RECOGNITION OF OVERLAPPING SPEECH USING DIGITAL MEMS MICROPHONE ARRAYS

Erich Zwyssig[1,2], *Friedrich Faubel*[3*], *Steve Renals*[1] *and Mike Lincoln*[4]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, Scotland UK
[2]EADS IW, Appleton Tower, Edinburgh, EH8 9LE, Scotland UK
[3]Spoken Language Systems, Saarland University, D-66123 Saarbrücken, Germany
[4]Quorate Technology, Appleton Tower, Edinburgh, EH8 9LE, Scotland UK

ABSTRACT

This paper presents a new corpus comprising single and overlapping speech recorded using digital MEMS and analogue microphone arrays. In addition to this, the paper presents results from speech separation and recognition experiments on this data. The corpus is a reproduction of the multichannel Wall Street Journal audio-visual corpus (MC-WSJ-AV), containing recorded speech in both a meeting room and an anechoic chamber using two different microphone types and two different array geometries. The speech separation and speech recognition experiments were performed using SRP-PHAT-based speaker localisation, superdirective beamforming and multiple post-processing schemes, such as residual echo suppression and binary masking. Our simple, cMLLR-based recognition system matches the performance of state-of-the-art ASR systems on the single speaker task and outperforms them on overlapping speech. The corpus will be made publicly available via the LDC in spring 2013.

Index Terms— MEMS microphones, microphone array, speech separation, WSJ, ASR

1. INTRODUCTION

This paper presents a new multiple microphone array corpus of single and overlapping speech (2012_MMA), together with speech separation and recognition experiments on the corpus. Recordings were made using microphone arrays in two conditions: a hemi-anechoic room and a meeting room. In each of these settings twelve participants were recorded reading Wall Street Journal (WSJ) sentences from prompts, both individually and overlapping in six same-gender pairs, exactly as in the experiments presented by Lincoln et al. for the second PASCAL Speech Separation Challenge [1]. Five circular microphone arrays were used to make simultaneous recordings: two different microphone types (digital MEMS and analogue) were used and the arrays had diameters of 20 cm (16kHz sampling rate) and 4 cm (96kHz and 48kHz sampling rates). We conducted speech separation and recognition experiments on this corpus to investigate the effect of the reduced SNR of the digital MEMS microphones compared to analogue microphones. In our experiments we looked at the effect of post-filtering, echo suppression, and binary masking.

These experiments are, as far as we know, the first ever recordings of single and overlapping speech in a meeting room and hemi-anechoic environment using both conventional analogue and newly available digital MEMS microphones, which are being used increasingly in modern consumer devices.

2. PRIOR WORK

Overlapping speech poses a serious challenge for modern ASR systems. The most systematic work in the field, using recordings of overlapped speech, has used the multi-channel Wall Street Journal audio visual (MC-WSJ-AV) corpus [1], released for the second PASCAL Speech Separation Challenge. Initial experiments on these recordings [2, 3] demonstrated that the ASR word error rate (WER) for overlapping speech can easily be double or triple that of a comparable single speaker scenario. More recent experiments on the single speaker part of the MC-WSJ-AV corpus have shown that it is important for distant speech recognition to use sophisticated front-end processing on multiple input channels [4] rather than just back-end compensation on a single distant channel [5, 6]. The ideal approach might therefore consist of a combination of the two [7].

Although there has been a lot of recent research activity in single speaker distant speech recognition, e.g. the CHiME challenge [8], this has typically involved the artificial creation of data by convolving close-talking speech recordings with a multi-channel room impulse response and then adding noise. Ideally, however, the corpora would be recorded in different natural environments in order to capture the way in which speakers change their speaking style in noise and reverberation [9], and this has motivated our collection of the 2012_MMA corpus.

[*] This author has been supported by the federal republic of Germany through the Cluster of Excellence for Multimodal Computing and Interaction.

3. MEMS MICROPHONE ARRAY

MEMS microphones are replacing analogue microphones at a fast pace in modern consumer devices. These MEMS microphones have the advantages of easier manufacturing and better sensitivity matching, at the cost of a significantly reduced signal to noise ratio (SNR). We have previously demonstrated that the reduced SNR of the MEMS microphone can be compensated for by using MLLR adaptation techniques in speech recognition, and that the automatic speech recognition performance of the MEMS microphones can match conventional analogue ones [10]. In those experiments we used circular arrays of a diameter of 20 cm, similar to those used for data collection in the AMI Meetings Corpus [11].

