IDIAP RESEARCH REPORT

The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments

Mike Lincoln [*] Iain McCowan [**]
Jithendra Vepa [**] Hari Krishna Maganti [**]

IDIAP–RR 05-69

October, 2005

[*] Centre for Speech Technology Research, Edinburgh, United Kingdom
[**] IDIAP, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-69

The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments

Mike Lincoln Iain McCowan Jithendra Vepa Hari Krishna Maganti

October, 2005

Abstract. The recognition of speech in meetings poses a number of challenges to current Automatic Speech Recognition (ASR) techniques. Meetings typically take place in rooms with non-ideal acoustic conditions and significant background noise, and may contain large sections of overlapping speech. In such circumstances, headset microphones have to date provided the best recognition performance, however participants are often reluctant to wear them. Microphone arrays provide an alternative to close-talking microphones by providing speech enhancement through directional discrimination. Unfortunately, however, development of array front-end systems for state-of-the-art large vocabulary continuous speech recognition suffers from a lack of necessary resources, as most available speech corpora consist only of single-channel recordings. This paper describes the collection of an audio-visual corpus of read speech from a number of instrumented meeting rooms. The corpus, based on the WSJCAM0 database, is suitable for use in continuous speech recognition experiments and is captured using a variety of microphones, including arrays, as well as close-up and wider angle cameras. The paper also describes some initial ASR experiments on the corpus comparing the use of close-talking microphones with both a fixed and a blind array beamforming technique.

1 Introduction

Meetings are an everyday occurrence in the workplace, used as a forum for problem solving, planning, and sharing of ideas within working teams. Within meetings, speech is the predominant mode of interaction between participants. If accurate, easily searchable records of meetings are to be maintained, automatic speech transcription is required. To this end, significant emphasis has recently been placed on adapting state-of-the-art automatic recognition systems to the meetings domain[1, 2].

One of the major problems with recognition of speech in meetings is that of robustly acquiring the speech signal given the adverse conditions (in terms of ASR performance) in which most meetings are held. Meeting rooms are often reverberant (e.g., the instrumented meeting room at the University of Edinburgh has a reverberation time in the region of 0.7s); they suffer from significant background noise, e.g. from projectors and computers within the room, and activities outside the room; and meetings often contain periods in which several people are speaking concurrently. Close-talking microphones alleviate many of these problems and give the highest accuracy from current ASR systems[3], however it is impractical to provide every participant in a meeting with a headset microphone for a number of reasons — the cost of such devices is prohibitive, participants find them obtrusive and feel self-conscious wearing them, and unless radio microphones are used, participants are effectively tethered to one location, unable to act or move naturally. Microphone arrays offer a potential solution to these problems.

A microphone array provides an enhanced version of the input speech based on the location of the speaker. A body of previous work, e.g.[4, 5], has shown that arrays can be an effective alternative to close-talking microphones for single speaker ASR in noisy environments. In addition, in a multi-speaker environment, the directional nature of the array allows discrimination between speakers leading to improved ASR performance for overlapping speech[6]. Primarily due to a lack of appropriate corpora, however, research into microphone array ASR has to date focused on digit recognition tasks, which obviously bear little similarity to speech in meetings. While some larger vocabulary tasks have been investigated, such as in[7], these have used simulated corpora rather than real recordings.

Recently, the NIST RTO4S and RT05S evaluations have provided a comparison of speech recognition on data recorded in real meetings with both close talking and array microphones. The ICSI-SRI system detailed in[1] shows that while microphone arrays achieve lower ASR performance than close-talking microphones, they can significantly improve performance over that of a single distant microphone. While it is desirable to conduct research into array processing techniques for use on real meeting data such as that used in the NIST evaluation, the recognition systems required to give acceptable performance on such tasks are extremely complex, representing the output of many man years of development. As such, it is not feasible for many research teams to test the performance of their algorithms on such data. Even when such systems are available, the time taken to conduct experiments with them can be impractical.

