Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

A HYBRID SPEECH ENHANCEMENT SYSTEM EMPLOYING BLIND SOURCE SEPARATION AND ADAPTIVE NOISE CANCELLATION

Siow Yong Low and Sven Nordholm

The University of Western Australia
Western Australian Telecommunications Research Institute(WATRI) *
Crawley, Western Australia 6009
AUSTRALIA
E-mail: {siowyong,sven}@watri.org.au
URL: www.watri.org.au

ABSTRACT

This paper presents a new hybrid speech enhancement scheme employing the blind source separation (BSS) and the adaptive noise cancellers (ANC). Since BSS primarily exploits spatial information, it has its limitation in its separation quality. To compensate for that, the ANC acts as a temporal decorrelator to perform further noise cancellation. Initially, the BSS separates the target signal from the interference using the L observed inputs. A higher order statistical method is then used to distinguish the speech dominant signal from the L BSS output(s). The remaining $L-1$ (interference dominant) outputs will serve as the reference signals for the ANC to perform further interference reduction on the speech dominant output. The novel structure bypasses a priori information such as the array geometry and source localisation needed by conventional beamforming based methods. Results show that the structure offers impressive enhancement capability with only a few microphones.

1. INTRODUCTION

Speech acquisition in adverse environments is a very challenging problem. The issue is not simply to achieve the maximum noise suppression possible, but also to maintain good signal integrity. Countless speech enhancement methods have been proposed over the years, many of which employ microphone arrays [1]. This is because microphone arrays provide the invaluable spatial diversity. A common method to perform spatial filtering (or beamforming) is to make use of the array geometry to form a beam towards the target signal. This technique has been widely studied and considerable interference suppression is reported in [1, 2]. However, beamforming based methods rely on a priori information about the array geometry and the source location. A promising alternative to beamforming is the blind source separation (BSS) [3]. With BSS, all the a priori information needed by conventional beamforming is not required at all. It uses independence as the adaptation criterion and one can view BSS as an approach that relies entirely on spatial diversity in the sense that different sensors receive different mixtures of the sources.

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

Basically, the scheme makes use of the BSS to separate the target signal from the interference. If there are L microphones and assuming that the algorithm converges then it is expected that there will be one speech dominant output and $L-1$ remaining interference dominant outputs from the BSS. With this in mind, the separated target signal can be further enhanced temporally by an adaptive noise canceller (ANC) using the interference dominant outputs as references. A higher order statistical method is proposed to distinguish the speech dominant output from the available BSS outputs. Finally, the ANC performs further temporal enhancement in the target dominant BSS output. The motivation behind the use of ANC as a post processor for the scheme can be understood by the fact that BSS exploits mainly spatial diversity (BSS only imposes spatial independence among the sources). Therefore there is room for improvement temporally. Here, the ANC performs temporal decorrelation by cancelling any components that are correlated with its references. One can view the hybrid system as a spatio-temporal processor with the BSS looking across the sensors (spatial) and the ANC looking across the time (temporal).

The major advantage of the proposed hybrid system is that it does not need any a priori information such as the array geometry model or source localisation. This offers great flexibility and avoids the deleterious effects of steering vector errors. Also, all processing is made in subbands which yields an efficient frequency spatio-temporal processing. Evaluations in both room and car hands-free situation reveal that even with a few microphones (two to five), the proposed structure manages to achieve a very good enhancement level.

2. THE PROPOSED STRUCTURE

Figure 1 illustrates the proposed hybrid system. Throughout the paper, we assume an L-element array. Each of the main blocks in the proposed structure is explained in the following sections.

2.1. Analysis & Synthesis Filter Banks

A uniform over-sampled analysis DFT filter bank is employed to decompose each of the L microphone input signals into M subbands with a decimation factor of $\frac{M}{2}$. The purpose of oversampling is to reduce the aliasing effects between the adjacent subbands and to ensure the sufficiency of data samples. The subbands are created in such a way that a prototype filter with a low pass

