2010

Speech enhancement via separation of sources from co-located microphone recordings

Muawiyath Shujau
University of Wollongong, mshujau@uow.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

I. Burnett
Faculty of Informatics, University of Wollongong, ianb@uow.edu.au

Speech enhancement via separation of sources from co-located microphone recordings

Abstract

This paper investigates multichannel speech enhancement for colocated microphone recordings based on Independent Component Analysis (ICA). Comparisons are made between co-located microphone arrays that contain microphones with mixed polar responses with traditional uni form linear arrays formed from omnidirectional microphones. It is shown that polar responses of the microphones are a key factor in the performance oflCA applied to co-located microphones. Results from PESQ testing show a significant improvement in speech quality of ICA separated sources as a result of using an A VS rather than other types of microphone arrays.

Disciplines

Physical Sciences and Mathematics

Publication Details

M. Shujau, C. H. Ritz & I. S. Burnett, "Speech enhancement via separation of sources from co-located microphone recordings," in International Conference on Acoustics, Speech, and Signal Processing, 2010, pp. 137-140.

SPEECH ENHANCEMENT VIA SEPARATION OF SOURCES FROM CO-LOCATED MICROPHONE RECORDINGS

M. Shujau, C. H. Ritz and I. S. Burnett
School of Electrical, Computer, and Telecommunications Engineering
University of Wollongong, Wollongong, NSW, Australia
[ms970, critz]@uow.edu.au

School of Electrical and Computer Engineering
RMIT University, Melbourne, VIC, Australia
ian.burnett@rmit.edu.au

ABSTRACT

This paper investigates multichannel speech enhancement for co-located microphone recordings based on Independent Component Analysis (ICA). Comparisons are made between co-located microphone arrays that contain microphones with mixed polar responses with traditional uniform linear arrays formed from omni-directional microphones. It is shown that polar responses of the microphones are a key factor in the performance of ICA applied to co-located microphones. Results from PESQ testing show a significant improvement in speech quality of ICA separated sources as a result of using an AVS rather than other types of microphone arrays.

Index Terms: Microphone arrays, Vector Sensors, Speech Enhancement

1. INTRODUCTION

Enhancing speech recorded from microphone arrays is an important stage in maximizing speech quality in hands free communication applications. One approach to such speech enhancement is to use Blind Source Separation (BSS) techniques [1, 2] and [2] further exploited intensity vector directions derived from a compact microphone array. While direction can be derived from many arrays, an Acoustic Vector Sensor (AVS) can directly record sound direction [3]. An array configured as an AVS has three velocity gradient microphones and one omni-direction microphone arranged orthogonally in an area occupying no more than $1cm^3$ (see Fig. 1). In [3], it was shown that an AVS can be used to accurately estimate the Direction of Arrival (DOA) of a source in air, and this work significantly extends on [3], demonstrating that an AVS is well suited to speech source separation and enhancement in anechoic and reverberant conditions.

In this paper, Independent Component Analysis (ICA) [4] is used as the basis of speech enhancement for AVS recordings. For ICA to work efficiently for spatial recordings, there are two essential criteria: a) sources should be statistically independent and b) the recordings are made with microphones located at different locations. The location of microphones is represented in ICA within the mixing matrix; this matrix incorporates information regarding distance and attenuation due to air absorption, and the effects of reverberations on each source to be separated. These characteristics are widely referred to as the acoustic transfer function for each captured signal [4]. In this paper, and similar to [5], it is proposed that the mixing matrix in the ICA algorithm should be extended beyond the acoustic transfer functions to include the polar patterns and frequency responses of the microphones used to capture signals. This paper investigates the importance of the latter in the ICA mixing matrix for co-located microphones and then considers the consequential impact on enhanced speech quality.

This paper is organized as follows: Section 2 of this paper describes the AVS used in this work and presents the ICA model used for source separation of microphone array recordings incorporating polar responses, frequency responses and the acoustic transfer function. Section 3 presents simulation and experiments investigating the relationship between the polar response, frequency response, microphone array type, statistical properties of the recorded signals, and speech quality performance. Conclusions are presented in section 4.

