Continuous Microphone Array Speech Recognition on Wall Street Journal Corpus

Hari Krishna Maganti [*] Jithendra Vepa [*]
Hervé Bourlard [*]

IDIAP–RR 05-47

August, 2005

[*] 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 05-47

CONTINUOUS MICROPHONE ARRAY SPEECH RECOGNITION ON WALL STREET JOURNAL CORPUS

Hari Krishna Maganti Jithendra Vepa Hervé Bourlard

AUGUST, 2005

Abstract. In this paper, we present a robust speech acquisition system to acquire continuous speech using a microphone array. A microphone array based speech recognition system is also presented to study the environmental interference due to reverberation, background noises and mismatch between the training and testing conditions. This is important in the context of smart meeting rooms of Augmented MultiParty Interaction (AMI) project which aims at significant development of conversational speech recognition. In this regard, an audio-visual database containing the Wall Street journal phrases was recorded in a real meeting room for the stationary speaker, moving speaker and overlapping speech scenarios. We carried out speech enhancement and continuous speech recognition experiments on stationary speaker data. Using a microphone array with beamformer followed by a postfilter enhances speech quality slightly inferior to that of close-talk headset,and better than lapel. We achieved a significant reduction in word error rates using models adapted based on maximum linear likelihood regression (MLLR) and maximum-a-posteriori (MAP) approaches. Though the error rates of the microphone array data are larger than those of headset data, they are significantly smaller compared to the error rates of lapel data.

1 Introduction

Most speech recognizers perform well only for small vocabularies under matched training and testing conditions. The present best speech recognition systems are able to achieve 1 % error rate in speaker-dependent isolated word recognition task with 20,000 word vocabulary [1] and less than 0.5 % word error-rate in speaker independent digit recognition task [2]. For speaker-independent continuous speech recognition word error-rate of 5 % and 10 % is possible [3]. This high level recognition accuracy is achievable only with matched training and test data. In real world conversational settings like meetings,the microphone arrays can be used for speech acquisition as they provide high SNR needed for speech recognition applications. However, the performance degradation can be observed when these are deployed in meeting rooms where input speech is degraded due to changes in acoustical and articulatory characteristics of the speech.

In this paper, we investigate the use of microphone arrays for speech enhancement and speech recognition in meetings. Earlier studies have shown microphone arrays based on beamforming techniques can replace close-talk headset for speech recognition in real reverberant environments [4, 5]. The effects of using a speech recognition system, trained on clean data when used for distant microphone array system are discussed here. While the previous studies signified the robustness of microphone array system to various background noise sources, the speech recognition experiments were limited to numbers and words. The performance of the current state-of-the-art systems trained on clean data deteriorate when used with microphone array speech data in meetings.A perfectly trained isolated word recognizer could recognize only 30 % when white noise with 10 dB SNR was added to the signal [6]. The word error-rate of the SPHINX speech recognition system raised from 15 % to 80 % when table-top microphone was used in place of close-talking microphone [7]. This degradation results from low quality speech signal, different training and testing scenarios and when the testing speaker's native language is not the same as with which the system was trained.

The microphone array based speech recognition research suffers from the lack of appropriate corpus as most of the speech corpora comprise of single-channel data. Hence, we have started recording an audio-visual database based on the Wall Street journal phrases in a real meeting room for the stationary speaker, moving speaker and overlapping speech scenarios. The database comprises of non native English speakers with no constraints on speaking styles and accents. It has been designed to suit speaker independent audio-visual based continuous speech recognition thus enabling realistic evaluation of multimodal component technologies.Three instrumented meeting rooms at IDIAP Research Institute, Switzerland, The Centre for Speech Technology Research, Edinburgh and TNO Human Factors, The Netherlands are involved in the recording of the data.

This work presents the results of initial experiments, using a part of the database to show continuous speech recognition based on array processing techniques. The clean training data was from Wall Street Journal-based CSR Corpus [8]. To reduce the mismatch between the training and the testing conditions baseline HMM models were adapted using maximum linear likelihood regression (MLLR) and maximum-a-posteriori (MAP) techniques to obtain better recognition results. We discuss performances of speech recognition on data obtained from different audio channels; *headset, lapel and microphone array*.

This paper is organized as follows. In section2 we describe the physical setup and database recording procedure. Section3 details the microphone array processing for speech enhancement, while continuous speech recognition and experimental results are described in Section4. Conclusions and outline for future work are given in section5.

