DEALING WITH UNCERTAINTY IN MICROPHONE PLACEMENT IN A MICROPHONE ARRAY SPEECH RECOGNITION SYSTEM

Ivan Himawan, Sridha Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology
Brisbane, QLD, Australia
i.himawan@qut.edu.au, s.sridharan@qut.edu.au

Iain McCowan

CSIRO e-Health Research Centre
Brisbane, QLD, Australia
iain.mccowan@csiro.au

ABSTRACT

This paper investigates robustness to uncertain microphone placements in an array beamformer front-end to a speech recognition system. There are two general approaches to handling the placement uncertainty: using the approximately known geometry in a robust beamforming technique, or using techniques that require no prior knowledge of geometry. Experiments in this paper compare the robustness of different techniques for both of these approaches in terms of speech recognition accuracy. To benefit from existing microphone array speech recognition data corpora for experimentation, microphone placement uncertainty is simulated by introducing random perturbations in the assumed geometry. Experimental results show that robust beamforming yields stable performance to a certain degree of placement error, but thereafter techniques such as automatic calibration are beneficial.

Index Terms— Microphone Array, Speech Recognition

1. INTRODUCTION

Microphone array technologies have been used in a wide range of applications such as teleconferencing, hearing aids, speaker localization and tracking, and as the front-end of speech recognition systems. Most microphone array processing algorithms employ spatial filtering, or beamforming, to capture the desired signal while minimising noise and interference.

With advances in sensor and sensor network technology, multimedia-capable devices and ad-hoc computing networks are becoming ubiquitous. In this new context, there is potential for applications that employ ad-hoc networks of microphone-equipped devices collaboratively as a virtual microphone array [1]. One example of data collected using arrays with unknown geometry is that used in the NIST rich transcription meeting recognition evaluation [2]. The data contains recordings of natural meeting interactions captured from a number of sites, each with their own microphone configurations. These configurations are unknown to the evaluation participants who must use the recorded data to perform speech recognition of the meetings. Unfortunately most traditional beamforming algorithms are based on assumptions that the array is stationary and has a known geometry. In the sensor network paradigm, these assumptions are inappropriate. Sensor locations are likely to be dynamic and unknown a priori.

Motivated by a desire to progress from traditional microphone arrays towards less constrained microphone networks, this paper investigates approaches that improve robustness of a microphone array beamformer to erroneous microphone placement. Such errors or uncertaintity occur when the position of microphones are estimated using array shape calibration approaches.

A first approach to this problem aims to design beamformers that are inherently insensitive to sub-optimal conditions and errors, rather than correcting the errors themselves. This is usually achieved using adaptive methods to estimate and minimise the noise and steering vector uncertainties from the input signals [3, 4]. Theoretical statistical analysis on the signal-to-interference-plus-noise (SINR) ratio performance of robust beamformers in the presence of random steering vector errors has been thoroughly presented in the literature [5, 6]. While such analysis provides valuable insight into the effectiveness and optimality of the beamformer, there is a need to confirm the practical relevance of the analysis, for instance when the beamformer is used as a front-end in speech recognition system. In such cases, eventual speech recognition accuracy is a more pertinent measure than the theoretical SINR.

A second approach is to consider that the geometry of the array is completely unknown. In this case, beamforming can be done by either direct estimation of the steering vector from the signals, or by first employing blind calibration of microphone location. One implementation of a direct estimation approach is blind beamformer employed in the Multiple Distant Microphone (MDM) speech recognition front-end for automatic meeting transcription reported by the AMI (Augmented Multi-party Interaction) project [7]. A solution for calibrating an unknown array geometry using only background noise statistics was proposed in [8], and was shown to give a reasonable estimate of the array geometry in a diffuse noise environment. Using knowledge of the estimated array geometry in conjunction with source localisation, the speakers position can be derived. Similarly to the blind beamforming approach, this automatic calibration technique determines both the microphone and speaker locations from the observed signals only, requiring no a-priori information about the array geometry or speaker location.

This paper investigates methods for dealing with microphone placement uncertainty in a microphone array front-end to a speech recognition system. Speech recognition accuracy is measured while varying the magnitude of random microphone placement errors. The aims of the experiments are to quantify the effect placement uncertainty has on speech recognition performance in general, and to identify promising approaches for improving system robustness to such uncertainty. The paper is organised as follows. Section II overviews the various techniques used in the experiments. Speech recognition experiments are presented and discussed in Section III and IV, followed by concluding remarks in Section V.

2. BACKGROUND PROBLEM

2.1. Beamforming with Uncertain Array Geometry

The first method of dealing with uncertain array geometry is to use a beamformer which is inherently robust to microphone placement, or steering, errors.

