IDIAP RESEARCH REPORT

Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech

Hari Krishna Maganti [*]
Daniel Gatica-Perez [**]

IDIAP–RR 06-29

May, 2006

to appear in
International Conference on Multimodal Interfaces 2006

[*] IDIAP Research Institute and University of Ulm, Ulm, Germany
[**] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL)

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 06-29

Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech

Hari Krishna Maganti   Daniel Gatica-Perez

May, 2006

to appear in
International Conference on Multimodal Interfaces 2006

Abstract. Accurate speaker location is essential for optimal performance of distant speech acquisition systems using microphone array techniques. However, to the best of our knowledge, no comprehensive studies on the degradation of automatic speech recognition (ASR) as a function of speaker location accuracy in a multi-party scenario exist. In this paper, we describe a framework for evaluation of the effects of speaker location errors on a microphone array-based ASR system, in the context of meetings in multi-sensor rooms comprising multiple cameras and microphones. Speakers are manually annotated in videos in different camera views, and triangulation is used to determine an accurate speaker location. Errors in the speaker location are then induced in a systematic manner to observe their influence on speech recognition performance. The system is evaluated on real overlapping speech data collected with simultaneous speakers in a meeting room. The results are compared with those obtained from close-talking headset microphones, lapel microphones, and speaker location based on audio-only and audio-visual information approaches.

1 Introduction

The recent aim of present innovations in multimodal technology components is to provide simpler interfaces for human-computer interaction applications in many conversational settings, including meetings. Meetings represent an effective mode of information exchange among humans. Automatic analysis of meetings have a potential of conveying detailed knowledge about human activities which can also be archived for later retrieval and review [4].

In the context of automatic analysis of meetings, robust localization and tracking of active speakers is of fundamental importance, particularly for enhancement and recognition of speech in microphone-array based ASR systems. Microphone arrays provide hands-free and high-quality distant speech acquisition through beamforming techniques, which rely on speaker location for speech enhancement [2]. In this article, we present a framework for systematic evaluation of an integrated system comprising speaker tracking and microphone array speech recognition. This evaluation refer to errors in speaker position coordinates and the corresponding influence on speech recognition performance.

Localization and tracking of active speakers, and speech enhancement and recognition from multiple far-field microphones are challenging tasks in smart room scenarios, where the speech signal is corrupted with noise from presentation devices and room reverberations. These tasks are further complicated in the case of overlapping speech, where multiple speakers talk simultaneously, which is a common situation in multi-party interaction [15]. Localization and tracking of active speakers have been investigated using computer vision systems [3] , audio source localization systems [10] and approaches based on audio-visual fusion [5, 13]. Microphone array speech recognition (i.e, the integration of beamformer with ASR for overlap speech in meeting rooms) has been investigated in [12].

In a closely related work, McCowan et al. presented a system combining audio-visual multi-speaker tracking and microphone array based speech enhancement for overlapping speech, however, speech recognition was not considered in this work [11]. Recently Asano et al. presented a system to detect, enhance, and recognize single speaker speech based on the fusion of sound localization from a small microphone array and vision tracking based on background subtraction from two cameras [1]. This work was limited to detection of speech events in noisy environments, and neither overlapping speech scenarios nor the evaluation of speaker localization errors for ASR were studied. A particle filter fusing audio from multiple large microphone arrays and video from multiple calibrated cameras was used in the context of seminar rooms, which also evaluated the effect of localization errors for ASR [16]. However, in this work neither multiple-person localization nor overlapping speech scenarios were considered. With the growing interest for system integration, it is important to analyze in detail the influence of speaker location on speech recognition performance. In the current paper, a systematic evaluation framework for overlapping speech from two speakers in meetings is presented. These two speaker locations are computed from manually labelled positions in two different camera views and triangulated to determine accurate speaker positions. One speaker position is kept fixed, then position errors systematically induced in the x,y,z coordinates of the other speaker, and speech recognition performance is finally measured. The performance is also compared with those obtained from close-talking headset microphones, close-talking lapel microphones, and speaker locations based on audio-only, and audio-visual based tracking methods.