2 Physical setup and Recording Procedure

The recording setup comprises of three widely spaced and four close-up CCTV cameras, two microphone arrays, headset, lapel, manikin connected to a fully synchronized capture devices. The meeting room dimensions are 8.2mx3.6mx2.4m containing a 4.8mx1.2m rectangular table [9]. The audio sensors are configured as an 2 eight-element, circular equi -spaced microphone array centered on the table and on the roof, with diameter 20cm, and composed of high quality miniature electret microphones. Additionally, lapel and close-talk microphones are available to each speaker. Two cameras on opposite walls record frontal views of participants, including the table and workspace area, and have non-overlapping fields-of-view (FOVs). A third wide-view camera looks over the top of the participants towards the white-board and projector screen. Four close-up cameras capture close views of different positions. The video sensor array provides 720x576 resolution images at 25 frames per second, while audio was recorded at 16 kHz with 16 bits per sample. Top view of the entire system is shown in fig. 1 and sample of pictures from the experimental environment are shown in fig. 2.

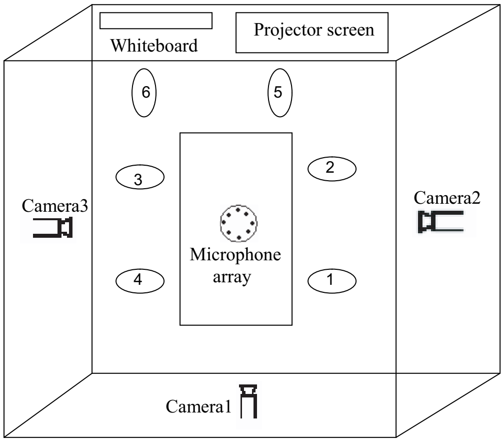

Figure 1: Top view of the physical setup

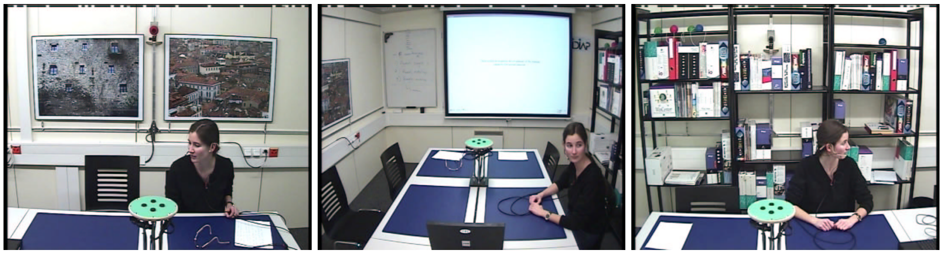

Figure 2: Right,center and left views of the meeting room

The major objective is to design a flexible database for realistic evaluation of component research areas in particular speaker localization, tracking, speech enhancement and recognition. The database consists of three major sections, first single speaker reading 1/6 sentences from 6 locations, second single speaker moving around the area of focus in the meeting room and the third pairs of overlapping speakers. To combat the effects of overlapping speech, which hinder the performance of speech recognition, the part of database is dedicated to the overlap speech. Each section comprises of 15 speakers hence resulting in a total of 45 recordings. The speakers were asked to read the phrases from the Wall Street journal (100 for each recording). The speakers were mainly non native English speakers and the recordings were as natural as possible with no constraints on speaking styles,accents and pronunciations.Around 20 hours of data including stationary, moving, overlapping speech was recorded.

3 Microphone Array Speech Enhancement

Microphone array speech enhancement processing involves transformation of low quality noisy speech into a high quality speech signal as close to close-talk headset. The microphone array speech enhancement system which includes a filter-sum beamformer followed by a post-filter stage as shown in fig. 3 is similar to the system presented in [10].

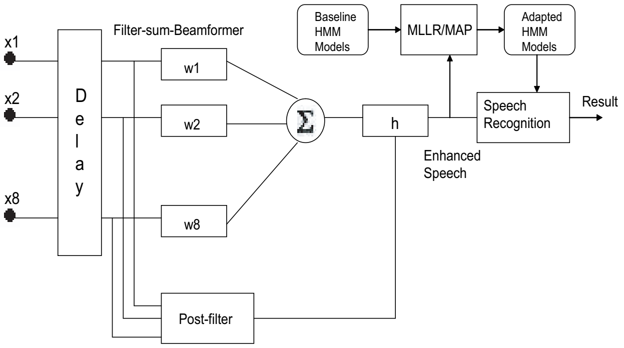

Figure 3: Integrated microphone array speech recognition system

3.1 Beamformer

Beamformer steer an array of microphones towards the active speaker direction through an array signal processing and geometry rather than physically moving the array. The Beamformer uses the superdirective technique to calculate the channel filters maximizing the array gain, maintaining a minimum constraint on the white noise gain. This detailed explanation of this technique is described in [11].