©2004 NORSIG 2004

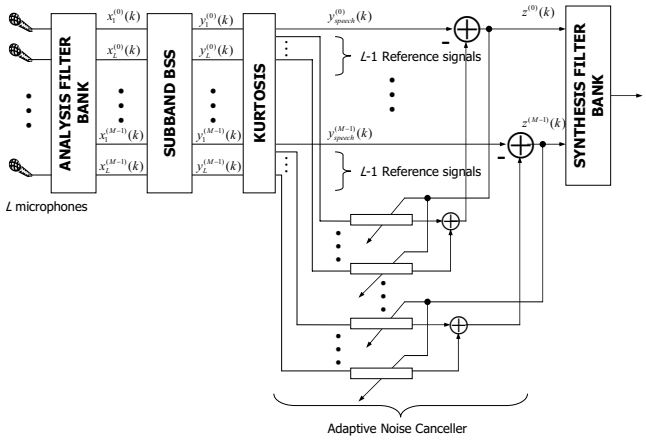

Fig. 1. *The proposed hybrid system with L microphones.*

characteristics also forms the response of the mth subband. In this case, the prototype filter is designed using a Hamming window with a cut off frequency $\frac{\pi}{M}$. Likewise, a synthesis filter bank is used to reconstruct the subband signals into fullband representation. Both filter banks are designed with minimum transformation and reconstruction aliasing effects.

2.2. Blind Source Separation

The goal of blind source separation is to recover sources from the observed mixtures using only the assumption of statistical independence among the sources [3]. The simplest BSS assumes an instantaneous mixing case, meaning that no time delays are present. In this case, N independent signals represented by the $N \times 1$ vector $\mathbf{s}(n) = [s_1(n) \cdots s_N(n)]^T$ are linearly mixed such that the observation of L number of mixtures is $x_i(n) = \sum_{j=1}^{N} h_{ij} s_j(n)$, for $i = 1, \cdots, L$ where h_{ij} is a scalar and $[.]^T$ denotes transposition. Without loss of generality, the observation signals can be expressed as

$$\mathbf{x}(n) = \mathbf{H}_{inst}\ \mathbf{s}(n), \tag{1}$$

where $\mathbf{x}(n)$ is a $N \times 1$ vector and $\mathbf{H}_{inst}$ is a $L \times N$ matrix containing the mixture coefficients. The objective here is to identify the inverse of the mixing matrix so that the sources can be recovered[1]. Much success has been reported using this simple model [4]. However, due to the multipath/reverberant environment, the sources are convolutively mixed as

$$\mathbf{x}(n) = \mathbf{H}_{conv} * \mathbf{s}(n), \tag{2}$$

where $\mathbf{H}_{conv}$ is a $L \times N$ mixing filter matrix in which each element of the matrix is a finite impulse response (FIR) filter and $*$ denotes the convolutive operator. The complicated inversion of the FIR matrix (Eqn. 2) can be bypassed by transforming the problem into the frequency domain [5, 6]. By doing so, the problem is then elegantly reverted to the simple instantaneous case. Nevertheless, the number of subbands in the filter bank must be sufficiently large for the convolutive mixture to be accurately modelled as instantaneous mixing in each of the subbands. Thus, equation (2) simplifies to

$$\mathbf{x}^{(m)}(k) = \mathbf{H}^{(m)}\mathbf{s}^{(m)}(k), \tag{3}$$

where $\mathbf{x}^{(m)}(k)$ and $\mathbf{s}^{(m)}(k)$ are the mth subband transformations of $\mathbf{x}(n)$ and $\mathbf{s}(n)$ respectively. $\mathbf{H}^{(m)}$ is a matrix containing the elements (scalar) of the mixing filters $\mathbf{H}_{conv}$ at the mth subband and k is the sample index.

Assuming that the mth subband of the mixing matrix is invertible, then the unmixing process is

$$\mathbf{y}^{(m)}(k) = \mathbf{V}^{(m)}\mathbf{x}^{(m)}(k), \tag{4}$$

where $\mathbf{V}^{(m)}$ is the desired unmixing matrix and $\mathbf{y}^{(m)}(k) = [y_1^{(m)}(k) \cdots y_L^{(m)}(k)]^T$ is the estimated source vector. Here, the unmixing matrix is determined such that the sources in the output vector are mutually independent. To find $\mathbf{V}^{(m)}$, we employ the information maximization approach using the natural gradient update [4],

$$\Delta\mathbf{V}^{(m)} \propto \eta \left[I - 2\varphi(\mathbf{y}^{(m)}(k))(\mathbf{y}^{(m)}(k))^H \right] \mathbf{V}^{(m)}, \tag{5}$$

where $(.)^H$ is the Hermitian transpose and η is the learning factor. The non-linear function φ is central to the update function as it minimizes the mutual information among the outputs if it is matched to the input probability density function (pdf) [4]. In other words, the non-linear function should approximate as close as possible the cumulative distribution of the input. For speech signal, it is generally chosen as [5],

$$\varphi(\cdot) = \tanh\left(\Re(\cdot)\right) + \jmath \tanh\left(\Im(\cdot)\right), \tag{6}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ represent the real and imaginary parts of a complex number respectively.