The paper is organized and presented in four sections: Section 2 gives an overview of the system setup, section 3 explains the evaluation methodology, section 4 presents experiments and results, and finally section 5 concludes the paper.

2 System setup

The observations are based on recordings from an instrumented meeting room. The audio-visual sensors include a circular eight-element microphone array centered on the table, headset and lapel microphones made of high quality electret type, and three CCTV video cameras to capture different views of the participants. The sensor configuration and calibration are similar to the system described by McCowan et al. [11]. Figure 1a shows the room layout, the positions of the microphone array and video cameras, and the typical speaker positions in the meeting room. To observe the microphone array sensitivity and spatial response across the meeting room, the speaker positions are designed to cover all possible seated regions: side-by-side (P1,P3) , opposite (P1,P2) and diagonally opposite (P1,P4). Sample images from the center camera are as shown in Figure 1b.

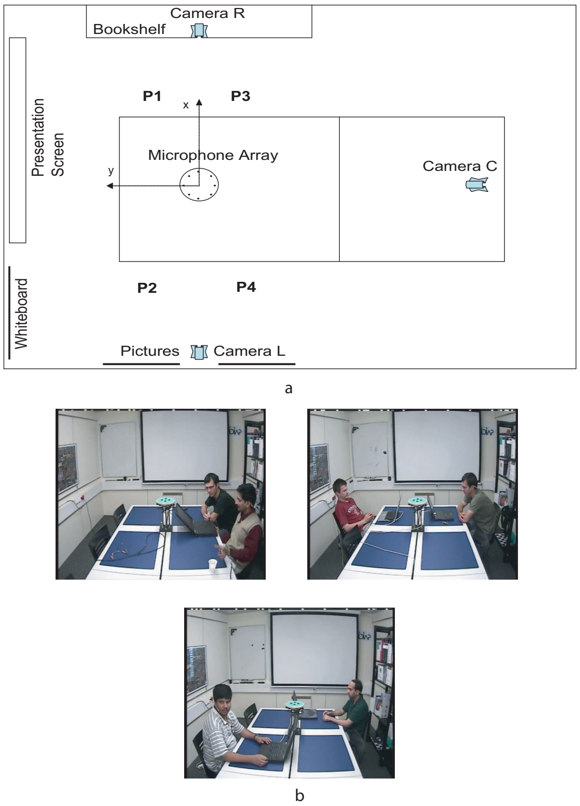

Figure 1: *(a) Schematic diagram of the meeting room; (b) sample images from the center camera. P1, P2, P3, P4 indicate the typical speaker positions.*

3 Evaluation Methodology

A schematic description of the evaluation methodology is shown in Figure 2. Each speaker is captured in two different camera views from the array of video cameras which are calibrated to a microphone-array 3-D reference of the meeting room. The frame-based ground truth was generated as follows. First, the 2-D point-based mouth position of each speaker was manually annotated in each camera plane. Then, each pair of 2-D points was reconstructed into a 3-D point using standard optimization methods [7]. The ground truth was produced at a rate of 1 frame/sec, i.e., every 25 video frames. The 3-D points for each speaker are used as input to the speech enhancement module.

The speech enhancement module comprises a beamformer followed by a post-filter. The beamformer uses a superdirective technique to calculate the channel filters maximizing the array gain, while maintaining a minimum constraint on the white noise gain as described by Cox et al. [2]. The post-filter is specifically designed to handle overlapping speech and is based on the speech signal energy, as fully described in [11]. At each time-step for which the distance between the tracked speaker location and the beamformer's focus location exceeds a small value, the beamformer channel filters are recal-