We have now developed a circular 8-channel microphone array with a diameter of 4 cm which would fit easily into many consumer devices, allowing mobile recording of 8 synchronous channels of audio and therefore enabling super-directive beamforming, state-of-the-art noise reduction, speech separation and dereverberation. Our digital MEMS microphone arrays are built using ADI ADMP441 omnidirectional MEMS microphones with bottom port and I^{2}S output and the Rigisystems USBPAL, a USB 2.0 multi-channel audio interface for Windows PC and MAC OS X. Detailed information on the MC-WSJ-AV and 2012_MMA corpora including the DMMA.3 can be obtained from `http://www.cstr.inf.ed.ac.uk/research/#corpora`.

4. SPEECH SEPARATION

We separate overlapping speech using a combination of spatial filtering and crosstalk cancellation methods [2, 3]. This is achieved via a two-stage approach in which an initial beamforming stage separates the speech based on spatial diversity (Section 4.1), followed by a cross-talk cancellation stage which post-processes the beamformer outputs in order to improve the separation (Section 4.2). We also discuss the speaker localisation system (Section 4.3).

4.1. Superdirective Beamforming

Consider two speakers located at directions

$$\mathbf{a}_k = [\cos\theta_k \cos\phi_k \quad \sin\theta_k \cos\phi_k \quad \sin\phi_k]^T, \qquad (1)$$

$k = 1, 2$, with θ_k and ϕ_k denoting the azimuth and elevation in relation to the array. The directions $\mathbf{a}_k$ translate to time delays $\tau_{k,i} = -\mathbf{a}_k^T \mathbf{m}_i / c$ at the microphone positions $\mathbf{m}_i$, $i \in \{1, \ldots, N\}$, where c denotes the speed of sound. Let $x_i(t)$ denote the signal at the i-th microphone and let $X_i(\omega, t)$ be the corresponding short-time Fourier transforms. Then defining $\mathbf{X}(\omega, t) = [X_1(\omega, t) \ \cdots \ X_N(\omega, t)]$, beamforming may be described as a multiplication by a weight vector $\mathbf{w}_k$:

$$Y_k(\omega, t) = \mathbf{w}_k^H(\omega) \cdot \mathbf{X}(\omega, t). \qquad (2)$$

For the delay-and-sum (DSB) beamformer, we set $\mathbf{w}_k(\omega) = \frac{1}{N}\mathbf{v}_k(\omega)$ where $\mathbf{v}_k$ denotes the array manifold vector

$$\mathbf{v}_k(\omega) = [e^{-j\omega\tau_{k,1}} \ \cdots \ e^{-j\omega\tau_{k,N}}]. \qquad (3)$$

To optimise spatial filtering with respect to reverberant environments, it has been proposed to minimise the total output power under the assumption of a diffuse noise field [12]. This leads to the superdirective beamformer (SDB) [13] whose weight vector is:

$$\mathbf{w}_k(\omega) = \frac{T^{-1}(\omega)\mathbf{v}_k(\omega)}{\mathbf{v}_k^H(\omega) T^{-1}(\omega)\mathbf{v}_k(\omega)} \, . \qquad (4)$$

$T_{i,j}(\omega)$ denotes the coherence of a spherically isotropic noise field: $T_{i,j}(\omega) = \mathrm{sinc}(\frac{\omega}{c}\|\mathbf{m}_i - \mathbf{m}_j\|)$, $i, j \in \{1, \ldots, N\}$. In order to use SDB for speech separation, a beamformer is pointed at each of the speakers. $Y_1(\omega, t)$ and $Y_2(\omega, t)$ are obtained according to (2), and the corresponding separated speech signals $y_1(t)$ and $y_2(t)$ are recovered through inverse Fourier transform followed by overlap-and-add.

4.2. Cross-Talk Cancellation

Since speakers tend to use different frequency bands at one time [14], a post-processing step may be employed in which the beamformer outputs $Y_k(\omega, t)$ are multiplied by a binary mask M_k whose components $M_k(\omega, t)$ identify which frequencies a speaker uses at time t [2, 3]:

$$\hat{S}_k(\omega, t) = M_k(\omega, t) \cdot Y_k(\omega, t), \quad k \in \{1, 2\}. \qquad (5)$$