2. INDEPENDENT COMPONENT ANALYSIS FOR AN AVS

The output of an AVS given in (1) consists of three components: an acoustic pressure component and two acoustic particle velocities. In 2D, this can be expressed in vector form as:

$$\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T \tag{1}$$

where $x_1(t)$ represents the acoustic pressure component measured by the omni-directional microphone and $x_2(t)$ and $x_3(t)$ represent the outputs from two gradient sensors that estimate the acoustic particle velocity in the x and y direction, relative to the microphone position. For the gradient microphones, the relationship between the acoustic pressure and the particle velocity is given by (2)[3] :

$$[x_2(t), x_3(t)] = f\big(p(t) - p(t-\Delta t)\big)\mathbf{u} \tag{2}$$

where f represents a function of the acoustic pressure difference and:

$$\mathbf{u} = [\cos\theta \quad \sin\theta]^T \tag{3}$$

is the source bearing vector with θ representing the azimuth of the source relative to the microphone array [3].

The traditional ICA model applied to a multichannel speech recording assumes that microphone frequency responses for each

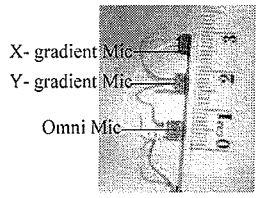

Fig 1: Acoustic Vector Sensor

channel are the same and that the mixing matrix is a result only of the acoustic transfer function. [4]. However, for the AVS, the microphones have directional polar responses. ICA for microphones with directional responses is described in [5]. Following [5] and considering the case of two sources and three microphones (see Fig 3 (a)), the recorded signals can be modeled using the mixing model:

$$\mathbf{x}(n) = \sum_{k=0}^{K-1} \mathbf{A_k}\mathbf{s}(n-k) \qquad (4)$$

In equation (1), $\mathbf{x}(n)$ represents the digitally sampled microphone signals of (1), $\mathbf{s}(n\text{-}k) = [s_1(\text{n-k}), s_2(\text{n-k})]^T$ represents the vector of source signal samples and $\mathbf{A_k}$ represents the convolutive mixing matrices, each of size 3×2. In [5], this model was used to perform ICA on a microphone array containing two closely spaced omni-directional microphones arranged to provide a figure-of-eight polar response and this model is adopted here. In contrast, this work applies ICA to recordings of the acoustic pressure gradient.

In this work, the gradient microphones represented by (2) are second order and result in figure-of-eight polar patterns [3]. In [6], it was shown that ICA can also be applied to gradient signals, represented by time-differentiated sources signals, and the final outputs are determined by integrating the outputs resulting from separation. The formation of the gradient signals can be modelled as a high frequency boost of the source signals of 6dB/octave for frequencies above 2 kHz [7]. This is similar to applying a pre-emphasis filter, which does not result in a significant change in the perceptual quality of a speech signal. Hence, to avoid approximation errors, the gradient microphone signals of (1) are not time-differentiated prior to applying ICA.

3. EXPERIMENTS AND RESULTS

Experiments were performed to compare the performance of ICA for speech enhancement using simulated and real recordings from various types of microphone arrays.

3.1 Experimental setup

Six female and six male sentences from IEEE speech corpus [8], each 10 s long with 1s of silence at the start and at the end, were used as the test database. Noise sources include 10s segments of babble, recordings of a factory floor, recordings of the background noise of a moving vehicle, white noise and pink noise [9]. Two scenarios for sources are used a) one source, one interferer b) one source and diffuse noise (synthesized using four interferers), as shown in Fig. 2 (a) & (b). Noisy speech signals were recorded with a range of signal-to-noise ratios ranging from 0 dB to 20 dB (0dB – the signal and noise levels are equal). Recordings were made at a sampling rate of 48 kHz. In total one hundred recordings were made for each of five SNR levels.

Anechoic recordings were processed using FastICA [4] while reverberant recordings were processed using a convolutive FastICA algorithm [10]. The resulting separated speech signals

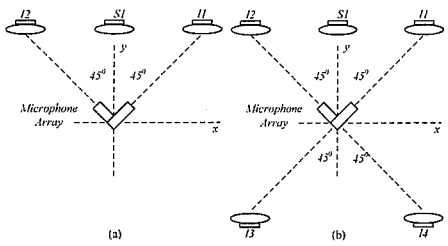

Fig 2: Arrangement of Sources and Microphones for simulation and Experimental recording.