For these reasons, it was decided to record a corpus offering an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. The corpus consists of read Wall Street Journal sentences taken from the test set of the WSJCAM0 database. As such, experiments may be carried out on the data using standard HMM recognition systems which are relatively straightforward to implement. The corpus is recorded in the instrumented meeting rooms constructed for the recording of the AMI Meetings Corpus[8]. The sentences are read by a range of speakers (some 45 in total) with varying accents (including a number of non-native English speakers). Sentences are read according to a number of scenarios including a single stationary speaker, a single moving speaker, and multiple concurrent speakers. During recordings, all speakers wear lapel and headset microphones, and audio from two eight element microphone arrays is also captured. The rooms also provide synchronised video recordings including close-up views of the speakers' faces, as well as wide-angle views of the entire room. As such, the data is suitable for a wide variety of research tasks including :

- development of microphone array ASR front-end processing systems,

- audio-visual ASR,
- audio-visual person tracking,
- integration of audio-visual person tracking with microphone array ASR processing,
- recognition of accented and non-native English speech,
- recognition of overlapped speech.

The paper is organised as follows. Section 2 specifies the new Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus. Section 3 then details two beamforming techniques — one utilising information concerning the position of the speaker and the array geometry; the other estimating the beamforming filters 'blind' from the recorded signals. Section 4 describes some initial experiments on a subset of the MC-WSJ-AV corpus by comparing recognition accuracy of these beamforming techniques to headset and lapel recordings. Section 5 presents concluding remarks and future plans.

2 The MC-WSJ-AV Database Specification

This section overviews the recording of the MC-WSJ-AV database, an audio-visual corpus suitable for, among many other tasks, development of microphone array algorithms.

2.1 Data Acquisition

Three sites are involved in the recording of the data: The Centre for Speech Technology Research, Edinburgh (UEDIN), The IDIAP Research Institute, Switzerland (IDIAP) and TNO Human Factors, The Netherlands (TNO). Instrumented meeting rooms installed at the three sites allow the capture of fully synchronised audio and video data as described in[8]. The layout of the UEDIN room, showing the positions of the microphone arrays and video cameras, plus the six reading positions, is shown in Figure 1. The room contains two eight-element circular microphone arrays, one mounted at the center and one at the end of the meeting room table. In addition, the speakers are provided with close-talking radio headset and lapel microphones. Close-up cameras give facial views of the speakers while in the seated positions, and wide angle cameras give views of the entire floor area of the room. The TNO and IDIAP rooms contain similar recording equipment, but differ in their physical layout and acoustic conditions.

2.2 Speaking Conditions

The data consists of recordings of Wall Street Journal sentences read in instrumented meeting rooms under three conditions:

1. *Single Speaker Stationary.* For this condition the speaker is asked to read sentences from six positions within the meeting room — four seated around the table, one standing at the whiteboard and one standing at the presentation screen. One sixth of each speaker's sentences are read from each position.

2. *Single Speaker Moving* For this condition the speaker is asked to move between the six positions while reading the sentences. The speaker begins reading at position 1 and moves to position 2 while reading the first sentence. They then move back to position 1 while reading the next, and continue alternating between these two positions with each sentence. When a sixth of their sentences have been read, the speaker then alternates between positions 2 and 3 for the next sixth, then positions 3 and 4 for the next, and so on.

