estimation of the frequency domain unmixing weights $\mathbf{W}(\omega)$, leads to arbitrary permutation of each frequency bin. One way to solve this problem is to impose a constraint on the time-domain filter size of the unmixing weights, D such that $\mathbf{W}(\tau) = 0, \tau > D \ll \Omega$. As demonstrated in [6], the constraint couples the otherwise independent frequencies, which provides a continuity of the spectra, hence effectively solving the permutation problem.

2.3. Target Signal Selection Strategy

Prior to the post-processing stage, the BSS outputs must be correctly channelled such that the target signal dominant output will be the input for the ANC and the remaining $L-1$ outputs will be the references. To achieve that, we propose to use the kurtosis. The kurtosis is a quantitative measure of non-gaussianity of a signal. A smaller value of kurtosis indicates that the distribution tends towards gaussian and a higher value kurtosis indicates that

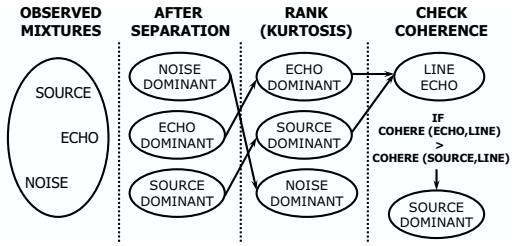

Fig. 2. *The desired signal selection strategy.*

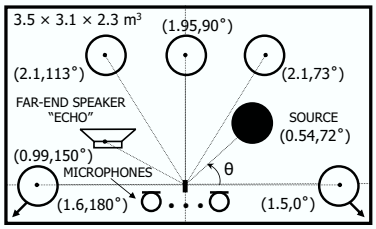

Fig. 3. *The hands-free experimental layout: the solid circle is the target signal and the hollow circles are the interference.*

the distribution tends towards supergaussian. Since speech signal has a Laplacian distribution, it belongs to the supergaussian case, which has a positive kurtosis value. This means that the speech dominant signal from the BSS will have a higher kurtosis value compared to noise dominant signal [5]. To rank the output signals according to the kurtosis, we propose to calculate the mean of the normalized kurtosis of the lth output, for all the Ω frequency bins,

$$\mathsf{Kur}^{(y_l(\omega))=} \sum_{\omega=0}^{\Omega-1} \frac{\mathsf{E}[|y_l(\omega,k)|^4]-2\mathsf{E}^2[|y_l(\omega,k)|^2]-|\mathsf{E}^2[(y_l(\omega,k))^2]|}{\sigma^4_{\mathbf{y}_l(\omega)}}. \quad (8)$$

$y_l(\omega,k)$ is one of the outputs from the BSS, $\sigma^2_{y_l(\omega)}$ is the variance of $y_l(\omega,k)$ and $|\cdot|$ denotes the absolute value operator.

However, under the presence of both the target signal and the acoustic echo, the kurtosis will not be able to function as desired since both signals are of similar distributions (comparable kurtosis value). To solve the problem, we make use of the far-end line echo by computing the coherence between the top two kurtosis ranked signals i.e. target signal dominant and echo dominant outputs respectively. Needless to say, the echo dominant BSS output will be more coherent to the far-end line echo and the target signal dominant output can be easily singled out (see Figure 2). Here, the coherence is calculated as

$$\mathsf{Coh}(y_l, line) = \sum_{\omega=0}^{\Omega-1} \frac{|P_{y_l,line}(\omega)|^2}{P_{y_l}(\omega)P_{line}(\omega)}, \quad (9)$$

where $P_{y_l,line}(\omega)$ is the cross power spectra of the line echo and one of the top two ranked BSS outputs, $P_{y_l}(\omega)$ and $P_{line}(\omega)$ are the power spectra of the corresponding BSS output and line echo respectively. Figure 2 summarizes the target signal selection strategy. Notationally, the selected target signal is labelled as $y_{target}(\omega,k)$ and the remaining $L-1$ outputs as $y_{l,ref}(\omega,k)$ where $l=1,\cdots,L-1$.

2.4. Post-Processing & Acoustic Echo Cancellation

In this stage, the ANC is employed to cancel any components that are temporally correlated to its $L-1$ references (i.e. non-target signal dominant BSS, $y_{l,ref}(\omega,k)$) from the target signal dominant BSS output, $y_{target}(\omega,k)$. To incorporate acoustic echo cancellation, the far-end line signal is used as an additional reference in the ANC making it the L-th reference ($y_{L,echo}(\omega,k)$). Note that even without the line-echo, the structure has the capability to suppress the echo. However, with the additional line-echo, it provides more temporal information for the ANC to achieve a much desirable performance.