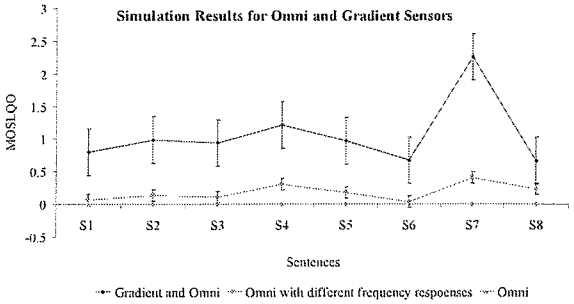

Fig 3: Simulation Results for Omni and Gradient Microphones

were analyzed using the ITU-PESQ software [12] (following low pass filtering and down-sampling to 16 kHz). When using PESQ, each output from ICA is compared with the original clean source signal to give a Mean Opinion Score for Listening Quality (MOSLQO) [12]; the highest MOSLQO corresponds to the target source. A difference MOSLQO is generated by subtracting the MOSLQO of an omni-directional recording of the mixed sources (used as the reference) from the highest MOSLQO of the ICA outputs [14].

3.2 Simulation experiments

This section examines the role played by the microphone characteristics on the quality of the output produced by ICA using simulated recordings. Simulated anechoic recordings were created using room-sim [13] and the test database of Section 3.1, with no attenuation due to air absorption and source-to-microphone distances set to 1 m. In all simulations the signal to noise ratios for the source and interferer were set at 0 dB, corresponding to the worst case scenario.

Three types of collocated microphone arrays were examined. The first array consists of two omni-directional microphones, each with flat frequency response. The second array consists of two omni-directional microphones, 1 with a flat frequency response and 1 with a frequency response having a 6 dB rise/octave above 2 kHz (matching that of a real gradient microphone [7]). The third array consists of one omni-directional and one gradient microphone having the same frequency response as the second microphone of array 2 but with the addition of a figure-of-8 polar response (matching that of a real gradient microphone [7]).

The results obtained from the simulations are shown in Fig 3. As expected there is no improvement in the MOSLQO when using co-located omni-directional microphones with identical frequency responses. For the second array, there is an improvement in the MOSLQO of 0.18. This shows that there is a small contribution to the ICA mixing matrix by the frequency response of the microphone. For the third array which is an AVS simulated with

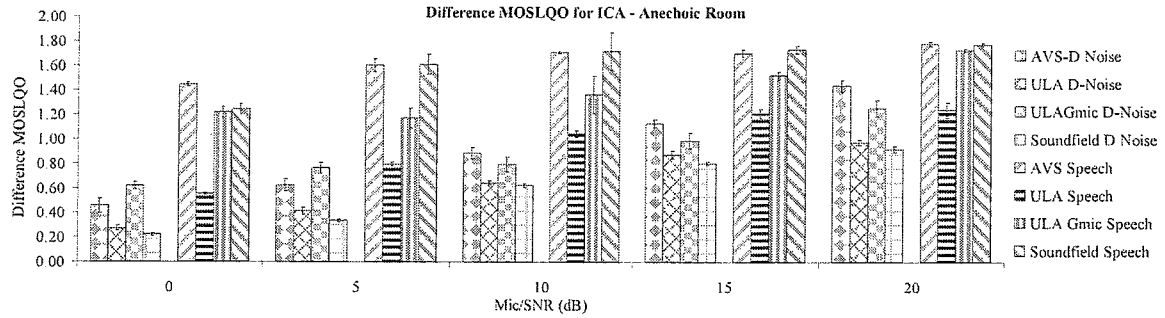

Fig 4: Results of PESQ MOSLQO for Anechoic room

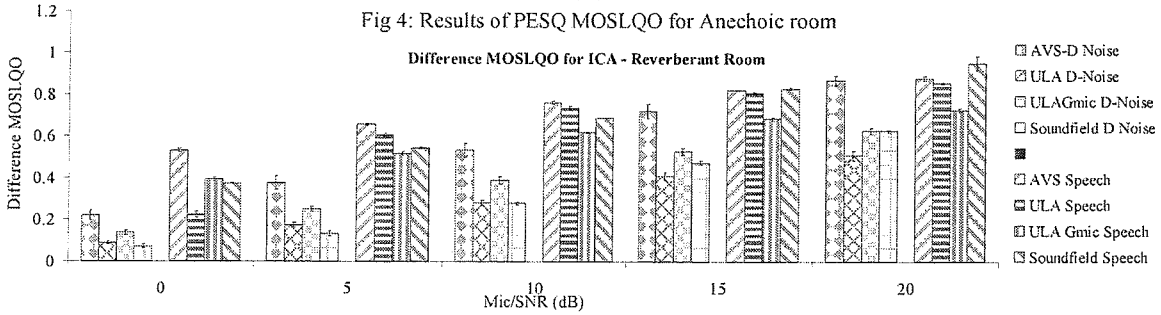

Fig 5: Results of PESQ MOSLQO results for Reverberant room

only omni and X sensor, the results of Fig. 3 show that there is a significant improvement in MOSLQO of 1.24. These results indicate that the main factor in the performance of ICA for speech enhancement from an AVS is the polar responses of the microphones.