Unfortunately, it is inherent in BSS to have scaling and permutation ambiguities. The scaling invariance causes different scaling for each subband and this results in spectral deformation during the reconstruction process. To resolve that, we force the determinant of the unmixing matrices to unity [5]. This effectively ensures volume conservation for every subband. The permutation invariance on the other hand, results in a serious separation performance loss. This is because if the unmixing matrices do not have the same permutation for all subbands, then the reconstructed signal will remain mixed. We avoid this problem by designing an initial value for the unmixing matrix [6]. Since the BSS acts like a beamformer, we can make a sharp null towards an arbitrary jammer direction. Thus the initial value for the mth subband is

$$\mathbf{V}^{(m)}_{initial} = \begin{bmatrix} \exp(\frac{2\pi f_s m}{M}\frac{d_1 sin\theta_1}{c}) & \cdots & \exp(\frac{2\pi f_s m}{M}\frac{d_1 sin\theta_L}{c}) \\ \vdots & \ddots & \vdots \\ \exp(\frac{2\pi f_s m}{M}\frac{d_L sin\theta_1}{c}) & \cdots & \exp(\frac{2\pi f_s m}{M}\frac{d_L sin\theta_L}{c}) \end{bmatrix}^{-1} \tag{7}$$

where f_s is the sampling frequency, $d_1, \cdots, d_L$ are the distances of the lth element from the centre of the array respectively. $\theta_1, \cdots, \theta_L$ are the angles of arrival of the sources which in this case are arbitrarily set. The outputs from the subband BSS are then fed into the temporal processing blocks for further enhancement.

2.3. Kurtosis

Since there are L outputs from the BSS, there is no telling which is speech dominant or interference dominant. To resolve that, we

[1] Note that the only information available is the mixed observation vector $\mathbf{x}(n)$.

propose to use the kurtosis. The kurtosis or the fourth order statistics is commonly used as a quantitative measure of non-gaussianity of a signal. It can be shown that the kurtosis of a Gaussian distribution is zero whereas the kurtosis of a supergaussian distribution is positive. Thus, a smaller value of kurtosis indicates that the distribution tends towards gaussian and a higher value kurtosis indicates that the distribution tends towards supergaussian. It is well known from Central Limit Theorem that the sum of several distributions will tend towards Gaussian. In practice, the interference is assumed to be spatially diffused, thus, it tends towards a Gaussian-like distribution. Since speech signal has a Laplacian distribution, it belongs to the supergaussian case which has a positive kurtosis value.

Having said so, we propose that the BSS output with the highest kurtosis value will be the speech dominant output and the remaining $L-1$ outputs will serve as reference signals for the ANC. To calculate the complex kurtosis, we calculate the mean of the kurtosis of the lth output ξ_l, for all the M subband signals as

$$\xi_l = \frac{1}{M}\sum_{m=0}^{M-1} \frac{\mathsf{E}[|y_l^{(m)}(k)|^4] - 2\mathsf{E}^2[|y_l^{(m)}(k)|^2] - |\mathsf{E}^2[(y_l^{(m)}(k))^2]|}{\sigma^4_{y_l^{(m)}(k)}}. \tag{8}$$

$\mathsf{E}[\cdot]$ denotes the statistical operator of expectation and $|\cdot|$ represents the absolute value operator. $y_l^{(m)}(k)$ is one of the outputs from the BSS and $\sigma^2_{y_l^{(m)}(k)}$ is the variance of $y_l^{(m)}(k)$. Notationally, the output that has the highest value of kurtosis will be labelled as $y^{(m)}_{speech}(k)$ and the remaining $L-1$ outputs will be denoted as $y^{(m)}_{l,ref}(k)$ with $l = 1, \cdots, L-1$.