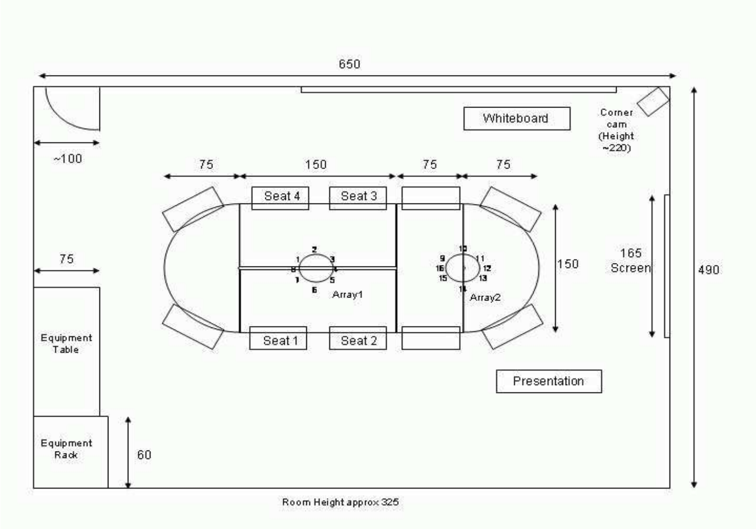

Figure 1: The layout of the UEDIN Instrumented Meeting Room (Measurements in cm). Array microphones are numbered 1-16. Cameras are mounted under Array 1 to give closeup views of participants in the seated locations. The six reading locations are indicated as Seat 1-4, Presentation and Whiteboard.

3. *Overlapping Speakers (Stationary)* Here, two speakers are asked to simultaneously read their sentences from different positions within the room. The speakers remain in the same positions for the entirety of these recordings and separate recordings are made from each of the 15 pairs of positions.

Speakers were asked to read naturally with no constraints placed on speaking style, pronunciation or accent. With the exception of the data recorded at the UEDIN site, which consists of all native speakers, the corpus contains many non-native English speakers.

2.3 Corpus Structure

MC-WSJ-AV is divided into one development set (DEV) and two evaluation sets (EVAL1 and EVAL2) for each of the three conditions. The selection of read sentences for these sets is based on the development and evaluation sets of the WSJCAM0 British English corpus[9]. Each speaker's prompts contain 17 adaptation sentences, 40 sentences from the 5000-word sub-corpus and 40 sentences from the 20000-word sub-corpus, giving a total of 97 sentences per speaker. The WSJCAM0 development set contains 20 speakers and the evaluation-1 and evaluation-2 sets 14 speakers each. Given that, for our overlap condition, the 6 speaking locations give 15 unique pairs of seats, it was decided that we would use the prompts from 15 WSJCAM0 speakers in each of our sets. THE MC-WSJ-AV DEV set sentences are therefore the first 15 speakers prompts from the WSJCAM0 development set. EVAL1 sentences are the prompts from the entire WSJCAM0 evaluation-1 set, plus one of the speakers from the WSJCAM0 development set not already used. Similarly the EVAL2 sentences are the prompts from the entire WSJCAM1 evaluation-2 set, plus one from the remaining WSJCAM0 development set. Table 4 shows the recording schedule for the DEV set.

2.4 Status

To date, the single speaker stationary data at UEDIN has been recorded, segmented and checked, and only this data was used for the current experiments. The single speaker moving data from the UEDIN site has been recorded and will be segmented and checked. The UEDIN overlap data is scheduled for recording in July 2005. All the data from the IDIAP site has been recorded and is currently being segmented and checked and will be completed by September 2005. The data from the TNO site will be recorded in late 2005. The data is expected to be released to the research community towards the end of 2005. A sample of the data is available from http://mmm.idiap.ch/MC-WSJ-AV

3 Beamforming Techniques

In this section we present two beamforming approaches that may be applied to the microphone array recordings in the corpus as the front-end to an ASR system. The first approach relies on knowledge of the recording environment and array geometry, while the second is a completely automatic system, requiring no knowledge of the speaker location or microphone placement. In both cases, the superdirective beamforming technique is used to calculate the channel filters w_n maximising the array gain, while maintaining a minimum constraint on the white noise gain. This technique is fully described in[10, 11]. The optimal filters are calculated as:

$$\mathbf{w} = \frac{\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}} \tag{1}$$

where $\mathbf{w}$ is the vector of microphone filters,

$$\mathbf{w}(f) = \begin{bmatrix} w_1(f) & w_2(f) & \ldots & w_N(f) \end{bmatrix}^T, \tag{2}$$

