SPEECH SIGNAL EXTRACTION UTILIZING PCA-ICA ALGORITHM WITH A NON-UNIFORM SPACING MICROPHONE ARRAY

Sven Nordholm and Siow Yong Low

Western Australian Telecommunications Research Institute (WATRI†)
39 Fairway, Crawley, WA 6100, Australia

ABSTRACT

Speech signal extraction is becoming more and more important as evidently displayed by its numerous applications such as mobile phones, conference equipments and surveillance. This paper presents a blind method to enhance a speech source of interest in noisy environments. The proposed technique consists of the principal component analysis (PCA) and the independent component analysis (ICA) to extract the speech signal. In an effort to overcome the small phase resolution due to the constraint on the inter-element distance, a non-uniform spacing PCA-ICA algorithm is suggested. By utilizing a different inter-element distance processing on each pair of microphones in a multistage fashion, a better separation is achieved. Results show better separation performance for the proposed method compared to the uniformly spaced microphone array.

1. INTRODUCTION

Speech signal extraction as the name implies, aims to extract speech signal (of interest) in adverse environments. Generally speaking, there are largely two approaches to perform the extraction or enhancement process, i.e., single channel and multi-channel techniques. Whilst single channel approach is simpler and "hardware appealing" compared to its multi-channel counterpart, there are however fundamental limitations as to how much it can really achieve, particularly in the presence of non-stationary noise. Multi-channel approach on the other hand, offers spatial diversity, which can be exploited to spatially pass or reject sources coming from a specific direction in space [1, 2]. However, most beamforming based approaches require information about array geometry and source localization.

Of late, blind signal separation (BSS) has emerged as an efficient tool to perform speech signal extraction [3, 4, 5]. Strictly, BSS as the name implies, is a technique for estimating original sources from observed mixed signals without information about the array geometry and source localization [6]. It is precisely this "blindness" to information needed by conventional beamformers, which makes BSS a very appealing tool to perform speech extraction. For instance, the uncoupling of geometric model results in the uncoupling of the disastrous steering vector errors i.e., no geometry model mismatch. In [7], it was shown that BSS is in fact similar to beamforming. This is because BSS forms spatial nulls to suppress sources. As pointed out in [8], spatial nulls can be best formed if phase differences are well exploited. This means that appropriate inter-element spacing should be set for certain frequency range.

†WATRI is a joint institute between Curtin University of Technology and the University of Western Australia. Research is also partially supported by the National ICT Australia (NICTA), funded by the Australian Research Council (ARC).

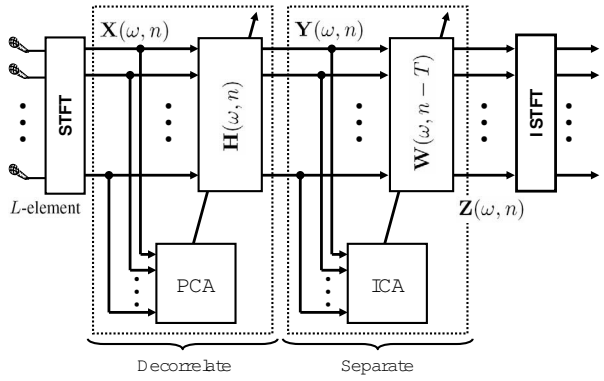

Fig. 1. The structure of the proposed combined PCA-ICA algorithm.

In an effort to overcome the poor phase resolution in the low frequency range, a non-uniform spacing principal component analysis - independent component analysis (PCA-ICA) algorithm is suggested. By utilizing a different inter-element distance processing on each pair of microphones in a multistage fashion, a better separation is achieved. The proposed method does not require bandpass filtering to split the observations into several subsystems or frequency groups. Instead, a simple sequential or multistage processing method is proposed, i.e., a two-stage BSS. The first stage involves separation in pairs of microphones with different spacing. As such, separation at lower frequency range and upper frequency range will benefit from a wider spacing and a smaller spacing, respectively. To recombine, the kurtosis selection strategy [3] and a second separation stage are used.

In Section 2, the PCA-ICA algorithm and its extension to an online system are described. Additionally, the application of BSS as a signal extraction tool is reviewed and explained. Following that, the pair-wise multi-resolution separation scheme is included, along with a comparison of the non-uniform spacing and uniform spacing array. Results show better noise suppression is achieved by using the non-uniform spacing compared to the uniform array.

