ON THE EFFECT OF SNR AND SUPERDIRECTIVE BEAMFORMING IN SPEAKER DIARISATION IN MEETINGS

Erich Zwyssig[1,2]*, Steve Renals*[1] *and Mike Lincoln*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, Scotland UK
[2]EADS IW, Appleton Tower, 6th Floor, Edinburgh, EH8 9LE, Scotland UK

ABSTRACT

This paper examines the effect of sensor performance on speaker diarisation in meetings and investigates the use of more advanced beamforming techniques, beyond the typically employed delay-sum beamformer, for mitigating the effects of poorer sensor performance. We present superdirective beamforming and investigate how different time difference of arrival (TDOA) smoothing and beamforming techniques influence the performance of state-of-the-art diarisation systems. We produced and transcribed a new corpus of meetings recorded in the instrumented meeting room using a high SNR analogue and a newly developed low SNR digital MEMS microphone array (DMMA.2). This research demonstrates that TDOA smoothing has a significant effect on the diarisation error rate and that simple noise reduction and beamforming schemes suffice to overcome audio signal degradation due to the lower SNR of modern MEMS microphones.

Index Terms— Speaker diarisation in meetings, digital MEMS microphone array, time difference of arrival (TDOA), superdirective beamforming

1. INTRODUCTION

Speaker diarisation is the process of determining *who spoke when* during a conversation. Diarisation systems typically identify both the number of speakers in the recording and the time intervals during which each individual is speaking. Speaker diarisation has recently been used in the analysis of meeting recordings which has shown that accurate diarisation is crucial to the performance of subsequent processes, such as speaker recognition and transcription [1].

Meetings are usually recorded using microphone arrays consisting of a number of high quality analogue microphones which provide a high signal to noise ratio (SNR). However, microphone arrays have recently been developed using digital MEMS microphones (so-called silicon microphones). MEMS microphone arrays have a number of advantages (e.g. lower price, smaller size), however the MEMS sensors have the disadvantage of significantly lower SNRs than their analogue counterparts.

In this paper we study the effect of the sensor performance on the diarisation task and investigate the use of superdirective beamforming to mitigate the effects of poorer sensor performance.

2. BACKGROUND

Time delay of arrival (TDOA) estimation seeks to identify the time difference between signals from a given source arriving at two different sensors in a sensor array and is an essential first step in most beamforming techniques. An established method for performing TDOA estimation is the generalised cross correlation with phase transform (GCC-PHAT [2],[3]) which can be used to determine the relative delay between signals arriving at two microphones in a microphone array.

The GCC-PHAT of two signals is defined as

$$\hat{G}_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{1}$$

where $x_i(t)$ and $x_j(t)$ are two discrete signals in the time domain and $X_i(f)$ and $X_j(f)$ their discrete Fourier transform. The TDOA $\hat{d}_{PHAT}(i,j)$ of the two signals $x_i(t)$ and $x_j(t)$ is estimated as the maximum value of the inverse Fourier transform $\hat{R}_{PHAT}$ of $\hat{G}_{PHAT}$:

$$\hat{d}_{PHAT}(i,j) = \arg\max_d(\hat{R}_{PHAT}(d)) \tag{2}$$

GCC-PHAT does not produce stable delay estimates when used in acoustically noisy environments (such as a typical meeting room), and smoothing techniques, such as Viterbi delay selection, can be used to obtained better estimates [4]. In our experiments we compare the performance of smoothed and un-smoothed delay estimates for beamforming in terms of the achieved diarisation error rate.

For acoustic beamforming, the delays between each microphone and a reference channel (typically taken as the channel with the highest energy level) are calculated, and these can be directly used for delay-sum beamforming. The output of a delay-sum beamformer is the weighted sum of all the microphone signals, with each channel delayed by it's

corresponding delay estimate:

$$y(n) = \frac{1}{M} \sum_{m=1}^{N} x_m(n - \tau_m) \quad (3)$$

where $y(n)$ is the output signal of the beamformer, M the number of microphones, x_m the input signal at microphone m and τ_m the delay of that input signal.