$\mathbf{d}$ is the propagation vector between the source and each microphone,

$$\mathbf{d}(f) = \begin{bmatrix} \alpha_1 e^{-2\pi f\tau_1} & \alpha_2 e^{-2\pi f\tau_2} & \ldots & \alpha_N e^{-2\pi f\tau_N} \end{bmatrix}^T, \tag{3}$$

and $\mathbf{\Gamma}$ is the noise coherence matrix. The two techniques used in the experiments differ only in the way in which the channel scaling factors α_n, delays τ_n, and noise coherence matrix are calculated, as described in the following sub-sections.

3.1 Fixed Beamformer

The fixed beamformer technique relies on knowledge of both the microphone array geometry, and a speaker location. Given $\mathbf{p}_n$ as the location vector of microphone n, and $\mathbf{p}^{(s)}$ as the location vector of the speaker, taking the first microphone as the reference for convenience, then the channel scaling factors, delays and noise coherence matrix can be calculated as [10, 11]:

$$\alpha_n = \frac{d_1^{(s)}}{d_n^{(s)}} \tag{4}$$

$$\tau_n = \frac{d_n^{(s)} - d_1^{(s)}}{c} \tag{5}$$

$$\Gamma_{nm} = \mathrm{sinc}\left(\frac{2\pi f d_{nm}}{c}\right) \tag{6}$$

where $d_n^{(s)} = \|\mathbf{p}^{(s)} - \mathbf{p}_n\|$ and $d_{nm} = \|\mathbf{p}_n - \mathbf{p}_m\|$.

As described in the previous section, for the stationary speaker scenarios in the corpus the speaker occupies one of 6 known locations in the room. For the experiments in this paper, we pre-calculate fixed beamforming filters for each of these locations. Then, for each utterance, we beamform simultaneously to each location and then select the one with the highest energy. In this way, the fixed beamforming technique used in the following experiments is in fact a simple tracking beamformer performing steered response power localisation over six discrete locations.

3.2 Blind Beamformer

The fixed technique described above relies on prior knowledge of the microphone array geometry, as well as accurate channel gain calibration and sample-synchronous acquisition. In many practical cases, these assumptions may not be valid, necessitating more general methods.

Assuming there is one dominant speaker during a given short-term speech frame, then α_n may be estimated as the ratio between the measured frame energy of the reference microphone and microphone n, and τ_n may be estimated by finding the peak in their generalised cross-correlation (GCC) function[12]. Assuming predominantly stationary background noise, The noise coherence matrix elements Γ_{nm} may be estimated using averaged spectra of low energy frames (assumes predominantly stationary background noise) or else simply set to an identity matrix (assumes predominantly incoherent noise, equivalent to delay-sum beamforming).

In the blind beamformer used in the following experiments, the above terms are estimated as described for each short-term input frame, and the beamforming filters updated according to Equation 1. This technique was used as an ASR system front-end for the multiple distant microphone (MDM) condition in the recent Spring 2005 NIST Rich Transcription evaluation[2], and was in part based on the approach taken in the ICSI-SRI-UW system from the 2004 evaluations[1].

4 Experiments and Results

One of the main uses of the MC-WSJ-AV corpus will be as a resource for research of microphone array processing ASR front-ends. This section presents some initial experiments comparing the beamforming approaches described in the preceding section with the headset, lapel and single distant microphone (SDM) recordings.

4.1 Recognition System and Task

A baseline speech recognition system was trained using HTK on the WSJCAM0 database. The training set consists of 53 male and 39 female speakers, all with British English accents. The system consists of approximately 11000 tied-state triphones with three emitting states per triphone and 6 mixture components per state. 52-element feature vectors were used, comprising 13 MFCCs (including the 0th cepstral coefficient) with their first, second and third order derivatives. The dictionaries used were generated from that developed for the AMI NIST RT05S system[3], and the language models are the standard MIT-Lincoln Labs 5k and 20k Wall Street Journal trigram language models. The baseline system, with no adaptation, gives 9.91% WER on the WSJCAM0 si_dt5a 5000-word task and 20.44% WER on the si_dt20a task 20000 word task. These results are comparable to those reported in the SQALE evaluation using the WSJCAM0 database[13].

