SPATIO-TEMPORAL PROCESSING FOR DISTANT SPEECH RECOGNITION

Siow Yong Low[†] *Roberto Togneri*[‡] *Sven Nordholm*[†]

[†]Western Australian Telecommunications Research Institute (WATRI) [*],
Crawley, WA 6009, Australia
[‡]School of Electrical and Electronic Engineering
The University of Western Australia, Crawley, WA 6009, Australia

ABSTRACT

A new subband based front-end processor for speech recognition is presented. It integrates both spatial and temporal signal processing methods to enhance noisy signals as a means to reduce the mismatch problem in speech recognition. The approach makes use of the popular blind signal separation (BSS) to spatially separate the target signal from the interference. Due to the multipath/reverberant environment, BSS has its fundamental limitation in the separation quality. To overcome that, an adaptive noise canceller (ANC) is employed to perform further interference reduction. Experimental results show that even in an adverse environment, the proposed structure improves the word recognition rate (WRR) by 70% for the connected digit recognition task.

1. INTRODUCTION

The performance of hands-free speech recognizers is highly susceptible to the surrounding environment. For instance, in a noisy and reverberant environment, the recognition rate decreases dramatically. This is typically caused by the severe mismatch of the received noisy signal to that of the "clean" training signal. One common method to solve that is by adapting the "clean" training signal similar to that of the environment. However, for mobile applications, this requires re-adaptation when the environment changes. A far more effective way to compensate for the deleterious mismatch problem is to employ microphone arrays [1, 2]. This is because microphone array or beamforming has the capability to spatially select the desired signal, thereby improving signal intelligibility and consequently less mismatch. Another popular array processing technique is the blind signal separation (BSS) [3, 4]. It offers a more flexible solution compared to the beamforming based methods in the sense that it requires no a priori information such as array geometry and source localisation. The separation algorithm relies only on the statistical independence of the sources to recover unobserved sources from several observed mixtures.

In this paper, we propose a new subband BSS with post processing for robust speech recognition. The new structure serves as a front-end processor for the speech recognition system as shown in Figure 1 and aims to minimize the commonly encountered mismatch problem. Initially, the scheme makes use of the BSS algorithm to separate the target signal from the interference. Following that, an adaptive noise canceller (ANC) is employed to further enhance the output. The motivation for the post processor (ANC) or the spectral decorrelator stems from the fact that spectral diversity is not fully exploited by the BSS. Since the processing is made in subbands, very short filter can be employed in the ANC. Connected digit speech recognition results reveal that the array processor manages to improve the recognition rate remarkably even in extremely adverse "babble" environment without the need for adaptation.

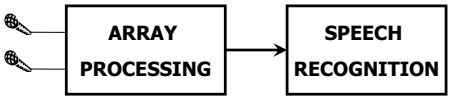

Fig. 1. *Microphone array speech recognition scheme.*

2. SYSTEM DESCRIPTION

Throughout this paper, we assume a two-element array and one source of interest in the presence of interference[1]. Figure 2 illustrates the different stages of the speech enhancement scheme. Firstly, a uniform over-sampled analysis DFT filter bank is employed to decompose each of the microphone input signals into M subbands. The purpose of over-sampling is to reduce the aliasing effects between the adjacent subbands and to ensure the sufficiency of data samples. The separation process (BSS) will then yield one output that consists of mainly the target signal plus residue of interference and the other output contains mostly interference. The kurtosis (4^{th} order statistics) of the two complex BSS outputs are then calculated to ascertain which output has the most dominant interference. This output will then serve as the reference signal for the ANC to cancel the residue of interference in the other BSS output. Finally, a synthesis filter bank is used to reconstruct the subband signals into fullband representation.

2.1. Blind Signal Separation

As the name suggests, BSS aims to recover sources from the observed mixtures using only the assumption of mutual independence among the sources [3, 4]. The simplest BSS assumes an instantaneous mixing case where 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 N number of mixtures is $x_i(n) = \sum_{j=1}^{N} a_{ij} s_j(n)$, for $i = 1, \cdots, N$ where a_{ij} is a scalar and $(.)^T$ denotes transposition. The observation signals can be

[*]WATRI is a joint venture between The University of Western Australia and Curtin University of Technology. The work has also been sponsored by ARC under grant no. A00105530.

[1]The interference in this case could be due to ambient noise or babble.