2. PCA-ICA ALGORITHM

2.1. Introduction

Figure 1 shows the block diagram of the proposed L-element PCA-ICA algorithm. The frequency domain based separation algorithm consists of two main parts. First, it decorrelates the data through PCA and subsequently separates the data via ICA. The PCA and ICA algorithms complement one another because if only the PCA is used, no separation can be accomplished, since the PCA only decorrelates

1-4244-0469-X/06/$20.00

$$\mathbf{H}(\omega,n)=\left\{\alpha^{-1}\mathbf{H}^H(\omega,n-1)\mathbf{H}(\omega,n-1)\left[\mathbf{I}-\frac{\mathbf{X}(\omega,n)\mathbf{X}^H(\omega,n)\mathbf{H}^H(\omega,n-1)\mathbf{H}(\omega,n-1)}{(\alpha^{-1}-1)^{-1}+\mathbf{X}^H(\omega,n)\mathbf{H}^H(\omega,n-1)\mathbf{H}(\omega,n-1)\mathbf{X}(\omega,n)}\right]\right\}^{\frac{1}{2}}, \quad (5)$$

the data (uncorrelated does not mean independence). Moreover, if ICA is applied alone, the problem becomes difficult for the ICA to solve. In other words, PCA first reduces the dimension of the problem and this gives a good initial conditions for ICA to perform the separation [9].

2.2. Principal Component Analysis (PCA)

In this paper, we suggest a recursive PCA to perform the decorrelation [10, 11]. The cost function, $\mathbf{E}(\omega,n)$ at instant, n and frequency, ω can then be written as

$$\mathbf{E}(\omega,n)=\mathbf{H}(\omega,n)\mathbf{R}_x(\omega,n)\mathbf{H}^H(\omega,n)-\Lambda_s(\omega,n), \quad (6)$$

where $\mathbf{H}(\omega,n)$ is the decorrelation matrix and $\mathbf{R}_x(\omega,n)$ is the covariance matrix of the received observations. The diagonal matrix, Λ_s is the sources' power and the subscript $(\cdot)^H$ denotes the Hermitian transposition. The objective is find $\mathbf{H}(\omega,n)$ such that the following is minimized

$$\hat{\mathbf{H}}(\omega,n)=\arg\min_{\mathbf{H}(\omega,n)} \parallel \mathbf{E}(\omega,n) \parallel_{\mathcal{F}}^2, \quad (7)$$

where $\parallel\cdot\parallel_{\mathcal{F}}^2$ denotes the squared Frobenius norm. To simplify derivation, we assume that the field is homogeneous, the matrix $\mathbf{H}(\omega,n)$ is full rank and that all sources have the same power, i.e., $\Lambda_s(\omega,n)=S(\omega,n)\mathbf{I}$, where $S(\omega,n)$ is the source power constant and $\mathbf{I}$ is an $L\times L$ identity matrix. Thus, the solution to the problem in (7) can be written as

$$\mathbf{H}(\omega,n)=\left[S(\omega,n)\mathbf{R}_x^{-1}(\omega,n)\right]^{1/2}. \quad (8)$$

The covariance matrix, $\mathbf{R}_x(\omega,n)$ can be estimated from the current observation vector as

$$\mathbf{R}_x(\omega,n)=\alpha\mathbf{R}_x(\omega,n-1)+(1-\alpha)\mathbf{X}(\omega,n)\mathbf{X}^H(\omega,n), \quad (9)$$

where $\mathbf{X}(\omega,n)=[X_1(\omega,n),\cdots,X_L(\omega,n)]^T$ is the observation vector at instant n and frequency, ω. Also, the superscript $(\cdot)^T$ represents the transposition operator and α is the forgetting factor, which controls the memory of the update. By setting the source power constant to unity and by using (9) and matrix inversion lemma [10], (8) can be conveniently expressed as (5).

From (5), it is clear that the recursive based PCA is a "sample-by-sample" algorithm, which is suitable for a real-time implementation. It is interesting to note that only the decorrelation matrix at the previous instant, $\mathbf{H}(\omega,n-1)$ and current observation vector, $\mathbf{X}(\omega,n)$ are needed. The output of the PCA is then given as

$$\mathbf{Y}(\omega,n)=\mathbf{H}(\omega,n)\mathbf{X}(\omega,n), \quad (10)$$

where $\mathbf{Y}(\omega,n)=[Y_1(\omega,n),\cdots,Y_L(\omega,n)]^T$. However, since the separation is performed independently for each frequency bin, there will be permutation indeterminacy [3, 4]. Therefore, the permutation needs to be aligned in each frequency bin correctly such that the reconstructed fullband signal contains frequency components only from the same source signal. One solution to solve this problem is shown in [12]. They propose the enforcement of a constraint on the weight matrix, which links the otherwise independent frequencies and hence solves the permutation problem. This is performed by making a DFT-IDFT cascade of the weight matrix [12] and by enforcing a time domain constraint according to