The test data used in the experiments are the Stationary Speaker EVAL1 sentences recorded at UEDIN. Results are reported on the 5000-word task, giving a total of 189 test sentences comprising approximately 23 minutes of speech. The experiments compare results on the headset, lapel, and single distant (from Array 1, as indicated in Figure 1) microphone data to the output of both the fixed beamformer and blind beamformer techniques described in Section 3.

4.2 Baseline Experiments

Recognition results using the baseline HMMs with no adaptation are shown in column 1 of Table 1. The performance using Headset data is comparable to that obtained on the original WSJCAM0 recordings, however the recognition deteriorates significantly when data from the other channels is used. It is encouraging to note, however, that the results on beamformed data are significantly better than the single distant microphone (SDM) case.

Channel	None	Channel adaptation	S.D. adaptation
Headset	14.8	14.0	12.3
Lapel	26.3	20.2	18.0
Fixed Beamformer	48.6	35.6	28.1
Blind Beamformer	55.2	36.5	31.6
SDM	87.6	73.3	66.5

Table 1: % Word error rates for the 5k closed vocabulary task using stationary speaker MC-WSJ-AV data from the UEDIN room.

4.3 Acoustic Channel Adaptation

It was hypothesised that the deterioration was in part due to channel mismatch between the training and test conditions and an experiment was conducted to verify this. To compensate for the mismatch the baseline models were adapted using a static, two pass, MLLR[14] adaptation. In the first pass a global transformation was performed, and in the second, specific transforms for speech and silence models are calculated. Adaptation data was taken from the UEDIN stationary speaker DEV set and consisted of 432 sentences comprising approximately 50 minutes of speech. Adaptation data was matched to the testing condition (that is, headset data was used to adapt models for headset recognition, lapel data was used to adapt for lapel recognition, etc.). Results for the matched adaptation case are given in column 2 of Table 1. As expected, the headset results show little improvement over the unadapted case since the training and test data would already be closely matched before adaptation. Both the fixed and blind beamformers show significant improvements, giving approximately the same performance. Other studies[6] have shown that recognition results from beamformed channels are comparable, or even better than those from lapel microphones. This is not observed in the current experiments — the lapel data showing approximately 15% improvement over the beamformer — likely due to the effects of high reverberation. The lapel will naturally have a higher ratio of direct-to-reverberant speech energy (due to proximity to the speaker), and may also benefit by being mounted on a large sound absorbing surface (the speaker), as opposed to the acoustically-reflective wooden array mount. This is borne out in the SDM results which show a 53% decrease in performance over the lapel. The SDM microphone is identical to the lapel, the only difference being that it is mounted in the array, and is thus located farther from the speaker's mouth.

Speaker	Gender	Accent	Headset	Lapel	Fixed	Blind	SDM
1	Male	British	7.6	13.6	23.4	26.6	70.3
2	Male	British	8.8	11.2	22.6	23.2	53.1
3	Female	British	10.5	20.4	33.9	31.0	72.1
4	Female	British	21.5	28.9	40.5	41.3	83.2
5	Male	Canadian	20.0	25.8	53.3	55.9	86.3

Table 2: % Word error rates per speaker - channel adaptation

A breakdown of the results per speaker for the Channel adapted case are shown in Table 2. These results show a large variation in performance between speakers — Male British speakers in general outperforming either the female, or non British males. As mentioned above, the majority of the training set were Males with British English accents and as such the models would better match the British English Males in the test set.