3.3 Experiments with Real Recordings

The microphone arrays were experimentally evaluated both in an anechoic chamber and a room with a RT_{60} of 30ms at the University of Wollongong [14]. The experiments used the speech and noise sources of Section 3.1.

The microphone arrays used for the experiment were: a) Acoustic Vector Sensor, b) Uniform linear Array with all omni-directional microphones, c) Uniform Linear Array with two orthogonally located gradient microphones in x and y planes and d) a Soundfield microphone [15] with the polar patterns set to figure of eight. The Uniform Linear arrays have a length of 300mm with 4 capsules (either omni or gradient depending on the array) each spaced 100mm apart. Both the AVS and the Soundfield Microphones are similar in that they record a 3D soundfield using a co-located array of microphones. The key difference between the AVS and soundfield is the type and arrangement of the capsules.

The results of the experiments are shown in Figs. 4 and 5. For anechoic conditions (Fig. 4), with 1 interferer the results from processing the AVS recordings with ICA show an average improvement in MOSLQO of 1.65, which is similar to the results obtained from the Soundfield microphone. However, the AVS with 1 speech interferer at an SNR of 0 dB, results in an MOSLQO of approximately 0.2 better than the Soundfield. For diffuse noise, the AVS produces an average improvement in MOSLQO over all noise scenarios of 0.9, which is similar to the next best performing array (in this case, the ULA with gradient microphones). However, the AVS is significantly better at high SNRs, while decreasing in performance at low SNRs.

The results for the reverberant room (Fig. 5) are different to those of the anechoic case. For the speech interferer, MOSLQO results for the AVS are on average 0.1 better over all SNR scenarios than the next best performing array, in this case the ULA. For diffuse noise, the AVS again performs better than all other arrays, with an average MOSLQO improvement of 0.14 higher than the next best performing array (again being the ULA). However, for both single interferers and diffuse noise at 0 dB, the AVS performs significantly better (on average 0.4) compared with the ULA, which is the next best performing array.

3.4 Experiments with changing microphone array orientation

Having established that the AVS is the best performing microphone array, compared with the ULA being best for most other scenarios, experiments were conducted to evaluate the impact of array orientation (relative to the source) on the resulting ICA performance. Since the gradient microphones are highly directional the performance of ICA may be due to directing the microphones directly at the source or interferer. The signal to noise ratio is set at 0 (the worst case scenario) and the recordings are made for a single speech interferer. The arrays are rotated in azimuth through 90^0 at 15^0 intervals and recordings made for each orientation. The results in Fig. 6 show MOSLQO results for outputs from ICA performed on these recordings in both anechoic and reverberant environments. The results show that there is very little or no effect on the performance of ICA by turning the array in azimuth.

3.5 Discussion

The work presented in this paper has shown that there is a significant impact on the performance of the ICA when the co-located microphones with different polar responses are used. This agrees with previous work investigating ICA for closely spaced microphone arrays with directional responses [7]. It is suggested that using directional microphone recordings results in increased