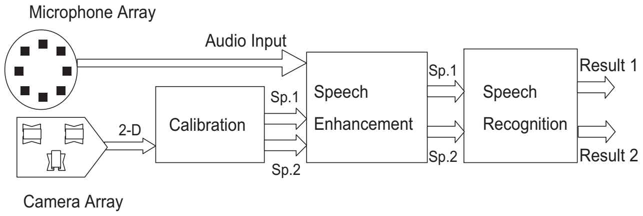

Figure 2: *System block diagram. The audio from the microphone array and 3-D estimates, reconstructed by the calibration module are used as inputs to the speech enhancement module. Sp.1 and Sp.2 indicate speaker 1 and speaker 2 respectively. Speech recognition is performed on the enhanced speech to obtain the corresponding result.*

culated. The enhanced speech is used as input to a standard HMM recognition system to evaluate the quality of the speech signal. For the baseline, a full HTK based recognition system, trained on the original Wall Street Journal database (WSJCAM0) is used [14]. 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 six mixture components per state. Fifty two element feature vectors comprising 13 MFCCs (including the 0th cepstral coefficient) with their first, second, and third order derivatives were used. Cepstral mean normalization is performed on all the channels. The dictionaries used are generated from that developed for the Augmented Multi-party Interaction (AMI) project and used in the evaluations of National Institute of Standards and Technology rich transcriptions (NIST RT05S) system [8], and the language models are the standard MIT-Lincoln Labs 5k and 20k Wall Street Journal trigram language models. To reduce the channel mismatch between the training and test conditions, the baseline HMM models are adapted using a maximum likelihood linear regression (MLLR) [9] and maximum-a-posteriori (MAP) adaptation [6]. 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).

The evaluation protocol is as follows. One speaker is kept fixed for all the cases and errors are induced in the x,y,z coordinates of the other speakers in a controlled fashion. The evaluation refers to errors in the speaker position coordinates in all the three axes of a cartesian coordinate system, and the corresponding influence on speech recognition performance. So for a range of R cm, each of the x,y,z coordinates are changed by time-steps of r cm and then speech recognition performance is measured.

Finally, to examine the approximate range of speaker location errors and their influence on speech recognition performance, audio-only localization and audio-visual based tracking systems are evaluated. The location estimates are directly computed from the audio-only speaker localization as proposed by Lathoud et al. [10] and the audio-visual tracking as proposed by Gatica-Perez et al. [5].

4 Experiments and results

For the experiments, an overlapping speech corpus was recorded which consisted of read Wall Street Journal sentences taken from the test set of the WSJCAM0 database. The sentences were read by two speakers simultaneously in a meeting room. The data comprised non-native English speakers with different speaking styles and accents. The audio data from the headset microphones, lapel microphones, and eight element circular microphone array was captured. The meeting room also provided synchronized video recordings, including frontal views of the participants and wide-angle view of the entire room. The data is divided into development (DEV) and evaluation (EVAL) sets with no common speakers in both sets. The DEV set consisted of 60 sentences, amounting to 11

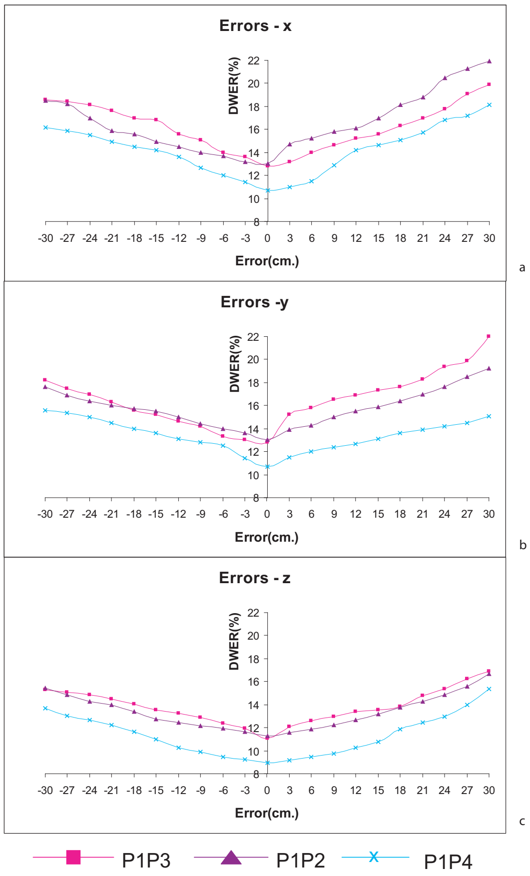

Figure 3: *Plots showing errors vs. degradation in WER for the three cases in* x, y, z *coordinates.*