Speaker	Gender	Accent	Headset	Lapel	Fixed	Blind	SDM
1	Male	British	8.7	13.4	20.4	25.1	66.2
2	Male	British	7.7	9.5	18.3	22.3	42.5
3	Female	British	9.2	19.5	22.5	20.8	59.6
4	Female	British	19.0	25.7	30.6	38.8	81.7
5	Male	Canadian	16.1	25.3	45.0	47.2	82.1

Table 3: % Word error rates per speaker - Speaker dependant adaptation

4.4 Speaker Dependant Adaptation

To overcome this, a second adaptation experiment was performed in which the 17 adaptation sentences recorded by the test speakers were used to generate speaker dependant transforms. Again these transforms were generated with data recorded from matched channels and therefore accounted for variations in both channel and the talkers speaking style. The results of this experiment are given in column 3 of Table 1 and show improvements over the channel adapted case, while retaining the same trend of results across channels. A breakdown of results per speaker for the Speaker adapted case is given in Table 3. As expected the largest improvements are seen in the females and, in particular, the Canadian speaker, whose speech would not have been represented in the training set at all. Since many of the speakers from recorded at the TNO and IDIAP locations will be non-native speakers (due to the lack of native English speakers at these sites) it is likely that speaker dependant adaptation will provide significant improvements on much of the rest of the corpus when it becomes available.

While the recognition accuracy of the array output is worse than that of the lapel data, even in the adapted case, it is far closer to the performance of the close talking microphones than that of a single distant microphone. The results also show that, while error rates from the fixed beamformer are consistently lower than those from the blind algorithm, the difference is small — less than 1% in the channel adapted case. This suggests that, if information about the array geometry and speaker locations is known, it should be used in estimating the beamformer filters. However if this information is not available then estimates of the delays using an automatic procedure still gives significant improvements over the SDM case. This information is of direct use in circumstances such as the NIST RT05s evaluation[15] in which layout information was only available for some of the recording rooms.

5 Conclusions and Future Work

In this paper, the specification of a new multi-channel audio visual read speech corpus has been presented. The database is currently being recorded and annotated. The corpus provides data recorded in number of instrumented meeting rooms suitable for a wide variety of tasks of application to the analysis of speech in meetings. Recognition results on a subset of the new corpus using data from headset and lapel microphones, two beamforming techniques and a single distant microphone have been compared. It has been shown that, when channel adaptation is applied to the acoustic models, the array techniques provide recognition accuracies far superior to those obtained using a single distant microphone. It has also been shown that, if information about the room layout is available, using it to estimate the beamformer filters can provide small improvements in accuracy over blind estimation of the filters. According to our current schedule, the corpus is expected to be completed and prepared for distribution by late 2005. Ongoing work will investigate more sophisticated beamforming techniques, in particular focussing on the problems of moving speakers and overlapping speech.

Stationary Speaker		Moving Speaker		Overlapping Speakers					
Speaker	Prompts	Speaker	Prompts	Speaker 1	Seat	Prompts	Speaker 2	Seat	Prompts
IDI-1	Dev 1	IDI-1	Dev 6	IDI-1	Seat 1	Dev 11	IDI-2	Seat 2	Dev 12
IDI-2	Dev 2	IDI-2	Dev 7	IDI-2	Seat 1	Dev 13	IDI-3	Seat 3	Dev 12
IDI-3	Dev 3	IDI-3	Dev 8	IDI-3	Seat 1	DEV 13	IDI-4	Seat 4	Dev 14
IDI-4	Dev 4	IDI-4	Dev 9	IDI-4	Seat 1	Dev 15	IDI-5	Wboard	Dev 14
IDI-5	Dev 5	IDI-5	Dev 10	IDI-5	Seat 1	Dev 11	IDI-1	Pboard	Dev 15
EDI-1	Dev 6	EDI-1	Dev 11	EDI-1	Seat 2	Dev 1	EDI-2	Seat 3	Dev 2
EDI-2	Dev 7	EDI-2	Dev 12	EDI-2	Seat 2	Dev 3	EDI-3	Seat 4	Dev 2
EDI-3	Dev 8	EDI-3	Dev 13	EDI-3	Seat 2	Dev 3	EDI-4	Wboard	Dev 4
EDI-4	Dev 9	EDI-4	Dev 14	EDI-4	Seat 2	Dev 5	EDI-5	Pboard	Dev 4
EDI-5	Dev 10	EDI-5	Dev 15	EDI-5	Seat 3	Dev 1	EDI-1	Seat 4	Dev 5
TNO-1	Dev 11	TNO-1	Dev 1	TNO-1	Seat 3	Dev 6	TNO-2	Wboard	Dev 7
TNO-2	Dev 12	TNO-2	Dev 2	TNO-2	Seat 3	Dev 8	TNO-3	Pboard	Dev 7
TNO-3	Dev 13	TNO-3	Dev 3	TNO-3	Seat 4	Dev 8	TNO-4	Wboard	Dev 9
TNO-4	Dev 14	TNO-4	Dev 4	TNO-4	Seat 4	Dev 10	TNO-5	Pboard	Dev 9
TNO-5	Dev 15	TNO-5	Dev 5	TNO-5	Wboard	Dev 6	TNO-1	Pboard	Dev 10