Beamforming is an effective method of spatial filtering. It differentiates desired signal from noise and interference based on its location. Consider a desired signal received by N omnidirectional microphones sampled at discrete time k, in which each microphone output is an attenuated and delayed version of the desired signal $a_n s(k-\tau_n)$ and noise v_n given by $x_n = a_n s(k-\tau_n) + v_n(k)$. In the frequency domain, the array signal model is $X(\omega) = S(\omega)\mathbf{d} + V(\omega)$, where $\mathbf{d}$ represents the array steering vector which depends on the actual microphone and source location. In the near field, $\mathbf{d}$ is given by [9]:

$$\mathbf{d} = [a_0 e^{-j\omega\tau_0}, a_1 e^{-j\omega\tau_1}, ..., a_{N-1} e^{-j\omega\tau_n}]^T,$$

$$a_n = \frac{d_{ref}}{d_n},\ \tau_n = \frac{d_n - d_{ref}}{c},$$

where d_n and d_{ref} denote the Euclidian distance between the source and the microphone n, or the reference microphone, respectively. To recover the desired signal, each microphone output is weighted by frequency domain coefficients $w_n(\omega)$ and the beamformer output is the sum of N weighted microphone outputs given by $Y(\omega) = \sum_{n=1}^{N} w_n^H(\omega) X_n(\omega)$. The inverse Fourier transform results in time domain output signal $y(k)$.

2.1.1. Delay-Sum Beamforming

Delay-sum beamforming compensates for the signal delay to each microphone output appropriately. After summing outputs, the desired signal will be reinforce, while the noise signals will be effectively reduced through destructive interference.

2.1.2. Optimum Beamforming

Optimum beamforming obtains filter weights according to a given optimisation criterion. Using the minimum variance distortionless criterion, the mean square of output noise power is minimised:

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_n \mathbf{w} \tag{1}$$

where $\mathbf{R_n}$ is the spatio-spectral covariance matrix of noise, subject to distortionless constraint given by:

$$\mathbf{w}^H \mathbf{d} = 1 \tag{2}$$

The well know solution is usually termed the Minimum Variance Distortionless Response (MVDR) weights, given by:

$$\mathbf{w_{MVDR}} = \frac{\mathbf{R}_n^{-1}\mathbf{d}}{\mathbf{d}^H \mathbf{R}_n^{-1}\mathbf{d}} \tag{3}$$

To increase the robustness of MVDR in the presence of array perturbation, quadratic constraint is applied to the beamforming weights by adding weighted identity matrix to the spatio-spectral covariance matrix [10].

Superdirective beamforming is derived from MVDR by applying theoretically well defined noise fields, such as a diffuse noise field. A diffuse noise field appears in several practical reverberant environments, such as inside offices or cars. The coherence function of a diffuse noise field can be modelled as:

$$\Gamma_{ij}(f) = sinc\left(\frac{2\pi f d_{ij}}{c}\right) \tag{4}$$

where d_{ij} is the distance between microphone i and j. This can be used in place of the noise covariance matrix in the MVDR solution to obtain the superdirective filter weights.

2.1.3. Robust Generalised Sidelobe Canceller

Optimum beamforming can be efficiently implemented using the Generalised Sidelobe Canceller (GSC) structure. However in standard GSC, the presence of array perturbation causes error in the steering vector, leading to target signal cancellation. To improve the robustness of GSC, [4] proposed Robust GSC (RGSC) with a coefficient constrained adaptive blocking matrix filter and a norm constrained adaptive noise canceller, to pick up the target signal with minimum distortion in the presence of array perturbation. The Robust GSC can be efficiently implemented using a block frequency domain method [11].

In a practical implementation of the RGSC, adaptation control is needed to prevent target signal cancellation. To achieve this for experiments in this paper, the adaptive noise canceller was adapted offline using a segment of noise samples taken from the speech recognition database, while the blocking matrix was adapted during noisy speech signal period.

Standard RGSC uses delay-sum as the fixed beamformer component. The experiments in this paper compare this (denoted RGSC_DS) to the use of a superdirective beamformer as the fixed beamformer (RGSC_SD), in order to improve array gain in diffuse noise fields and increase the spatial selectivity in the low frequency range.

2.2. Beamforming with Unknown Array Geometry

An alternative to robust beamforming from an approximately known array geometry, is to beamform with no prior knowledge of the array geometry. In this case, beamforming may be performed in one of two ways: by direct estimation of the array steering vector $\mathbf{d}$, or by explicitly estimating the array geometry through automatic calibration technique.

2.2.1. Beamforming from Estimated Propagation Vector

One beamformer that directly estimates the propagation vector is the blind beamformer employed by the AMI-MDM system [7]. The system first measures the energy of each microphone signal output, and selects the highest energy microphone as reference for Time-Delay-on-Arrival (TDOA) estimation. The attenuation factor which corresponds to the ratio of energies between each channel and the reference channel is also calculated. The TDOA delay between each microphone and a reference microphone is determined by finding the peak of generalised cross correlation function [12]. From these, the array steering vector $\mathbf{d}$ is constructed, and the beamforming weights calculated.