3.2 Postfilter

Beamformer followed by post-filtering significantly improve the enhancement of the speech signal by reducing the background noise. In Marro et al [12], it is mathematically demonstrated that by associating post-filter with the output of the microphone array suppresses the uncorrelated components and improves the performance of the beamformer.Assuming that the noise and reverberations form a diffuse noise field and hence are uncorrelated at each microphone of the array. Wiener filter transfer function [11] can be estimated from the cross- and auto-spectral densities of the input channels. This estimate can be made more robust by averaging the spectral densities over all possible sensor combinations and can be written as,

$$W(f) = \frac{\frac{2}{P(P-1)} \sum_{i=1}^{P-1} \sum_{j=i+1}^{P} \hat{\phi}_{x_i x_j}(f)}{\frac{1}{P} \sum_{i=1}^{P} \hat{\phi}_{x_i x_i}(f)} \quad (1)$$

Where f is the frequency index, P is the number of microphones, $\hat{\phi}_{x_i x_j}$ is the estimated cross-power spectral density between the time compensated microphone signals i and j and $\hat{\phi}_{x_i x_i}$ the estimated power spectrum density of the delayed signal of the i-th microphone.

In [13], a detailed analysis of the Zelinski array post-filter modified to handle wide range of noise-fields has been shown to improve the broadband noise reduction of the array, and lead to better performance in speech recognition applications. The average segmental signal to noise ratio (SNR) is calculated to evaluate the effectiveness of the microphone array in acquiring clean speech. All the results are calculated with respect to the level on single table-top microphone to normalise for different levels of individual speakers and hence represent SNR enhancement(SNRE). The SNRE results in Table 1 show that the signal quality of the microphone array is slightly inferior to that of close-talk headset but better than lapel.

Table 1: SNRE results for different data channels

Channel	**SNRE(dB)**
Headset	29.7
Lapel	24.3
Array	27.6

4 Experiments and Results

In our recognition experiments, a full HTK based recognition system [15] trained on original close talking Wall Street Journal corpus was used. The acoustic feature set has 39 elements, which comprises of 13 Mel frequency cepstrum coefficients (MFCC) with mean removal and their first and second derivations computed using a 25-ms Hamming window shifting every 10 ms. Our system used 3 state per HMM context-dependent triphone models with around 9000 tied states, each modelled by a mixture of 16 Gaussians. The baseline system gave a WER of 20 % using the clean test data from the original Wall Street journal corpus.

Table 2: Adaptation and test data description

Data	**Number of speakers**	**Number of utterances**
Adaptation	7	91
Testing	5	65

As mentioned before, we only used stationary speaker data for these experiments. The details of this data is presented in Table 2. We used around 12 minutes data of 7 speakers for adaptation and 8 minutes data of 5 speakers for testing. All of our speakers, except two, were non-native English speakers. The adaptation data have been recorded under the same conditions as the test data. In Table 3 , we show word error rates of headset, microphone array and lapel data using baseline models, i.e. models trained on original WSJ corpus. The headset data produces a much lower word error rate than microphone array data and lapel data.

Table 3: WERs obtained from baseline models

Channel	**WER (%)**
Headset	42.5
Lapel	67.0
Array	64.2

To compensate the mismatch between the training and the test conditions, we adapted the baseline HMM models. For adaptation we used both maximum likelihood linear regression (MLLR) and maximum-a-posteriori (MAP) approaches. In MLLR, we used a static two-pass approach, where in the first pass a global transformation was performed. In the second pass a set of more specific transforms was estimated using the global transform . Since we have a small amount of adapation data, we used MLLR transformed means as the priors for MAP adaptation. The performace of each of these three data channels were improved when we used models adapted using MLLR and MAP techniques.

Table 4: WERs obtained from MLLR adapted models

	WER (%)		
Channel	Headset-adapted models	Lapel-adapted models	Array-adapted models
Headset	34.5	–	–
Lapel	60.4	53.5	–
Array	51.3	–	46.1

The word error rates for headset, microphone array and lapel datasets obtained using MLLR adaptation on headset data are presented in Table 4. A significant reduction in WERs can be seen in all of the three data channels. We also adapted models using microphone array data as well as lapel data and used them for respective datasets. Using these channel specific adapted models, we achieved further reduction in word error rates.