2.4. The Adaptive Noise Canceller

The ANC is employed to cancel any components that are correlated to $y^{(m)}_{l,ref}(k)$ from $y^{(m)}_{speech}(k)$ for each of the M subbands. For ease of computation, the least mean square (LMS) algorithm is used to update the coefficients in the subband adaptive filters. Regrettably, the price to pay for the simplicity of the LMS algorithm is that its steady-state excess mean square error (MSE) increases linearly with target signal power [7]. To remedy the problem, the following modified subband leaky LMS algorithm based on [7] for the mth subband is used instead

$$\mathbf{w}_l^{(m)}(k+1) = (1-\beta)\mathbf{w}_l^{(m)}(k) + (z^{(m)*}(k)\mathbf{y}^{(m)}_{l,ref}(k) f_l^{(m)}(k), \tag{9}$$

for $l = 1, \cdots, L-1$, $(\cdot)^*$ is the conjugation operator and the lth weights are

$$\mathbf{w}_l^{(m)}(k) = [w_{l,1}^{(m)}(k) \;\; w_{l,2}^{(m)}(k) \;\; \cdots \;\; w_{l,K}^{(m)}(k)]^T. \tag{10}$$

The lth reference signal is

$$\mathbf{y}^{(m)}_{l,ref}(k) = [y^{(m)}_{l,ref}(k) \;\; y^{(m)}_{l,ref}(k-1) \;\; \cdots \;\; y^{(m)}_{l,ref}(k-K+1)]^T, \tag{11}$$

and the non-linear function $f_l^{(m)}(k)$ is given as

$$f_l(k) = \frac{\alpha}{K[\hat{\sigma}^2_{z^{(m)}}(k) + \alpha\sum_{l=1}^{L-1}\|\mathbf{y}^{(m)}_{l,ref}(k)\|^2]}. \tag{12}$$

The constants β and α are the leaky factor and the step size respectively. K is the order of the filter and $\hat{\sigma}^2_{z^{(m)}}(k)$, is a time-varying estimate of the output signal power $z^{(m)}(k)$ 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 [7]. When this happens, the function in (12) will effectively reduce the step size. Here, the output signal power is estimated using the square of vector norm of length K. This estimate is then exponentially averaged as

$$\hat{\sigma}^2_{z^{(m)}}(k) = (1-\lambda)\hat{\sigma}^2_{z^{(m)}}(k-1) + \lambda\hat{\sigma}^2_{z^{(m)}}(k), \tag{13}$$

where λ is the smoothing parameter. Quite simply, the method exploits intervals of weak and strong of the target signal like a energy detector. In addition, since the processing is made in subbands, very short filter (in the order of 1 to 5 taps) can be employed in the ANC. Finally, the output of the ANC is given by

$$z^{(m)}(k) = y^{(m)}_{(speech)}(k) - \sum_{l=1}^{L-1}\mathbf{w}_l^{(m)H}(k)\mathbf{y}^{(m)}_{l,ref}(k). \tag{14}$$

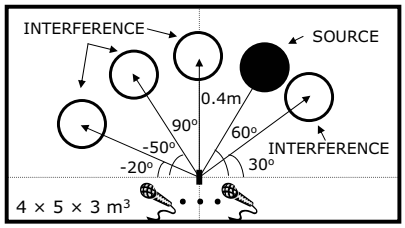

Fig. 2. *Experimental setup for the multi-babble scenario.*

3. EVALUATIONS

3.1. Car and Room Environments

The performance evaluation of the proposed structure were made in a real hands-free car situation and in a room ($4 \times 5 \times 3$ m^3). For the car environment, the source was 30 cm from the centre of the array. The array was mounted on the visor at the passenger side in a Volvo station wagon. Data was gathered on a multi-channel DAT-recorder with a sampling rate of 12 kHz and the car was moving at a constant speed of 110 km/h. As for the room environment, it was created using the image model method [8] with a reverberation time of 100 ms and the speech source was located 40 cm from the array at an angle 60°. The setup consisted of a few directional babble sources in the background as shown in Figure 2. The intention of using babble background (as opposed to white noise) was to test how robust the scheme was to interference with a characteristics similar to speech. Note also that from the experimental setup in Figure 2, the target signal was placed at an angle rather than directly in front of the array. The babble source were intentionally put close to the source to test the robustness of the structure. The inter-element distance for both environments was 5 cm. All simulations were performed with 64 subbands and the number of taps in the adaptive filters was 1 and 5 for the car and room scenarios respectively. The parameters α and the leaky factor β were set to 0.05 and 10^{-5} respectively.