In the interest of simplicity, the following modified frequency domain leaky LMS algorithm for the frequency ω is used instead

$$\mathbf{H}(\omega,k+1) = (1-\beta)\mathbf{H}(\omega,k) + z^*(\omega,k)\mathbf{Y}_{ref}(\omega,k)f(\omega,k), \quad (10)$$

where the $LQ\times 1$ stacked reference weights are

$$\mathbf{H}(\omega,k) = [\mathbf{h}_1(\omega,k),\cdots,\mathbf{h}_{L-1}(\omega,k),\mathbf{h}_{L,echo}(\omega,k)]^T, \text{and} \quad (11)$$

$$\mathbf{h}_l(\omega,k) = [h_l(\omega,k),\cdots,h_l(\omega,k-Q+2),h_l(\omega,k-Q+1)]^T. \quad (12)$$

Similarly, the $LQ\times 1$ stacked reference signals are

$$\mathbf{Y}_{ref}(\omega,k) = [\mathbf{y}_{1,ref}(\omega,k),\cdots, \mathbf{y}_{L-1,ref}(\omega,k),\mathbf{y}_{L,echo}(\omega,k)]^T, \text{where} \quad (13)$$

$$\mathbf{y}_{l,ref}(\omega,k) = [y_{l,ref}(\omega,k),\cdots, y_{l,ref}(\omega,k-Q+2),y_{l,ref}(\omega,k-Q+1)]^T. \quad (14)$$

The non-linear function $f(\omega,k)$ is given as

$$f(\omega,k) = \frac{\gamma}{Q\hat{\sigma}_z^2(\omega,k)+\gamma\mathbf{Y}^H_{ref}(\omega,k)\mathbf{Y}_{ref}(\omega,k)}, \quad (15)$$

where the constants β and γ are the leaky factor and the step size respectively. Q is the order of the filter and $\hat{\sigma}_z^2(\omega,k)$ is a time-varying estimate of the output signal power $z(\omega,k)$ that adjusts the step size according to the target signal level. It is built upon the fact that excess MSE increases with both the step size and the target signal [5]. When this happens, the function in (15) will effectively reduce the step size. The output signal power is estimated using the square of vector norm of length Q and then exponentially averaged as

$$\hat{\sigma}_z^2(\omega,k) = (1-\lambda)\hat{\sigma}_z^2(\omega,k-1)+\lambda\|\mathbf{z}(\omega,k)\|^2, \text{where} \quad (16)$$

$$\mathbf{z}(\omega,k) = [z(\omega,k),\cdots,z(\omega,k-Q+2),z(\omega,k-Q+1)]^T, \quad (17)$$

λ is the smoothing parameter, $\|\cdot\|$ denotes the Euclidean norm and the output of the ANC is

$$z(\omega,k) = y_{target}(\omega,k) - \mathbf{H}^H(\omega,k)\mathbf{Y}_{ref}(\omega,k). \quad (18)$$

3. EXPERIMENTS AND DISCUSSIONS

The proposed speech enhancement scheme was evaluated in a real room of dimensions $3.5\times3.1\times2.3$ m^3 using a four-element linear array with a spacing of 0.04 m, sampled at 8 kHz. Two loudspeakers emitting babble noise were placed facing the front two corners of the room to create diffuseness and three other loudspeakers (also babble) were randomly placed in the middle of the room facing the array. The exact positions of the speech source (female, English), far-end loudspeaker (male, English) and interference are illustrated in Figure 3. All simulations were performed with signal to noise ratio SNR $=-0.5$ dB, signal to echo ratio SER $=0$ dB, $\Omega=512$,

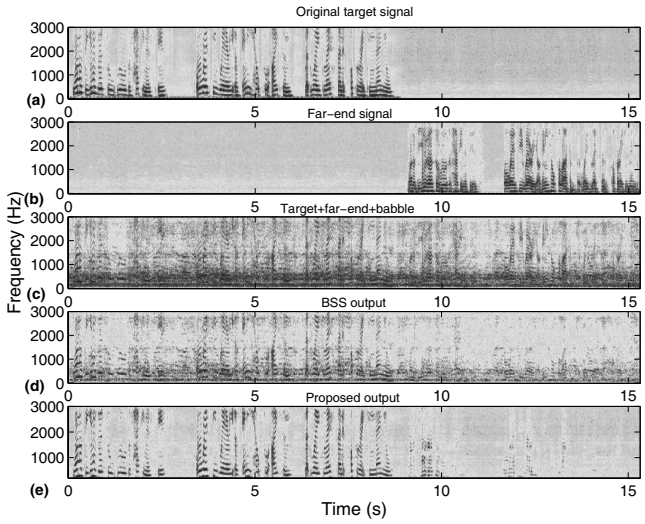

Fig. 4. *The spectrograms of (a) target signal, (b) far-end signal, (c) corrupted signal, (d) BSS output and (e) proposed output for non double-talk situation.*