2.2.2. Beamforming from Blindly Derived Positions

First, microphone positions are revealed using technique proposed in [8]. It assumes no prior knowledge of microphone placement, using only background noise statistics to achieve shape calibration.

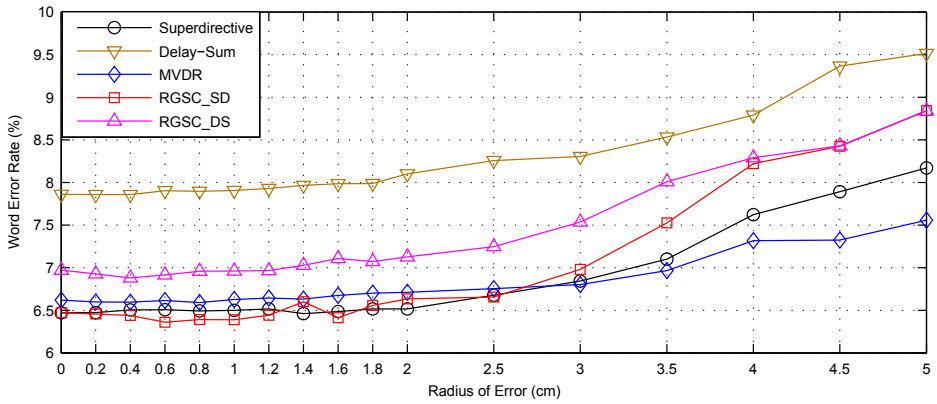

Fig. 1. *Word Error Rate versus Microphone Placement Error for various robust beamforming approaches.*

The technique works by fitting the theoretical diffuse noise coherence function (Eq. 4) to the measured noise coherence to obtain the distance between microphone pairs.

Given these distances, classical multidimensional scaling is performed to estimate the array geometry. Having derived microphone positions automatically, location of the source can then be estimated using source localisation algorithms in relative to revealed positions. Steered Power Response Phase Transform weighted (SRP-PHAT) is chosen in this work due to its robustness in reverberant environment [13].

3. EXPERIMENTS

Speech recognition experiments were conducted on the single speaker (S1) portion of the Multichannel Overlapping Numbers Corpus (MONC) [14][1]. The MONC contains digit utterances spoken around a circular meeting room table and captured by a fixed 8-element, equally spaced, table-top circular microphone array. The noise conditions in the recording was predominantly diffuse background noise.

The clean MONC S1 training data set was used to train baseline Hidden Markov Models (HMM) with standard Mel-Frequency Cepstral Coefficient (MFCC) parameters (including 0th cepstral coefficient) and their first and second derivatives. The baseline system achieves a word error rate (WER) of 4.37% on the test set. In the following experiments, MLLR followed by MAP adaptation of the models is performed for each technique using the MONC development set. The WER for each robust beamformer using ground truth microphone placements is given in Table 1.

In this article, in order to investigate the effect of uncertain microphone placement on beamforming accuracy only, the following experiments all assume a known speaker location. This allows us to ignore effects of microphone placement uncertainty on automatic speaker localisation in the current study. Clearly this is unlikely in a practical system with uncertain microphone locations, however, and the use of speaker localisation forms the focus of our ongoing research.

To simulate the placement uncertainty for the robust beamforming experiments, the true microphone locations were perturbed with random angle for a specified radius r. The radius of error was increased step by step, initially with 0.2 cm increment from the actual microphone positions up to 2 cm, followed by 0.5 cm increments from 2 cm to 5 cm. Due to randomisation, results are averaged over 15 experiments for each increment of error. The mean WER results are plotted against placement error in Figure 1.

[1]Available from http://www.cslu.ogi.edu/corpora/

Table 1. *Speech recognition performance of robust beamforming techniques using ground-truth microphone positions.*

Techniques	WER (%)
Superdirective	6.47%
Delay-Sum	7.86%
MVDR	6.62%
RGSC_SD	6.47%
RGSC_DS	6.97%

Results comparing the different approaches from unknown geometry are presented in Table 2 for delay-sum, MVDR and superdirective beamformers. As the superdirective technique requires inter-microphone distances to calculate the noise coherence matrix, this could not be implemented for the AMI-MDM approach, however in automatic calibration approach, the technique uses inter-microphone distances from estimated array geometry. Calibration results are averaged over 15 runs using different noise segments in the calibration procedure. The table also gives the intersection point with Figure 1, indicating the degree of microphone placement uncertainty at which the AMI-MDM or automatically calibrated array performance becomes better than the robust beamformer alone.