One commonly used measure of the performance of beamforming techniques is the array gain, G, which shows the improvement of the signal to noise ratio of the array compared to an individual sensor:

$$G = \frac{SNR_{array}}{SNR_{sensor}} \quad (4)$$

Delay-sum beamforming achieves a signal amplification of 3dB for every doubling of the number of microphones. Enhancement is achieved by constructively adding the signals from the look direction and suppressing interference from other sources. By optimising the array gain, more sophisticated methods, known as superdirective beamformers, can be used to improve the beamformers directional selectivity at lower frequencies, further cancelling undesired sources. A number of superdirective beamformers have been developed. Examples include filter-sum, differential, eigen, generalised sidelobe cancelling, and minimum variance distortionless response (MVDR) beamformers, each being differentiated by the method employed to optimise G.

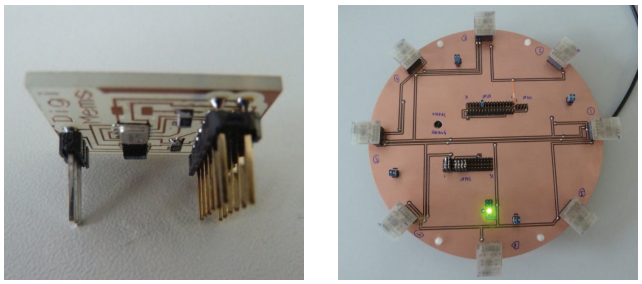
In this work we employ an MVDR superdirective beamformer [5]. The aim of MVDR beamforming is to minimise the power of the output signal of the array, while maintaining unity gain in the look direction and additional constraints (such as maximum white noise gain). MVDR beamforming is based on filter-sum beamforming and its frequency domain output signal, Y_b, is defined as:

$$Y_b(e^{j\Omega}) = \sum_{m=0}^{M-1} {W_m}^*(e^{j\Omega}) X_m(e^{j\Omega}) = \mathbf{W}^H \mathbf{X} \quad (5)$$

where $W_m(e^{j\Omega})$ denotes the filter coefficients of the beamformer for sensor m at frequency Ω, $\mathbf{X_m}(\mathbf{e^{j\Omega}})$ are the microphone input signals and $[\cdot]^H$ denotes the matrix transpose conjugate.

3. DMMA.2

Most microphone arrays to date have been composed of high quality, but expensive and relatively bulky, analogue microphones. A digital MEMS (micro electro mechanical system) microphone is a microphone on a chip containing a pressure sensitive membrane, a matched pre-amplifier, and integrated analogue-digital conversion (ADC) and downsampling. We have previously constructed a prototype digital MEMS microphone array, DMMA.1 [6], and preliminary experiments produced promising results for a task based on the adaptation of WSJ acoustic models. DMMA.1 has a number of limitations, most significantly the inability to directly record all channels individually at 48kHz sample rate. In order to overcome this problem, a second microphone array has been constructed which allows the recording of 8 microphone channels at sample rates from 8 kHz to 48 kHz.



(a) Microphones on daughter boards (b) Complete microphone array

Fig. 1. The digital MEMS microphone array

In this work we have designed a new array, DMMA.2 (Figure 1), which like DMMA.1 is an 8 channel circular microphone array with a diameter of 20 cm. It is built using ADI ADMP441 omnidirectional MEMS microphones[1] with bottom port and I^2S output and the Rigisystems USBPAL[2], a USB 2.0 multi-channel audio interface for Windows PC and MAC OS X.

Digital MEMS microphones have significantly lower intrinsic signal to noise ratios compared to analogue microphones. Initial tests on the microphones used in the DMMA.2 suggest that this sensor noise is not white as would be expected. While SNR and THD measurements carried out show the microphones to be within specification, the MEMS microphones output a non-white chirping noise, which we suspect originates from the DSP built into the microphones. Further tests, including sensor measurements in a vacuum enclosure are being conducted.