minutes of speech data. The EVAL set comprised 150 sentences from two speakers for all the cases (side-by-side, opposite, and diagonally opposite speakers). The speaker at P1 position is fixed and the errors are induced in the x,y,z coordinates of the other speakers. The positive x axis is towards the bookshelf and the positive y axis towards the presentation screen, as shown in the Figure 1a, while the positive z axis is towards the roof of the meeting room. The errors are introduced in a step of r = 3 cm for a range of R = 30 cm in both positive and negative directions for all the coordinates. Speech recognition experiments are performed only for the speakers where errors have been introduced, hence the total number of sentences considered are 75, amounting to 14 minutes of data.

Figure 3 shows the results for errors in cm, in x, y, z coordinates against word error rates (WER) in percentage for all the cases. The y-axis is the absolute degradation in WER (DWER) with respect to the headset microphone, WER(array) - WER(headset)), which are the best results and are considered as baseline for all the results, as shown in Table 1. The values at zero on the error axis indicate the degradation in WER obtained from the ground truth.

It can be observed from the plots that WERs are roughly linear with respect to the localization errors, indicating that the errors in computing the speaker location proportionally degrade the quality of the desired speech signal. From the plots, it is clear that the WERs for P1P3 (side-by-side) speakers are normally the highest in all the three coordinates due to the close proximity of the speakers. As expected, in the case of P1P4 (diagonally opposite) speakers, the WERs are low in all the three coordinates due to the comparatively large distance between the speakers. For the errors in the positive x direction, P1P2 case is the worst performing because positive x is the direction that brings the two speakers closer. Also, for the errors in positive y direction, P1P3 case is the worst because positive y is the direction that brings the two speakers close. These are clearly evident from Figures 3a and 3b.

From Figure 3, it is also observed that the WERs are lower in the case of z coordinate than the x y coordinates, and the errors in positive and negative z coordinate are symmetrical. This indicates that the quality of the speech signal is comparatively less affected by the errors in the z coordinate and also that the inaccuracies in the positive and negative directions distort the speech signal similarly.

Table 1 also shows the comparison of WERs for ground truth, lapel microphone, and speaker locations based on recently proposed by Gatica-Perez et al. [5] and Lathoud et al. [10] for all the two-speaker methods.

Table 1: *Comparison of WER for different channels. Headset values are the absolute values and the remaining are relative to the headset.*

	WER (%)		
Signal	**P1P3**	**P1P2**	**P1P4**
Headset	**41.6**	**43.2**	**44.7**
Ground truth	12.8	13.0	10.7
Lapel	13.4	13.8	11.4
Audio-visual	17.2	16.2	16.6
Audio-only	34.6	36.0	36.5

As expected, in the case of P1P4 , the ground truth and lapel microphone have less WER degradation than P1P3 and P1P2 which is again due to the comparatively long distance between the speakers. It is also clear, that the lapel and the ground truth based microphone-array have relatively similar performances indicating the accuracy of the audio-visual calibration method. It is also observed that for all the cases, the WER for audio-only estimates are much higher than audio-visual tracking estimates, which are higher in turn than the ground truth speaker locations. For the estimates from the audio-visual tracking method, the Euclidean distance error is between 18-24 cm, which is partly due to the fact that the tracker estimate in each camera view corresponds to the center of a person's head, rather than to the center of the mouth, and that the two head centers in each camera view do not correspond to the 3-D same physical point. In contrast, the audio-only estimates are discontinuous and are available only in approximately 60% of the frames. Errors are computed only on those frames for which there is at least one audio estimate, and the Euclidean distance error is between 100-120 cm. This is clearly reflected on the WER as observed in Table 1. This confirms that the speaker location estimates based on audio-visual fusion information are more accurate than the audio-only estimates, which is consistent with earlier studies [16]. In summary, our experiments suggest what the limits of ASR are even when localization is perfect, and where some current technologies stand with respect to this given standard.