Near perfect demixing would be possible if the true masks were known [14]. In practice, $M_k(\omega, t)$ needs to be estimated. This can be achieved by comparing the power at the beamformer outputs $Y_1(\omega, t)$ and $Y_2(\omega, t)$ and then allocating the time-frequency unit (ω, t) to the stronger output [2]:

$$\hat{M}_k(\omega, t) = \begin{cases} 1, & |Y_k(\omega, t)|^2 \geq |Y_l(\omega, t)|^2 \quad \forall l \\ 0, & otherwise \end{cases} . \qquad (6)$$

In this work, the $|Y_k(\omega, t)|^2$ are smoothed in time by convolving with a triangular filter kernel. The resulting masks $\hat{M}_k(\omega, t)$ are further processed by Welsh averaging:

$$\bar{M}_k(\omega, t) = \alpha \bar{M}_k(\omega, t-1) + (1 - \alpha)\hat{M}_k(\omega, t) \qquad (7)$$

with $\alpha = 0.9$. Since the optimum window length for time-frequency masking is about 1024–2048 samples (at a sampling rate of F =16kHz) [14], we use an FFT of length $L = 2^{\log_2(F/32)}$, with a window shift of $L/32$.

4.3. Speaker Localisation with a Superdirective SRP

Speaker localisation is carried out using a superdirective variant of the steered response power (SRP-PHAT) method from [15, 16]. The main idea of this approach is to (1) steer an

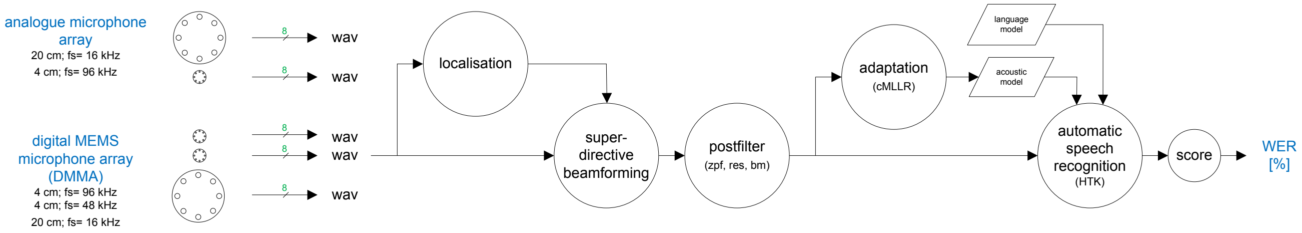

Fig. 1. Speech separation and ASR experiment

SDB in every possible direction and then (2) find the speaker at that position where the output power is maximised. PHAT-weighting is accomplished by applying the SDB to the pre-whitened $\tilde{X}_i(\omega,t) = X_i(\omega,t)/\|X_i(\omega,t)\|$ instead of $\mathbf{X}$.

Once the location of the first speaker has been found, we perform a second SRP iteration in which one beamformer w_1 is fixed on the position of the first speaker. A second beamformer w_2 scans all possible directions for the second speaker. During the calculation of the response power $\int |Y_2(\omega,t)|^2 d\omega$ in a particular direction, the effect of the first speaker is cancelled by processing the output $Y_2(\omega_t) = w_2(\omega)\tilde{X}_2(\omega,t)$ with the binary masking method from Section 4.2. This effectively restricts the localisation to those time-frequency units which are not used by the first speaker.

5. EXPERIMENTS

The 2012_MMA corpus contains recordings of two sets of six male and six female speakers reading sentences from the WSJCAM0 [17] test and development sets in a meeting room ($T_{60} = 180ms$) and a hemi-anechoic room (virtually no reverberation), first alone and then in same-gender pairs. All participants were native British English speakers. The set of prompts for each speaker was selected from one of the sets used in WSJCAM0 and typically contained 17 TIMIT style sentences (for adaptation), 40 sentences from the 5k word (closed vocabulary) sub corpus of WSJCAM0 and 40 sentences from the 20k word (open vocabulary) sub corpus. Recordings were made using five circular microphone arrays (diameter d, sampling rate Fs) in each environment:

- Analogue, d = 20 cm, Fs = 16 kHz
- Analogue, d = 4 cm, Fs = 96 kHz
- Digital, d = 20 cm, Fs = 16 kHz
- Digital, d = 4 cm, Fs = 96 kHz
- Digital, d = 4 cm, Fs = 48 kHz

The recordings were processed as follows (Figure 1). First, sound source localisation was carried out using the audio signal from the 8 channels. Beamforming and post-filtering was then performed and two speakers were extracted from the audio inputs. Speech recognition was carried out on the post-filtered signal, and acoustic model adaptation was performed using the adaptation recordings. Recognition and scoring were conducted using a context-dependent HMM-GMM system using the HTK toolkit [18].