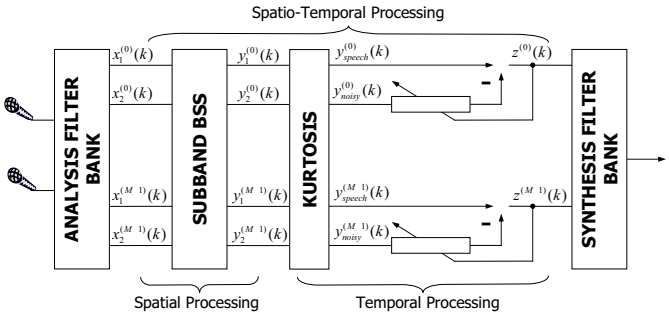

Fig. 2. *The proposed oversampled subband BSS with post processing scheme.*

represented as

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

where $\mathbf{x}(n)$ is a $N \times 1$ vector and $\mathbf{A}_{inst}$ is a $N \times N$ matrix containing the mixture coefficients. The task at hand is to identify the inverse of the mixing matrix $\mathbf{A}_{inst}$ so that the sources can be recovered. Many methods have been reported with great success to solve the instantaneous model [5, 6]. However, instantaneous model is not practical in real life. Due to the multipath/reverberant environment, the sources are convolutively mixed as

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

where $\mathbf{A}_{conv}$ is a $N \times N$ 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 [7, 8]. By doing so, the problem is then elegantly transformed 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

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

where $\mathbf{x}^{(m)}(k)$ and $\mathbf{s}^{(m)}(k)$ are the mth subband transformations of $\mathbf{x}(n)$ and $\mathbf{s}(n)$ respectively. $\mathbf{A}^{(m)}$ is a matrix containing the elements (scalar) of the mixing filters $\mathbf{A}_{conv}$ at the mth subband and k is the subband 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 $\mathbf{y}^{(m)}(k)$ 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 [5],

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

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) [5]. In other words, the non-linear function should approximate as close as possible the cumulative distribution of the input. For speech signals, it is generally chosen as [7]

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

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 solve that, we force the determinant of the unmixing matrices to unity [7]. This effectively ensures volume conservation for every subband. The permutation problem on the other hand is avoided by making use of the directivity pattern [9]. Theoretically, the directional nulls only exist in the direction of the sources for all subbands. Therefore the permutation problem is solved by flipping the rows of $\mathbf{W}^{(m)}$ if there is any deviation from the null location. In other words, the set of unmixing matrix is grouped in such a way that each row of $\mathbf{W}^{(m)}$ (for all subbands) will have a common null. In this paper, we only consider one source in a noisy environment, so the grouping is done such that there are two common nulls pointing towards the source and arbitrary interference source respectively.

2.2. Kurtosis

If the separation algorithm in Section 2.1 converges, then there will be two outputs from the BSS. Needless to say, one will be target signal dominant and the other will be the interference dominant. However, there is no telling which output is which. To resolve the issue, we propose the use of 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 (speech signal has a Laplacian distribution which belongs to the supergaussian case). Therefore if the kurtosis of any of the two BSS outputs has a smaller value, then that particular output will serve as the reference signal for the ANC. This is because a smaller kurtosis indicates that the distribution tends towards gaussian.

To calculate the complex kurtosis, we propose to calculate the mean of the kurtosis ξ for all the M subband signals given as

$$\frac{1}{M}\sum_{m=0}^{M-1}\frac{\mathrm{E}[|y^{(m)}(k)|^4] - 2\mathrm{E}^2[|y^{(m)}(k)|^2] - |\mathrm{E}^2[(y^{(m)}(k))^2]|}{\sigma^4_{y^{(m)}(k)}}. \tag{7}$$

$\mathrm{E}[\cdot]$ denotes the statistical operator of expectation and $|\cdot|$ represents the absolute value operator. $y^{(m)}(k)$ is one of the outputs from the BSS and $\sigma^2_{y^{(m)}(k)}$ is the variance of $y^{(m)}(k)$. From here onwards, the output from the subband BSS with a smaller value of ξ will be labelled $y^{(m)}_{ref}(k)$ and $y^{(m)}_{speech}(k)$ for the other.

2.3. The Adaptive Noise Canceller

The ANC is employed to cancel any components that are correlated to $y^{(m)}_{ref}(k)$ 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. The price to pay for the simplicity of the LMS algorithm, however,is

that its steady-state excess mean square error (MSE) increases linearly with target signal power [10]. To mitigate the problem, the following modified subband leaky LMS algorithm based on [10] for the mth subband is used instead,

$$\mathbf{w}^{(m)}(k+1) = (1-\beta)\mathbf{w}^{(m)}(k) + (z^{(m)*}(k)\mathbf{y}^{(m)}_{ref}(k)f(k), \quad (8)$$

where $(\cdot)^*$ denotes conjugation and

$$\mathbf{w}^{(m)}(k) = [w^{(m)}_1(k) \;\; w^{(m)}_2(k) \;\; \cdots \;\; w^{(m)}_K(k)]^T \quad (9)$$

$$\mathbf{y}^{(m)}_{ref}(k) = [y^{(m)}_{ref}(k) \;\; y^{(m)}_{ref}(k-1) \;\; \cdots \;\; y^{(m)}_{ref}(k-K+1)]^T \quad (10)$$

and the non-linear function $f(k)$ is given as

$$f(k) = \frac{\alpha}{K[\hat{\sigma}^2_{z^{(m)}}(k) + \alpha\|\mathbf{y}^{(m)}_{ref}(k)\|^2]}. \quad (11)$$