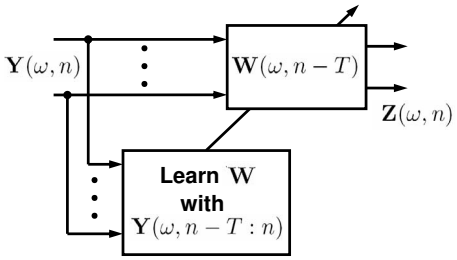

Fig. 2. The block based ICA algorithm, where T is the block delay.

$$\mathbf{H}_t(\tau)=0 \quad (11)$$

where $\tau>Q\ll M$ with $\mathbf{H}_t(\tau)$ denoting the IDFT of $\mathbf{H}(\omega,n)$, Q is the time domain constraint on the filter size and M is the number of frequency bin.

2.3. Independent Component Analysis (ICA)

In keeping with the on-line nature of the PCA algorithm explained in Section 2.2, we propose a block based ICA algorithm. Depending on the block size, T, the block based algorithm introduces a block delay of $T+1$ samples. A simple illustration is shown in Figure X. From the figure, the separation matrix, $\mathbf{W}(\omega,n)$ is actually learnt by using the block of data at instant, $n-T$ up to instant, n, i.e., $T+1$ samples, $[\mathbf{Y}(\omega,n-T):\mathbf{Y}(\omega,n)]$. Due to the block delay, the separation matrix is only updated every $T+1$ samples. This means that the block of data at instant, $n-T$ to instant, n is separated with $\mathbf{W}(\omega,n-T)$ (see Figure 2).

Since the algorithm is applied in frequency-domain, any instantaneous ICA-algorithm can be used [9, 13]. In this paper, the information maximization approach with a natural gradient is applied. The separation matrix, $\mathbf{W}(\omega,n)$ is given as [13]

$$\mathbf{W}^{(k+1)}(\omega,n)=\mathbf{W}^{(k)}(\omega,n)+\mu\left\{\mathbf{I}-\varphi[\mathbf{U}(\omega,n)]\mathbf{U}^H(\omega,n)\right\}\mathbf{W}^{(k)}(\omega,n), \quad (12)$$

where $\mathbf{U}(\omega,n)=\mathbf{W}^{(k)}(\omega,n)\mathbf{Y}(\omega,n)$, μ is the step-size parameter, the index k represents the epoch training and φ is the non-linear function. The nonlinear function is given as [8]

$$\varphi\left[\mathbf{U}(\omega,n)\right]=\tanh\left[\beta\cdot|\mathbf{U}(\omega,n)|\right]\cdot e^{j\{\arg[\mathbf{U}(\omega,n)]\}}, \quad (13)$$

where β is a regulator constant and $|\cdot|$ denotes the absolute value operator. The output can then be written as

$$\mathbf{Z}(\omega,n)=\mathbf{W}(\omega,n-T)\mathbf{Y}(\omega,n). \quad (14)$$

3. SPEECH SIGNAL EXTRACTION

Speech signal extraction is a closely related area to speech enhancement. The two areas differ only in their terminologies, but share the

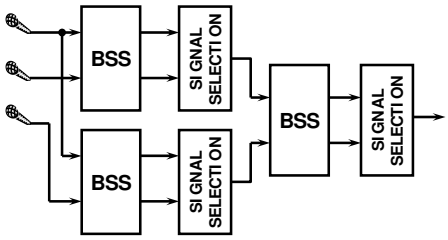

Fig. 3. Configuration of the separation process with pairs of non-uniform array. Note that the three-element array is uniformly spaced.

commonality of extracting or enhancing the signal of interest. Likewise, BSS can be mildly viewed as a signal extraction/enhancement process. The major difference is that conventional BSS yields L outputs as opposed to a single desired signal. For instance, from (14), one observe that there are L outputs from the PCA-ICA algorithm. Consider the situation where there exists only one speech signal in a noisy environment or there is only one speech signal closest to the array, i.e., a strong directional signal exists and coupled with the fact that the separation algorithm converges. Then the speech signal or the directional signal will be in the separated outputs [3, 4, 7]. The task at hand is to identify the speech signal of interest without subjecting to listening test.