Table 1. Overlapping speaker WER [%] of the ASR experiments on the MC-WSJ-AV corpus

Adaptation	None WER [%]	channel WER [%]	speaker & channel WER [%]
SDB	90.3	67.2	67.2
SDB+ZPF	87.6	63.2	63.16
SDB+RES	81.7	55.3	58.9
SDB+BM	73.8	46.3	48.6

There is an acoustic mismatch between the WSJCAM0 training data and the microphone array recordings which form the test data. To address this we used the adaptation sentences recorded to carry out a two pass constrained maximum likelihood linear regression (cMLLR) adaptation [19] of the model means and variances, similar to our previous experiments [10]. We adapted the models to the individual channels and to the speakers, pooling the 17 adaptation sentences recorded by each speaker. The recognition experiments were then performed on the WSJ-5k data from the matched array. Modifications were necessary for the overlapping speaker experiments because the identity and position of the individual speakers were not known. cMLLR adaptation was therefore carried out for a speaker pair and not the individual speakers.

6. RESULTS AND DISCUSSION

The results presented here were produced following the exact setup described in [10] to ensure validity of the experimental data and in order to be able to compare the results. Baseline experiments were also carried out with the MC-WSJ-AV corpus. This data was recorded with the 8-channel analogue array with a diameter of 20 cm, the same array as used for a subset of the new recordings presented here. The word error rates (WER) achieved are presented in Table 1.

State-of-the-art speech recognition accuracy (WER) using the single stationary speaker data of the MC-WSJ-AV corpus is 12.2% [4]. For the overlapping speaker scenario Himawan et al. [2] achieved 58% WER (40% for the better speaker). McDonough et al. [3] achieved 39.6% WER using a different ASR system.

Table 2. Results from the ASR experiments on the single (WSJ) and overlapping speaker (MSWSJ) corpus in a meeting room and anechoic chamber

	Corpus	**WSJ (IMR)**					**WSJ (anechoic)**				
	Microphone array	Analogue		Digital			Analogue		Digital		
	diameter [cm]	20	4	20	4	4	20	4	20	4	4
	Fs [kHz]	16	96	16	96	48	16	96	16	96	48
	Adaptation	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]	WER [%]
SDB	None	23.2	26.3	45.3	32.3	29.4	18.0	20.6	37.1	21.1	20.8
	cMLLR (channel)	17.9	18.2	29.7	21.4	20.0	16.4	17.6	26.3	17.9	17.9
	cMLLR (speaker & channel)	16.1	17.3	25.6	19.7	18.2	14.4	15.8	24.9	15.0	15.6
SDB+ZPF	None	21.8	26.3	35.3	33.0	29.6	18.0	20.5	36.1	21.0	20.7
	cMLLR (channel)	16.8	18.1	19.3	21.7	20.0	17.0	16.8	25.9	17.9	18.0
	cMLLR (speaker & channel)	13.9	17.0	18.7	20.1	18.2	14.7	14.9	23.8	14.9	15.6

	Corpus	**MSWSJ (IMR)**					**MSWSJ (anechoic)**				
SDB	None	93.4	105.0	97.2	108.8	108.6	93.7	104.8	97.8	107.9	104.7
	cMLLR (channel)	66.7	81.5	64.1	80.9	82.1	67.6	79.4	60.0	81.7	80.0
	cMLLR (speaker & channel)	67.7	83.6	63.0	85.8	85.9	67.4	81.4	59.4	83.1	82.3
SDB+ZPF	None	88.2	102.7	90.2	105.4	107.2	90.4	102.9	94.2	106.3	102.8
	cMLLR (channel)	56.2	77.1	43.2	78.7	79.5	64.3	76.7	59.1	78.9	77.8
	cMLLR (speaker & channel)	55.8	80.5	43.5	81.5	83.4	64.5	78.7	58.4	79.6	80.2
SDB+RES	None	65.3	66.2	72.5	66.9	64.9	58.8	65.2	71.8	72.0	63.9
	cMLLR (channel)	35.4	36.3	39.4	31.9	34.1	30.9	37.6	44.5	49.0	37.8
	cMLLR (speaker & channel)	36.1	37.0	40.8	35.0	36.1	32.4	43.1	45.2	50.8	39.1
SDB+BM	None	59.9	63.2	58.4	60.3	60.3	61.9	75.8	66.6	71.8	62.9
	cMLLR (channel)	31.9	35.8	32.7	33.5	33.5	40.3	47.0	42.4	46.2	42.6
	cMLLR (speaker & channel)	34.3	38.7	34.9	35.4	35.2	39.4	48.0	42.8	48.5	44.0

Our results on the single speaker data (2012_MMA, WSJ), ranging from 13-25%, are in line with those with all five microphone arrays using simple cMLLR adaptation[1]. Speech recognition experiments on the recordings from the hemi-anechoic chamber achieve similar results, as presented in Table 2.