The constants β, α and K are the leaky factor, the step size and the order of the filter respectively. $\hat{\sigma}^2_{z^{(m)}}(k)$, on the other hand 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 [10]. When this happens, the function in (11) effectively reduces the step size. In other words, the method exploits intervals of weak/strong of the target signal just like a voice activity detector (VAD). Greenberg [10] proposes to estimate the output signal power by smoothing the instantaneous output power using a low pass filter with a time constant of approximately 10 ms. Here, we propose to continuously update the instantaneous output power in the tapped delay line with a first order auto-regressive (AR) smoothing. Specifically, the output signal power is estimated using the square of vector norm of the output signal with length K. This estimate is smoothed as,

$$\hat{\sigma}^2_{z^{(m)}}(k) = \lambda\hat{\sigma}^2_{z^{(m)}}(k) + (1-\lambda)\hat{\sigma}^2_{z^{(m)}}(k-1) \quad (12)$$

where λ is the smoothing parameter.

2.4. Speech Recognition

Evaluation of the proposed speech enhancement scheme was undertaken for the task of connected digit recognition. The speech recogniser was based on the HMM paradigm as implemented by the HTK software [11] and operates as shown by Figure 3 where the input to the system is the speech enhanced by the proposed BSS scheme. There was no adaptation of the models, only the initial models trained on the clean data were used.

It was found that for connected digit recognition best results were achieved with 15-state, 5-mixture HMMs for the individual digits (1-9, zero and "oh"), a 3-state model for the beginning and end silence regions and a single state model for the inter-word pauses. The acoustic features used were the standard 39 dimensional static, delta and delta-delta MFCC vectors [12] derived from speech data frames of duration 25 ms and a frame advance of 10 ms. To compensate for the severe mismatch conditions imposed by the residual additive and convolutional noise present in the enhanced speech, the standard cepstral mean normalisation was applied to the acoustic features.

The training data consisted of 220 isolated digit utterances (11 digits per speaker) and 70 connected digit sequences across all speakers spoken by 10 male and 10 female speakers taken from the NIST TIDIGITS database. The system was tested using 20 different connected digit sequences (5 per speaker) uttered by 4 other speakers (2 male and 2 female).

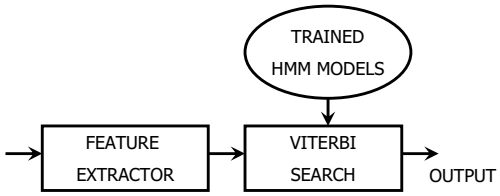

Fig. 3. *The block diagram of the speech recognition system.*

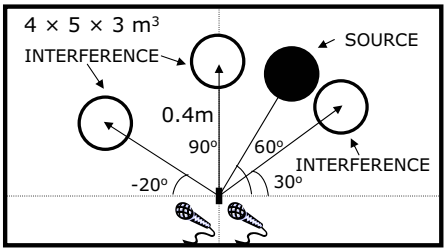

Fig. 4. *The experimental layout.*

3. RESULTS

The performance evaluation of the proposed structure was made in a room ($400 \times 500 \times 300$ cm) with a two element microphone array using the image model method [13]. The reverberation time of the room and the inter-element distance were 100 ms and 4 cm respectively. The speech source was located 40 cm from the centre of the array at an angle 60° in the presence of multiple babble sources with a signal to noise ratio (SNR) of 0 dB as shown in Figure 4. 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. The processing was made using 64 subbands with a decimation factor of 32 and the number of taps in the adaptive filters was 5. The parameters η, α, β and λ were set to 10^{-4}, 0.05, 10^{-5} and 0.99 respectively.

Figure 5 shows the original signal, the corrupted signal, the two BSS outputs and the output of the proposed system. Clearly from the plots, the separation process yields one output that consists of mainly the interference and the other predominantly speech. Numerically, the kurtosis measurements for each of the BSS outputs are 5.4 and 16.5 for Figures 5(c) and 5(d) respectively. The final output in Figure 5(e) reveals an impressive enhancement quality with most of the noisy residues removed. It is also worth mentioning that the proposed structure achieves a uniform noise suppression across the speech spectrum. This has the added advantage in reducing the mismatch problem in speech recognition.