Several techniques were proposed in [3, 4, 5] to overcome the output indeterminacy problem. In [4], the signal of interest, i.e., the closest speech source to the array was identified by matching the separation matrix to a direct-path mixing model. Whereas, [5] incorporated some information about the human auditory system to intelligently select the speech signal that is closest to the array. Similar to [3], we use the kurtosis as the signal selection strategy. The selection strategy makes use of the fact that the speech dominant BSS output has the highest kurtosis since its distribution approaches Laplacian. The other $L-1$ outputs will naturally consist of contributions from noise, in which their distributions will tend towards Gaussian, yielding a lower kurtosis. Similar to the previous block based ICA algorithm, the kurtosis method is extended to cope with on-line demand. A detailed analysis is given in Section 5.

4. NON-UNIFORM SPACING

The motivation behind non-uniform spacing microphone array is evident from [7, 8], i.e., the physical understanding of the relationship between wavelength, λ and frequency, ω, $c = \omega\lambda$, where c is the sound wave velocity. Naturally, a better phase resolution is achieved for a wider element spacing for lower frequency range. On the contrary, a closer element spacing is required for higher frequency components. This paper suggests a straightforward method to perform separation by utilizing a different combination of elements from a uniform linear array configuration.

Consider a three-element linear array with inter-element spacing d. Instead of performing standard BSS directly on the observations, the separation can be performed according to Figure 3. From the figure, the separation is performed pair-wise, i.e., the first pair consists of separation involving elements with spacing d and the second pair with spacing $2d$. Since it is assumed that there is only one speech signal in a noisy environment, the kurtosis signal selection strategy is then used to select the speech dominant components from the two separation processes. Following that, a second stage separation is performed on the two speech dominant outputs. The purpose of having the second stage is to realign the desired signal components as each pair of microphones (with different spacing) will have different resolution at each frequency bin, i.e., spacing $2d$ will achieve better separation at lower frequency compared to spacing d [8]. Thus, to separate the desired signal out, another separation is performed. By doing so, the approach bypasses the need of bandpassing the observations to cater for different spacing. Finally, the desired speech signal is then selected through the kurtosis.

5. EVALUATION

The proposed speech enhancement scheme was evaluated in a real room of dimensions $3.5 \times 3.1 \times 2.3$ m^3 using a two-element linear array with a spacing of $d = 0.04$ m. Two loudspeakers emitting babble noise were placed facing the front two corners of the room to create diffuseness and three other babble sources were randomly placed in the middle of the room facing the array. The target signal was positioned approximately 0.5 m from the centre of the array at angle 60°. All simulations were performed with signal to noise ratio SNR $= -0.5$ dB and sampling frequency, $f_s = 8$ kHz. The frequency transformation was performed by short-time Fourier transform with an overlapping factor of four.

A simple experiment was conducted to investigate the frequency binning length effect on the kurtosis. Figure 4 shows the kurtosis of the separated outputs ($L = 2$) with different frequency binning lengths. The solid and the dotted lines show the speech dominant output and the babble dominant output, respectively. Clearly, as the frequency binning length increases, the overlap between the speech and babble outputs decreases. In this experiment, it was found that a 30-point frequency binning length provides the necessary data to compute the kurtosis. For comparison purpose, Figure 5 shows the kurtosis of the actual speech and babble signals with the separated outputs for frequency binning length of 30. The comparable kurtosis values between speech dominant BSS output and the actual speech signal suggests the applicability of the proposed on-line kurtosis.

Table 1 tabulates the suppression and distortion level of the speech dominant output for the uniform 3-element array and the proposed method with different number of frequency bins, M. Results show that the proposed non-uniform approach achieves an average of $3-4$ dB improvement in suppression compared to the case of uniform array. Figure 6 illustrates the relevant spectrograms for the uniform and non-uniform spacing cases. An experiment was also carried out with an adaptive noise canceller (ANC) post-processor [3], labelled as PCA-ICA-ANC in Table 1. Evidently, the ANC benefits from the non-uniform approach with negligible expense on the target distortion, with an average suppression of more than 13 dB.

6. CONCLUSION

This paper presented an on-line PCA-ICA algorithm, which is suitable for real-time implementation. The PCA algorithm is recursively updated whilst the ICA algorithm is block based. The block based kurtosis signal selection strategy is also incorporated to transform the separation process into an extraction process. A non-uniform spacing approach is also suggested for the algorithm and experimental results show higher suppression level is achieved compared to its uniform counterpart.

7. REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech and Signal Process. Magazine*, vol. 5, pp. 4–24, April 1988.