3.2. Results

Figure 3 shows the spectrograms of the original target signal, the corrupted signal and the processed output for the real car environ-

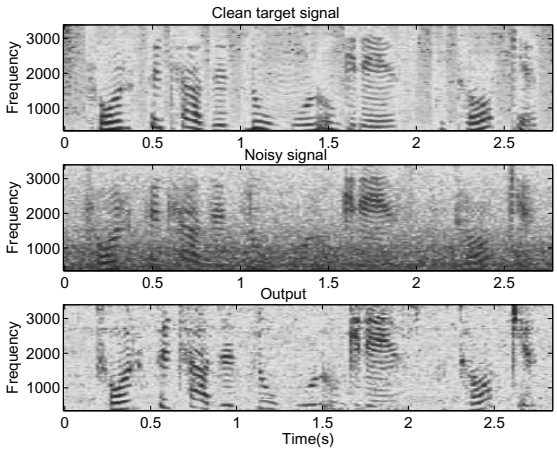

Fig. 3. *Spectrograms of the clean target signal, noisy signal and processed output for the car environment using four elements.*

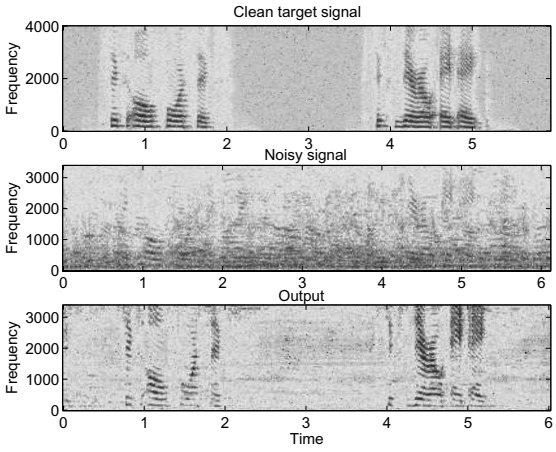

Fig. 4. *Spectrograms of the clean target signal, noisy signal and processed output for the room environment using four elements.*

ment using four elements. The signal to noise ratio (SNR) of the corrupted signal was measured to be -7 dB. The SNRs were calculated using the segments of the speech/non-speech signals. From the spectrograms, it is clear that even with such adverse condition, the noise is removed significantly whilst maintaining good target signal integrity.

The spectrograms of the original target signal, the corrupted signal and the processed output for the room environment using four elements are shown in Figure 4. The noisy target signal in this case has a SNR of 0 dB. Evidently, the spectrograms show that the "drowned" target signal in the sea of interference has been "rescued". The results demonstrate that even when the target speech signal is not in the centre of the array and in close proximity with the other babble sources, the proposed structure still manages to pick it up correctly.

No. Microphones.	Car Env.	Room Env.
2	8.7 dB	11.6 dB
3	13.4 dB	15.6 dB
4	15.5 dB	17.7 dB
5	17.2 dB	20.6 dB

Table 1. *The SNRs (dB) of the processed signal using different number of microphones in both car and room environments.*

Numerically, the SNRs of the processed output for both the car and room scenarios using different number of microphones are tabulated in Table 1. As expected, the more microphones the structure employs, the better the SNR is. However, even with as little as two microphones, the structure still manages to achieve an impressive SNR improvement of more than 10 dB. An improvement of SNR up to 20 dB is registered for the case of 5 elements. Informal listening tests suggest a very good output quality.

4. CONCLUSIONS

A novel speech enhancement scheme has been presented. The subband based design integrates the powerful BSS and a simple ANC with improved algorithm as an efficient hybrid system. As such, the hybrid system makes full use of both the spatial (BSS) and temporal (ANC) domains. The structure does not require a priori information such as the array geometry or source localisation. Really, it is just reusing the information from the BSS to achieve further enhancement in the ANC. Results show impressive interference suppression whilst maintaining good speech intelligibility.

5. REFERENCES

[1] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Jun. 1999.

[2] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for handsfree speech enhancement," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 1, pp. 885–888, 2002.

[3] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[4] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computing*, vol. 7, pp. 1129–1159, Nov. 1995.

[5] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," *Proc. IEEE Apps. of Signal Process. to Audio and Acoust.*, pp. 19–22, 1997.

[6] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H.Saruwatari, "Subband based blind source separation for convolutive mixtures of speech," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 5, pp. 509–512, Apr. 2003.

[7] J. E. Greenberg, "Modified LMS algorithms for speech processing with an adaptive noise canceller," *IEEE Trans. on Speech and Audio Process.*, vol. 6, no. 4, pp. 338–350, Jul. 1998.

[8] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *Journal of Acoust. Soc. America*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.