For evaluation of the speech recognition performance, four test environments were defined: Clean, Babble, BSS, and Proposed defined as follows:

Clean same conditions as the training data.

Babble added babble sources as shown by Figure 4, without any enhancement

BSS the speech dominant output of the subband BSS stage.

Proposed the output of the proposed speech enhancement system.

The recognition accuracy results for each of the test environments is shown in Table 1. The Word Recognition Rate (WRR) was de-

	WRR	SRR
Clean	100%	100%
Babble	7.1%	0%
BSS	10.7%	0%
Proposed	76.4%	20.0%

Table 1. WRR and SRR Recognition Results

fined as the total number of digits minus the number of incorrect digit substitutions, digit deletions and digit insertions, all divided by the total number of digits. The Sentence Recognition Rate (SRR) was defined as the number of correctly transcribed digit utterances divided by the total number of utterances.

From Table 1 under matched conditions (Clean) there are no errors in the recognition. However with added babble, for the unenhanced case, there is a severe degradation in the recognition performance, with only a 7.1% WRR and 0% SRR (i.e. no utterances are correctly transcribed). Upon enhancement of the speech by the BSS stage there is a slight but noticeable improvement in the WRR to 10.7%. However, it is only with the addition of the post-processing stage proposed in this paper that we observe a significant improvement in performance by a 66% increase in the WRR from 10.7% with BSS only, to 76.4% with the proposed system. Furthermore 20% of the utterances are now correctly transcribed without any insertion, deletion or substitution errors. Given that no adaptation was applied to the HMMs and only standard cepstral mean normalisation was used, these results clearly indicate the efficacy of the proposed enhancement scheme in producing speech that is close to acquisition under matched conditions compared to the unenhanced case.

4. CONCLUSIONS

A novel spatio-temporal front-end processor for speech recognition has been presented. The proposed system employs a subband blind signal separation algorithm with a post processor. Essentially, the scheme recycles the outputs from the BSS using the ANC to achieve further enhancement. One can view the structure as an efficient combination of both spatial (BSS) and temporal (ANC) processing. Evaluation results with connected digit speech recognition show that there is a remarkable improvement in the WRR without the need for adaptation.

5. REFERENCES

[1] J. Kleban and Y. Gong, "HMM adaptation and microphone array processing for distant speech recognition," *IEEE Int. Conf. on Acoust. Speech and Signal Process.*, vol. 3, pp. 1411–1414, 2000.

[2] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Experiments of HMM adaptation for hands-free connected digit recognition," *IEEE Int. Conf. on Acoust. Speech and Signal Process.*, vol. 1, pp. 473–476, 1998.

[3] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

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

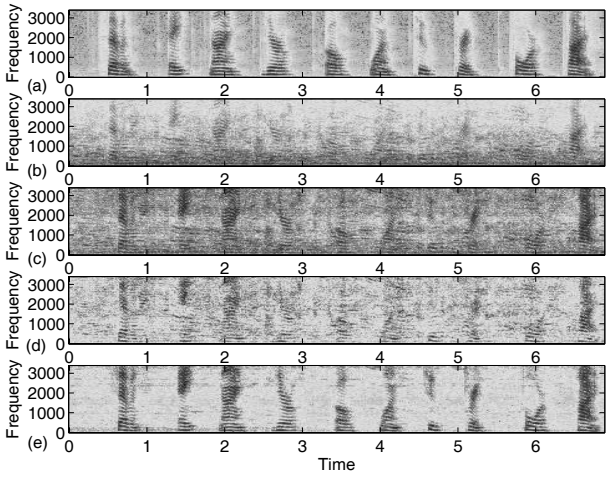

Fig. 5. *The spectrograms of (a) the original speech sequence, (b) the corrupted signal, (c) BSS first output, (d) BSS second output and (e) Processed output.*

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

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Adaptive Learning Systems for Signal Processing, Communication and Control. John Wiley & Sons, 2001.

[7] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," *Proc. IEEE Applications of Signal Processing to Audio and Acoustics*, pp. 19–22, 1997.

[8] 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.

[9] H. Saruwatari, T. Kawamura, and K. Shikano, "Fast-convergence al-gorithm for ica-based blind source separation using array signal processing," *Proc. 11th IEEE Statistical Signal Processing Workshop*, pp. 464–467, 2001.

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

[11] T. Hain, P.C. Woodland, G. Evermann, and D. Povey, "New features in the CU-HTK system for transcription of conversational telephone speech," *IEEE Int. Conf. on Acoust. Speech and Signal Process.*, vol. 1, pp. 57–60, 2001.

[12] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, vol. 13, pp. 45–57, Sep. 1996.

[13] 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.