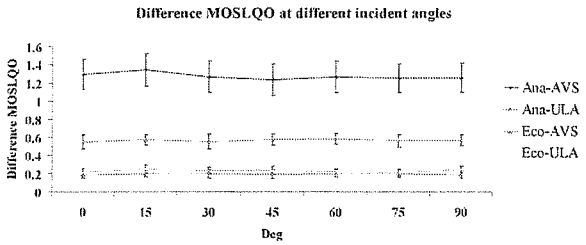

Fig 6: The MOSLQO results for different azimuth angles

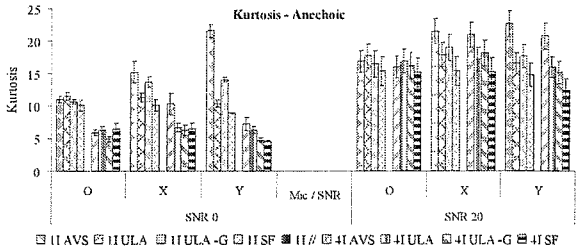

Fig 7: Kurtosis results for SNR of 0dB and 20dB for anechoic recordings

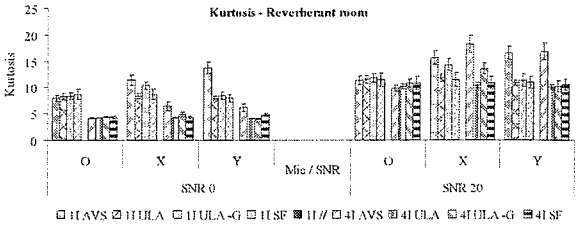

Fig 8: Kurtosis results for SNR of 0dB and 20dB for reverberant recordings

statistical independence between the recorded signals, in turn this results in improved separation performance using ICA. Using the database of recordings described in Section 3.1, the kurtosis of each of the microphone recordings was measured, with the results shown in Figs. 7 and 8. The results show that when the microphones have directional polar responses, the recorded signals from different channels will have significantly different kurtosis values. When results from Figs. 7 and 8 are compared with the performance results of ICA in Figs. 4 and 5, it is seen that where there is a large variation in the kurtosis values between the channels the performance of ICA as measured by PESQ is improved. This indicates that directional microphones result in signals that are more suitable for separation via ICA, compared with arrays of non-directional microphones.

4. CONCLUSION

This paper has investigated speech enhancement using source separation techniques applied to an Acoustic Vector Sensor (AVS). Source separation is based on a convolutive ICA model applied to co-located microphones that both record omni-directional with directional gradient signals. Perceptual quality results (measured using PESQ) show a significant improvement in speech enhancement using ICA applied to an AVS compared to ICA applied to a traditional linear microphone array, in both anechoic and reverberant environments. Results also show that a key factor in the performance improvement is the use of directional polar responses, which lead to recorded signals that are statistically independent. Future work will investigate alternative techniques to speech enhancement using an AVS as well as alternative AVS configurations and the use of mutual information to analyse recordings from different arrays.

Acknowledgement: This project was partially supported by the Australian Research Council Grant DP0772004. We acknowledge the anonymous reviewers comments including the suggestion to use mutual information to analyse recordings from different arrays.

5. REFERENCES

[1] S. Nordholm, and S. Y. Low, "Speech Extraction Utilizing PCA-ICA Algorithm with a non-uniform spacing microphone array," *IEEE ICASSP 2006*, France, May 2006

[2] B. Gunel, H. Hacihabiboglu, A.M. Kondoz, "Intensity vector direction exploitation for exhaustive blind source separation of Convolutive mixtures," *Proc. ICASSP 2009*, Apr. 2009.

[3] M. Shujau, C.H. Ritz, I.S. Burnett, "Designing Acoustic Vector Sensors for localization of sound sources in air", *EUSIPCO 2009*, UK, August 2009

[4] A. Hyvärinen, E. Oja, "Independent Component Analysis: Algorithms and Applications", *Neural Networks*, Elsevier Science Ltd, Vol. 13 (4-5), pp411- 430, June 2000.

[5] M. S. Pedersen, D. Wang, J. Larsen, U. Kjems, Two-microphone Separation of Speech Mixtures, *IEEE Trans. on Neural Networks*, vol. 19(3), pp. 475-492, IEEE Press, 2008.

[6] M. S. Pedersen, C. M. Nielsen, "Gradient flow convolutive blind source separation," *IEEE Machine Learning for Signal Proc., 2004*,October 2004

[7] J. Eargle, "The Microphone Book," Elsevier, UK, 2004

[8] IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio and Electro-acoustics*, AU-17(3), 225-246

[9] The Signal Processing Information Base (SPIB), Available online: http://spib.rice.edu/spib.html

[10] S. C. Douglas, H. Sawada, S. Makino, "A spatio-temporal fastICA algorithm for separating convolutive mixtures," *IEEE ICASSP05*, Vol.5, pp 165-168, March 2005

[11] ITU P.862 (2000). Perceptual evaluation of speech quality (PESQ) and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P. 862

[12] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions ", *J. Acoust. Soc. Am.*, pp-3387-3405, May 2009.

[13] D. R. Campbell, K. J. Palomäki and G. J. Brown, "Roomsim, a MATLAB Simulation of "Shoebox" Room Acoustics for use in Teaching and Research," *Computing and Information Systems Journal*, Vol. 9, 2005.

[14] C. H. Ritz, G. Schiemer, I. S. Burnett., E. Cheng, D. Lock, T. Narushima, S. Ingham, D. W. Conroy, "An Anechoic Configurable Hemispheric Environment for Spatialised Sound", *Proc. of the 2008 Australia Computer Music Conf.*, July 2008

[15] Soundfield Research, Soundfield: An Introduction, Available online: www.soundfield.com/feature.htm.