4. AD_IMR CORPUS

The DMMA.2 and an array with identical geometry constructed using high signal to noise ratio analogue microphones have been used to simultaneously record six research meetings of around one hour in length. The recordings were made in a typical meeting room at the University of Edinburgh. The analogue array is identical to that used in the AMI meeting corpus recordings and is fully documented in [1]. From each of the recordings, a continuous ten to fifteen minute segment containing lively discussion has been selected, creating a total of approximately 78 minutes of recordings. These extracts were transcribed to show speech/non-

[1] http://www.analog.com/en/mems-sensors/microphones/admp441/products/product.html

[2] http://www.rigisystems.net/

Recording	Length [s]	# of speakers
rec14june2011	825	5
rec15june2011	804	7
rec21june2011	630	4
rec22june2011	856	4
rec28june2011	607	4
rec29june2011	914	6

Table 1. Summary of AD_IMR recordings

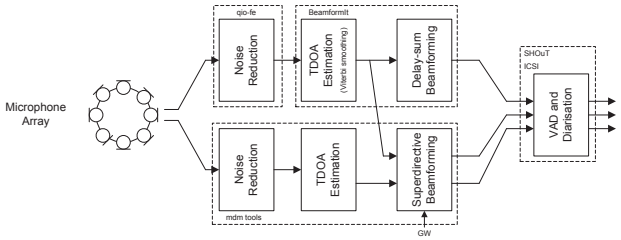


Fig. 2. Data flow for the experiments

speech events and for each speech segment the speaker ID was annotated. Both overlapping speech (where more than one speaker is talking simultaneously) and back channels (short interjections from listeners, typically indicating agreement or disagreement with the main speaker) were included in the transcription. The transcription was formatted using the RTTM specification, as defined by NIST [3], allowing scoring of automatically generated diarisation annotations using the standard NIST evaluation tools. Details of the meeting recordings contained in the corpus, named AD_IMR, are listed in Table 1.

5. METHODS

Experiments were conducted to investigate the effect on the diarisation task of using the digital array and superdirective beamforming. Using two state-of-the-art diarisation systems, we compared the error rates achieved using the low SNR recordings from the DMMA.2 with recordings of the same meeting from the analogue array. Using both smoothed and un-smoothed delay estimates, we then compared diarisation errors using the MVDR beamformer and the currently used delay-sum beamformer.

Figure 2 shows the data flow for the experiments. Initially, Wiener-filter-based noise reduction is applied to the analogue and digital microphone signals [7] and both smoothed and unsmoothed TDOA values for each of the channels calculated [4]. Enhanced signals are then generated using three techniques: (1) Delay-sum beamforming using smoothed delay estimates; (2) Superdirective beamforming using un-smoothed delay estimates; (3) Superdirective beamforming using smoothed delay estimates. We used the open source BeamformIt toolkit[4] [8] and the AMI project beamforming tools [9].

Speaker diarisation is then performed on the three enhanced signals using two diarisation systems—the SHOuT speech recognition toolkit [10] and the ICSI speaker diarisation system [11][5].

6. RESULTS

Two metrics are used to verify the performance of speaker diarisation systems—the voice activity detection error rate (VER) and the diarisation error rate (DER). The VER is calculated from missed speech and false alarms—that is, speech segments that are not detected as speech and non-speech segments that are identified as speech. In addition to missed speech and false alarms, DER (see [8], chapter 6.1.3, page 162ff) also takes into account the speaker to whom each segment is assigned, and penalises segments assigned to the wrong speaker. In order to account for errors in the reference labels and slight variations in automatic processing, a tolerance of ± 250ms is permitted at the edge of each speech segment.

The VER and DER results for the six meetings in the AD_IMR corpus are given in Table 2. The results show that, for diarisation, the new digital microphone array compares well with the analogue array despite the reduced SNR, producing only marginally increased error rates. This result suggests that MEMS microphone technology provides a viable alternative to expensive analogue devices for speech data capture, and further experiments will be conducted on a variety of speech processing tasks using the DMMA.2

Table 2 also shows that Viterbi smoothing of the TDOA coefficients and delay-sum beamforming provide better results than superdirective beamforming using either smoothed or un-smoothed delays. This may in part be due to the fact that the TDOA smoothing method was optimised for diarisation performance using a delay-sum beamformer, and alternative optimisation may be required in the superdirective case. Also, it is possible that the superdirective beamformer actually removes vital acoustic information from the sidelobes, leading to an increased DER due to the diarisation tools being tuned to acoustic output from a delay-sum beamformer. Analysing the effect of the superdirective beamformer white noise gain constraint GW on the diarisation error rate, it was found that, by tuning GW, the performance gap between the digital and analogue arrays could be reduced. In general, reducing the GW leads to a decrease of the difference in the DER between the analogue and digital arrays, as shown in Figure 3, with

[3] http://www.itl.nist.gov/iad/mig/tests/rt/

[4] http://www.xavieranguera.com/beamformit/

[5] The implementation of the ICSI system evaluated here only uses acoustic features, in contrast to the system used in the ICSI submission to the NIST RT09 evaluation which incorporates TDOA features directly as an input to the diarisation system.