No. of Frequency bins (M)	Uniform				Proposed Non-Uniform			
	PCA-ICA only		PCA-ICA-ANC		PCA-ICA only		PCA-ICA-ANC	
	Supp.	Dist.	Supp.	Dist.	Supp.	Dist.	Supp.	Dist.
128	5.19 dB	-21.69 dB	7.37 dB	-23.65 dB	7.96 dB	-21.42 dB	13.10 dB	-23.64 dB
256	6.23 dB	-21.41 dB	8.61 dB	-21.10 dB	9.77 dB	-22.48 dB	13.56 dB	-22.39 dB
512	7.48 dB	-20.81 dB	10.05 dB	-20.54 dB	10.82 dB	-22.59 dB	14.10 dB	-22.40 dB
1024	5.92 dB	-22.01 dB	7.83 dB	-20.31 dB	9.90 dB	-23.80 dB	13.55 dB	-23.34 dB

Table 1. The suppression (Supp.) and distortion (Dist.) levels of uniform and non-uniform PCA-ICA and PCA-ICA-ANC schemes with different number of frequency bins, M.

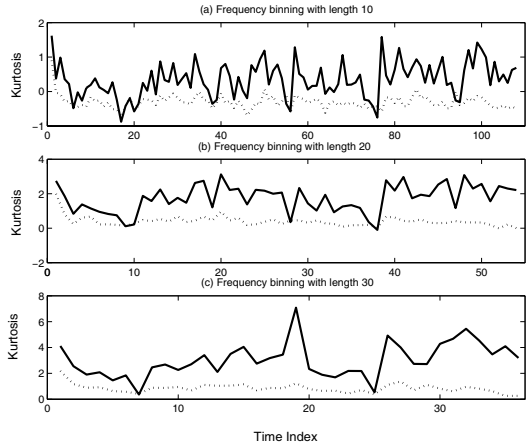

Fig. 4. The kurtosis values of speech dominant (solid line) and babble dominant (dotted line) outputs when updated with a frequency binning length of (a) 10-point, (b) 20-point and (c) 30-point.

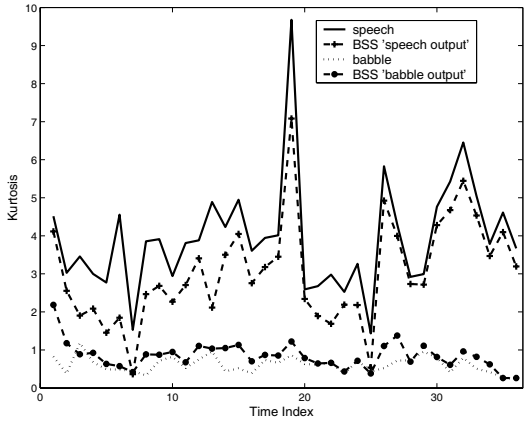

Fig. 5. A comparison of the kurtosis of the speech and babble dominant outputs with clean speech and babble.

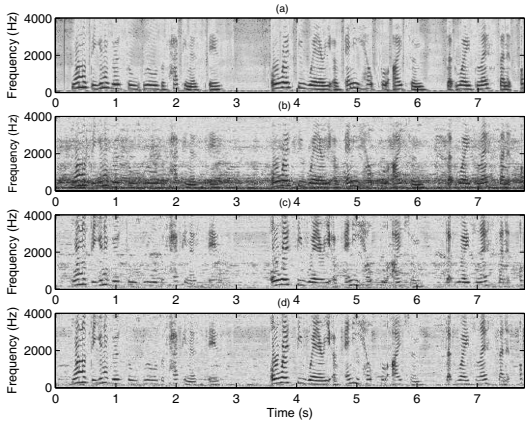

Fig. 6. Spectrograms of (a) the original speech signal, (b) noisy observation, (c) speech dominant output of the uniform PCA-ICA and (d) speech dominant output of the non-uniform PCA-ICA.

[2] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of hands-free speaker input in cars," *IEEE Trans. on Vehicular Technology*, vol. 42, pp. 514–518, November 1993.

[3] 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.

[4] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 3, pp. 61–64, March 2005.

[5] A. K. Barros, F. Itakura, T. Rutkowski, A. Mansour, and N. Ohnishi, "Estimation of speech embedded in a reverberant environment with multiple sources of noise," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 1, pp. 629–632, 2001.

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

[7] 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.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," *IEEE Int. Workshop on Neural Networks for Signal Process.*, pp. 465–474, September 2002.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[10] M. Berg, E. Bondesson, S. Y. Low, S. Nordholm, and I. Claesson, "A combined online PCA-ICA algorithm for blind source separation," *Asia-Pacific Conference on Communications*, pp. 969–972, October 2005.

[11] S. Ding, T. Hikichi, T. Niitsuma, M. Hamatsu, and K. Sugai, "Recursive method for blind source separation and its applications to real-time separations of acoustic signals," *Proceedings of ICA*, pp. 517–522, April 2003.

[12] 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.

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