4. DISCUSSION

At true microphone positions as presented in Table 1, superdirective and RGSC_SD achieve the lowest WER among all techniques, as the MONC database was recorded in approximately diffuse noise conditions (i.e. a moderately reverberant room with no significant localised noise sources). MVDR gives similar performance, albeit slightly degraded due to filter initialisation on real noise samples. RGSC_DS shows significant reduction of WER compare to the standard delay and sum beamforming techniques due to the adaptive noise cancelling structure.

In the presence of small array perturbations, for radius error from 0.2 cm to 2 cm, all beamforming techniques exhibit relatively stable performance. This indicates a certain degree of robustness in the presence of array mismatch. The performance is stable for such

Table 2. *Speech recognition performance of beamformers from unknown array geometry. Mean and standard deviation over 15 runs are given for automatic calibration technique. Intersection indicates placement error for equivalent performance in Figure 1.*

Techniques	Word Error Rate WER(%)		Intersection (cm)
AMI-MDM Delay-Sum	8.88%		4.09
AMI-MDM MVDR	6.88%		3.24
	Mean WER(%)	Std. WER(%)	
AutoCal.+Delay-Sum	8.04%	0.09%	1.9
AutoCal.+Superdirective	6.62%	0.09%	2.33
AutoCal.+MVDR	6.63%	0.05%	1.4

small errors even for delay-sum beamforming, indicating that the robustness at this level is an inherent characteristic of the array and speaker configuration, rather than being due to any explicitly robust beamforming techniques.

As the radius of error increases, the speech recognition accuracy for all beamforming techniques begins to decline. In relative terms, most techniques degrade at a similar rate, apart from the MVDR and RGSC_SD approaches. The MVDR approach shows the greatest robustness to erroneous microphone placement. While RGSC_SD performs significantly better than RGSC_DS for small placement errors, it degrades more rapidly as the error increases. This can likely be attributed to the additional errors in the inter-microphone distances used in the noise coherence matrix, compounding the effect of the steering errors.

The results using a completely unknown array geometry in Table 2, show that the AMI-MDM achieve higher WER compared to automatic calibration. This is might be caused by inacurracy in TDOA delay calculations. Automatic calibration techniques in the other hand achieve WER comparable to a system with approximately 2cm of uncertainty in the microphone placements. In both cases, the MVDR beamformer yields significantly better performance than the delay-sum beamformer, due to its greater robustness to error seen in Figure 1.

The good performance of automatic calibration techniques might be due to the closer integration between microphone and source position estimations, in which speaker position is estimated from calibrated microphone positions. The estimated position of the speaker may compensate errors in microphone positions which require further investigation in the approach.

5. CONCLUSION

This paper has examined the effect of uncertain geometry on the accuracy of a microphone array speech recognition system. Robust beamforming approaches are compared, as well as two techniques that require no prior geometry information. For the investigated configuration, results show the beamformers are robust up to 2cm of microphone placement error. When uncertainty exceeds this, it is better to estimate the required delay or geometry information directly from observed signals.

6. ACKNOWLEDGMENT

This work is partly supported by the European IST Programme Project FP6-033812 (AMIDA).

7. REFERENCES

[1] V.C. Raykar, I.V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 70–83, 2005.

[2] J.G. Fiscus et al., "The rich transcription 2006 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction*, S. Renals, S. Bangio, and J.G. Fiscus, Eds., vol. 4299, pp. 309–322. Springer, 2006.

[3] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 1365–1376, 1987.

[4] 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 Processing*, vol. 47, pp. 2677–2684, 1999.

[5] M. Wax and Y. Anu, "Performance analysis of the minimum variance beamformer in the presence of steering vector errors," *IEEE Trans. on Signal Processing*, vol. 44, pp. 938–947, 1996.

[6] O. Besson and F. Vincent, "Performance analysis of beamformers using generalized loading of the covariance matrix in the presence of random steering vector errors," *IEEE Trans. on Signal Processing*, vol. 53, pp. 452–459, 2005.

[7] T. Hain et al., "The 2005 AMI system for the transcription of speech in meetings," in *Proc. NIST MLMI Meeting Recognition Workshop*, 2005.

[8] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array calibration in diffuse noise fields," *To Appear in IEEE Trans. on Audio, Speech and Language Processing*, 2007.

[9] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds., chapter 2, pp. 19–38. Springer, 2001.

[10] J. E. Hudson, *Adaptive Array Principles*, Peter Peregrinus Ltd., New York, 1981.

[11] W. Herbordt and W. Kellermann, "Efficient frequency-domain realization of robust generalized, sidelobe cancellers," in *Proc. IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 377–382.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, pp. 320–327, 1976.

[13] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds., chapter 8, pp. 157–180. Springer, 2001.

[14] D. Moore and I. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003, vol. 5, pp. 497–500.