Table 5: WERs obtained from MAP adapted models

	WER (%)		
Channel	Headset-adapted models	Lapel-adapted models	Array-adapted models
Headset	33.5	–	–
Lapel	59.6	52.2	–
Array	49.7	–	47.9

Table 5, presents word error rates of all the three data channels using MAP adapted models. We obtained further reduction (around 1% absolute) using MAP adapted models, except in the case of microphone array data. This may be due to insufficient data for MAP adaptation, compared to other data channels (headset and lapel), this data may probably contain reverbation and background noises. In MAP adapatation also, we obtained good performance using models adapted using the respective datasets.

5 Conclusions and Future Work

In this paper we present a framework for microphone array based speech acquisition and recognition. For this we recorded an audio-visual database based on the Wall Street journal phrases in a real meeting room for the stationary speaker, moving speaker and overlapping speech scenarios. The database comprises of non native English speakers with no constraints on speaking styles and accents. The speech enhancement experiments showed that using a beamformer followed by a postfilter enhanced speech quality slightly inferior to that of close-talk headset,and better than lapel.

A full HTK based recognition system trained on original close talking Wall Street Journal corpus was used for our experiments. When using these models, the headset data produced a much lower error rate compared to lapel data and microphone array data,however microphone array data performance was better than the lapel data. The MLLR and MAP adaptations helped in reduction of word error rates significantly. The MAP adapted models resulted in slightly less word error rates compared MLLR adapted models. With more data to follow from the partner institutions we hope to improve the recognition performances further.

Number of issues have been raised by our studies; to study the recognition performances on the moving speaker and overlapping speech data, noise robust features and approaches to deal with microphone array data. In the future we also plan to study the effects of tracking beamformer on recognition performances.

6 Acknowledgments

This work was supported by the EU 6th FWP IST integrated project AMI (FP6-506811). IM2 Swiss National Center of Competence in Research. The authors want to thank Artem Peregoudov for providing baseline HMM models, Iain McCowan for helpful discussions and Bastien Crettol for technical support. We also thank the participants in the recordings for their time.

References

[1] S. Das, R. Bakis, A. Nadas, D. Nahamoo and M. Picheny, ”Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system”, Proc. ICASSP,pp. 71-74, 1993.

[2] R. Cardin, Y. Normandin and E. Millie, ”Inter-word coarticulation modeling and MMIE training for improved connected digit recognition”,Proc. ICASSP, pp. 243-246, 1993.

[3] L.R. Rabiner, ”Applications of voice processing to communications”, Proc. IEEE, Vol. 82, No. 2, pp. 199-228, Feb. 1994.

[4] Fischer S. and Simmer K.U., An Adaptive Microphone Array for Hands-Free Communication, Proc. IWAENC95, pp.44-47, June 1995.

[5] Sullivan T., Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition, Ph.D.thesis, August 1996.

[6] B.H. Juang and K.K. Paliwal,”Hidden Markov models with first-order equalization for noisy speech recognition”, IEEE Trans. Signal Processing, Vol. 40, pp. 2136-2143, Sept. 1992.

[7] A. Acero and R.M. Stern, ”Environmental robustness in automatic speech recognition”, Proc. ICASSP pp. 849-952, 1990.

[8] Paul, D.B., Baker, J.M., ”The Design for the Wall Street Journal-based CSR Corpus”, DARPA Speech and Language Workshop, Morgan Kaufmann Publishers, SanMateo, CA, Feb 1992.

[9] D. Moore. ”The IDIAP smart meeting room”. IDIAP Com-02-07, Martigny, Switzerland, Nov. 2002.

[10] D. Moore and I. McCowan. ”Microphone array speech recognition:Experiments on overlapping speech in meetings”. In Proc. ICASSP, Apr. 2003.

[11] H. Cox, R. Zeskind, and M. Owen. ”Robust adaptive beamforming”.IEEE Trans. on Acoustics, Speech and Signal Processing,35(10):13651376, Oct. 1987.

[12] K. U. Simmer, J. Bitzer, and C. Marro. ”Post-filtering techniques”. In M. Brandstein and D. Ward, editors, Microphone Arrays, chapter 3, pages 3660. Springer,2001.

[13] I. McCowan and H. Bourlard. ”Microphone array post-filter based on noise field coherence”. IEEE Trans. on Speech and Audio Processing, 11(6), November 2003.

[14] C. Leggetter and P. Woodland, ”Maximum likelihood linear regression for speaker adaptation of continuous density HMMs,” Computer Speech and Language, vol. 9, pp. 171– 185, 1995.

[15] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK-Hidden Markov Model Toolkit V2.1, Entropic Research, Cambridge, 1997.