Table 2. % VER and DER for delay-sum (DSB) and superdirective (SDB) beamforming using the ICSI and SHOuT diarisation systems, for analogue and digital arrays. FA denotes false alarms, MS denotes missed speech.

		SHOuT				ICSI			
		DER	VER	FA	MS	DER	VER	FA	MS
DSB	analogue	20.54	2.3	1.3	1	22.49	2.2	1.3	0.9
(TDOA smoothing)	digital	21.89	3	1.5	1.5	22.81	2.9	1.5	1.4
SDB	analogue	29.21	4.8	3.5	1.3	28.17	4.7	3.5	1.2
GW=0.6	digital	35.16	4.9	3	1.9	30.31	4.8	3.1	1.7
modified SDB	analogue	23.11	3.6	1.9	1.7	21.58	3.5	1.9	1.6
GW=0.6 (TDOA smoothing)	digital	25.45	3.7	1.6	2.1	28.82	3.7	1.7	2

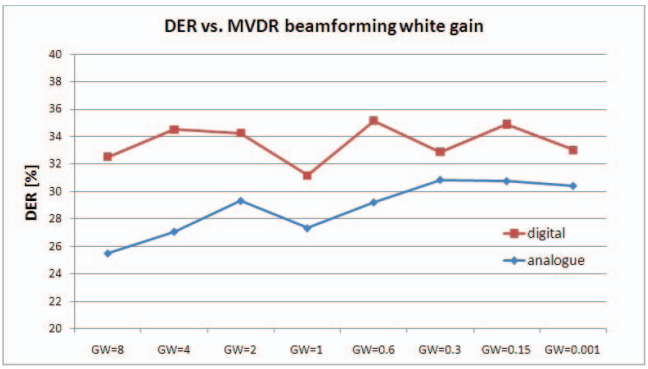


Fig. 3. Effect of white noise gain constraint GW on DER

best performance achieved by setting $GW < 0.15$.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the development of a new digital MEMS microphone array. We have recorded a new corpus of 6 meetings using both the digital array and an analogue array, and annotated 78 minutes of data extracted from the recordings for speech/non-speech and speaker identification. We have compared the performance of two state-of-the-art diarisation systems using both the analogue and digital recordings, and a number of delay estimation and beamforming techniques.

We found that the digital MEMS microphone array approaches the performance of the analogue array when using superdirective beamforming, if the white noise gain constraint of the beamformer is correctly adjusted. In addition, we found that superdirective beamforming, even when using delay estimation smoothing, is unable to match the diarisation performance of delay-sum beamforming and believe this may be caused by mismatch between the beamformer output and the diarisation systems used.

Future work will investigate optimising the TDOA estimation and diarisation system for such a beamformer in order to increase performance. We also plan to record more meetings with the DMMA.2, some with improved speaker tracking and others in an anechoic chamber, to investigate the effects of SNR and reverberation on diarisation performance.

8. REFERENCES

[1] T. Hain, L. Burget, J. Dines, P.N. Garner, A.E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "The AMIDA 2009 meeting transcription system," in *Interspeech*, 2010.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976.

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *ICASSP*, 1997.

[4] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[5] J. Bitzer and K. Uwe Simmer, "Superdirective microphone arrays," in *Microphone arrays: signal processing techniques and applications*, M. Brandstein and E Ward, Eds. Springer Verlag, 2001.

[6] E.P. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *ICASSP*, 2010.

[7] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI front end archive," `http://www.icsi.berkeley.edu/Speech/papers/qio/`, 2002.

[8] X. Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politècnica de Catalunya, 2006.

[9] G. Lathoud, I. McCowan, and D. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Eurospeech*, 2003.

[10] M. Huijbregts, "SHOuT Speech Recognition Toolkit," `http://shout-toolkit.sourceforge.net/`, 2006.

[11] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.