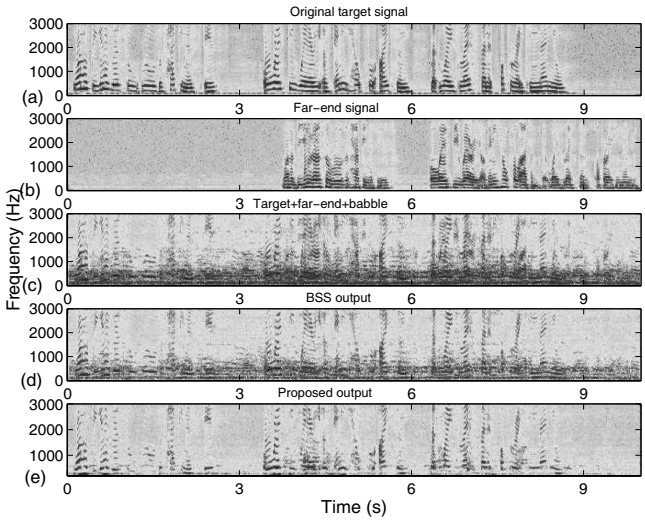

Fig. 5. *The spectrograms of (a) target signal, (b) far-end signal, (c) corrupted signal, (d) BSS output and (e) proposed output for double-talk situation.*

Operation Mode	BSS only		Proposed	
	NS	ES	NS	ES
Non double-talk	3.35 dB	6.47 dB	23.36 dB	21.33 dB
Double-talk	5.35 dB	2.51 dB	20.81 dB	16.34 dB

Table 1. *The noise (NS) and echo (ES) suppressions of the BSS and the proposed scheme for non double-talk and double-talk.*

$D = 128$, $K = 5$, and the number of taps in the adaptive filters was $Q = 4$. The parameters α, γ, λ and the leaky factor β were set to 1, 0.2, 0.99 and 10^{-6} respectively.

Figures 4 and 5 show the relevant spectrograms for the noisy non double-talk and the double-talk situations respectively. The plots reveal the superior performance of the structure in enhancing the corrupted target signal. Clearly from the plots, there is a limitation to the separation capability of the BSS, given such under-determined situations. Here, the post-processor efficiently compensates the limitation by exploiting the temporal information. To quantify the performance, the following suppression measure is calculated

$$S = 10\log_{10}\left(\frac{\sum_{\omega=0}^{\Omega-1}\hat{P}_{in}(\omega)}{\sum_{\omega=0}^{\Omega-1}\hat{P}_{out}(\omega)}\right) - 10\log_{10}(C), \qquad (19)$$

where $\hat{P}_{in}(\omega)$ and $\hat{P}_{out}(\omega)$ are the spectral power estimates of the observation and the output respectively and the constant C normalizes to the target signal gain. Table 1 presents the noise and echo suppressions compared to using BSS only. Results indicate that the post-processing achieve significant suppression improvement over BSS, yielding more than 20 dB of noise and echo suppressions.

The experiment also verifies the proposed target signal selection strategy. For the case of non double-talk situation, the kurtosis of the four BSS outputs were $\mathsf{Kur}^{(y_1(\omega))} = 15.42$, $\mathsf{Kur}^{(y_2(\omega))} = 8.07$, $\mathsf{Kur}^{(y_3(\omega))} = 16.18$ and $\mathsf{Kur}^{(y_4(\omega))} = 8.78$ respectively. Markedly, the "speech dominant" outputs (i.e. target signal and echo) were the two highest kurtosis at the first and third BSS outputs. The coherence of these outputs against the line echo were calculated to be $\mathsf{Coh}(y_1, line) = 0.10$ and $\mathsf{Coh}(y_3, line) = 0.41$. From the results, the coherence indicates that the line echo is more coherent to y_3 and this means that the target signal dominant is the first BSS output y_1. Informal listening test confirms the validity of the proposed selection method.

4. CONCLUSIONS

A novel blind joint noise and echo cancellation scheme has been presented. The structure takes advantage of the lack of a priori information of the BSS whilst boosting its suppression capability through a post-processor. A new signal selection strategy is incorporated to distinguish the target signal from noise and echo sources. The selection method efficiently singles out the target signal dominant output even under the presence of acoustic echo (speech). Results show impressive noise and echo suppressions with good target signal integrity.

5. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Process. Springer-Verlag, Berlin, 2001.

[2] S. Haykin, Ed., *Unsupervised Adaptive Filtering*, vol. 1: Blind Source Separation, Wiley & Sons, New York, 2000.

[3] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. on Speech and Audio Process.*, vol. 11, no. 2, pp. 109–116, March 2003.

[4] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual crosstalk components in blind source separation using LMS filters," *IEEE Workshop on Neural Networks for Signal Process.*, pp. 435–444, September 2002.

[5] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. on Speech and Audio Process.*, vol. 12, no. 5, pp. 539–548, September 2004.

[6] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.