Results using superdirective beamforming (SDB) are similar to our previous results [10], where we demonstrated that the WER gap between the digital and analogue arrays can be compensated for by channel (i.e. microphone array type) adaptation. These results can be improved by a few percent using Zelinski postfiltering [20] (SDB+ZPF). Using speaker and channel adaptation the WERs obtained from the different microphone arrays are almost identical.

For the multi-speaker WSJ speech separation task we achieved a lowest WER of around 35%, again only using simple cMLLR adaptation to the channel. These results were obtained with both residual echo suppression [21] and binary masking. The MSWSJ results are presented in Table 2.

The best results are achieved by using SDB and residual echo suppression (SDB+RES) or binary masking (SDB+BM). Residual echo suppression appears to be more efficient for analogue microphones, while binary masking works better for the MEMS microphones. Speaker and channel adaptation is not efficient for overlapping speech recognition due to the data not being from one, but two speakers. Channel-only adaptation is more efficient as there is more adaptation data.

Results reported here are averages of 6 speaker pairs. We observed that the WER for one speaker is usually significantly better then for the other one, e.g. the reported WER of 31.9% for the analogue microphone array of 20 cm diameter is a product of the average of 24.4% WER for the first better speaker and 39.3% WER for the second speaker. This was already observed during speech separation experiments on the MCSJAV corpus [2].

7. CONCLUSIONS AND FUTURE WORK

In this paper we have demonstrated that the 2012_MMA corpus is a valuable extension to the existing MC-WSJ-AV corpus, allowing research in speech separation on natural speech using recordings from five different microphone arrays, including (digital) MEMS microphones. Using state-of-the-art speech separation, acoustic beamforming techniques, post-filtering and simple constrained MLLR adaptation, we have obtained baseline WERs in line with the state-of-the-art on the distant single speaker task, and demonstrated improved recognition accuracy on the overlapping speech separation and recognition task.

We are currently working with the Linguistic Data Consortium (LDC) to publish the 2012_MMA corpus in Spring 2013.

[1] Note that the digital MEMS microphone array (d=20 cm, Fs = 16 kHz) is a prototype only and shows increased noise and therefore also increased WER. The issues have been resolved with the new array (d = 4 cm)

8. REFERENCES

[1] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005.

[2] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Machine Learning for Multimodal Interaction*. 2008, Springer.

[3] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow, "To separate speech: A system for recognizing simultaneous speech," in *Machine Learning for Multimodal Interaction*. 2008, Springer.

[4] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, 2012.

[5] M.J.F. Gales and Y.Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011.

[6] A. Krueger, O. Walter, V. Leutnant, and R. Haeb-Umbach, "Bayesian feature enhancement for asr of noisy reverberant real-world speech," in *Interspeech Conference*, 2012.

[7] D. Kolossa et al., "CHiME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *CHiME Workshop on Machine Listening in Multisource Environments*, 2011.

[8] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech & Language*, 2012.

[9] D. Pelegrín-García, B. Smits, J. Brunskog, and C.H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *The Journal of the Acoustical Society of America*, vol. 129, pp. 1981–1990, 2011.

[10] E. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[11] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, pp. 181–190, 2007.

[12] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of the Acoustic Society of America*, 1955.

[13] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 39–62. Springer, 2001.

[14] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, 2004.

[15] J.H. Dibiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, Rhode Island 02912, USA, 2000.

[16] K.U. Simmer, J. Bitzer, and C. Marro, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 155–180. Springer, 2001.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 1995.

[18] Cambridge University Engineering Department (CUED), "HTK Speech Recognition Toolkit," `http://htk.eng.cam.ac.uk/`, 2012, [Online; accessed 30-November 2012].

[19] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech and language*, 1998.

[20] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1988.

[21] C. Siegwart, F. Faubel, and D. Klakow, "Improving the separation of concurrent speech through residual echo suppression," in *ITG Symposium Speech Communication*, 2012.