Table 4: Recording schedule for Development set. For the single speaker cases, the EVAL1 and EVAL2 sets follow an identical schedule, using different speakers and prompts. For the overlap case, the positions are rotated across sites for the EVAL1 and EVAL2 sets, such that data is recorded from every pair of positions at every site.

6 Acknowledgments

Many thanks are due to all the recording participants who undertook a difficult task with patience and good humour.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication).

References

[1] Stolke, A. et al., "Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system," NIST RT04 Workshop, 2004.

[2] Hain, T. et al., "The 2005 AMI System for the Transcription of Speech in Meetings," NIST RT05 Workshop, Edinburgh, UK, 2005.

[3] Hain, T. et al., "The Development of the AMI system for the Transcription of Speech in Meetings," Proc. MLMI 05, Edinburgh, UK, 2005.

[4] Omologo, M. and Matassoni, M. and Svaizer, P "Speech recognition with microphone arrays" in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. 2001, chapter 15, pages 331–353. Springer, 2001.

[5] McCowan, I. and Marro, C. and Mauuary, L. "Robust speech recognition using nearfield superdirective beamforming with postfiltering," Proc. ICASSP 2000, 2000, vol.3, pages 1723-1726.

[6] D. Moore and I. McCowan "Microphone Array Speech Recognition: Experiments on Overlapping Speech in Meetings," Proc. ICASSP 2003, 2003.

[7] Seltzer, M. and Raj, B. "Calibration of microphone arrays for improved speech recognition," Proc. of Eurospeech 2001, 2001.

[8] Carletta, J. et al., "The AMI Meeting Corpus: A Pre-Announcement," Proc. MLMI 05, 2005.

[9] T. Robinson and J. Fransen and D. Pye and J. Foote and S. Renals "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," Proc. IEEE ICASSP, pages 81–84, 1995.

[10] Cox, H. and Zeskind, R. and Owen, M. "Robust adaptive beamforming," IEEE Transactions on Acoustics, Speech and Signal Processing, pages 1365-1376, October 1987.

[11] Cox, H. and Zeskind, R. and Kooij, I. "Practical supergain," IEEE Transactions on Acoustics, Speech and Signal Processing, pages 393-397, June 1986.

[12] Knapp, C. H. and Carter, G. C. "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics, Speech and Signal Processing, pages 320-327, August 1976.

[13] Young, S. et al., "Multilingual large vocabulary speech recognition: the European SQALE project," Computer Speech and Language, pages 73-89, 1997.

[14] Leggetter, C. and Woodland, P. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," Computer Speech and Language, pages 171-185, 1995.

[15] "The Rich Transcription Spring 2005 Evaluation (RT-05S)," http://nist.gov/speech/tests/rt/rt2005/spring/