5 Conclusions

We presented a framework to evaluate the effect of speaker location errors on a microphone array-based ASR system in the context of meetings. The system is evaluated on real data collected with overlapping speech from simultaneous speakers in a meeting room. The evaluation refers to the errors in speaker position coordinates in the three axes of a cartesian coordinate system, and the corresponding influence on speech recognition performance. The results indicate that in a range of 30 cm range, the errors in computing the speaker location degrade the quality of the desired speech signal in a roughly propotional way. We also observed that the quality of the speech signal is less affected by errors in the z coordinate, and that the inaccuracies in positive and negative directions distort the speech signal similarly. We also compared the results with those obtained with lapel microphones and speaker locations based on audio-only speaker localization and audio-visual tracking methods and confirmed that the errors in speaker location estimates by audio-visual tracking are promising but still open to improvement.

Acknowledgment

This work is supported by the EC projects AMI (Augmented Multi-party Interaction, pub. AMI-191), DIRAC (Detection and Identifications of Rare Audio-visual Cues), and the Swiss NCCR IM2. We thank M. Lincoln and I. McCowan for the collaboration in designing the MC-WSJ-AV corpus, S. Ba for his help with the audio-visual sensor array calibration, and B. Crettol for his support to collect the data.

References

[1] F. Asano, Y. Motomura, H. Asoh, T. Yoshimura, N. Ichimura, K. Yamamoto, N. Kitawaki, and S. Nakamura. "Detection and Separation of Speech Event using audio and video information fusion," *Journal of Applied Signal Processing*, vol.11, pp. 1727–1738, 2004.

[2] H. Cox, R. Zeskind, and M. Owen. "Robust adaptive beamforming," *IEEE Trans. on Acoustics, Speech and Sig. Processing*, 35(10):1365–1376, Oct., 1987.

[3] J. Crowley and P. Berard. "Multi-modal tracking of faces for video communications," *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Juan, June, 1997.

[4] R. Cutler, Y.Rui, A. Gupta, JJ Cadiz, I.Tashew, L.He, A.Colburn, Z.Zhang, Z.Liu, and S.Silverberg. "Distributed meetings: A meeting capture and broadcasting system". *Proc. ACM MM*, Oct, 2002.

[5] D. Gatica-Perez, G. Lathoud, , J. -M. Odobez, and I. McCowan. "Multimodal Multispeaker Probabilistic Tracking in Meetings," *Proc. of the IEEE Conf. on Multimedia Interfaces (ICMI)*, Trento, Oct., 2005.

[6] J.-L. Gauvain and C.-H. Lee. "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 2(2):291–298, April, 1994.

[7] R. Hartley and A. Zisserman. "Multiple View Geometry in Computer Vision," *Cambridge University Press*, second edition, 2001.

[8] T. Hain et al. "The Development of the AMI System for the Transcription of Speech in Meetings," *Proc. of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, July, 2005.

[9] C. J. Leggetter and P. C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9(2):171-185, 1995.

[10] G. Lathoud and I. McCowan. "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays," *Proc. of the ISCA Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Oct., 2004.

[11] I. McCowan, H.K. Maganti, D. Gatica-Perez, D. Moore, and S. Ba. "Speech Acquisition in Meetings with an Audio-Visual Sensor Array," *Proc. of the IEEE Conf. on Multimedia Expo (ICME)*, Amsterdam, July, 2005.

[12] D. Moore and I. McCowan. "Microphone array speech recognition: Experiments on overlapping speech in meetings," *Proc. of Int.Conf. on Acoustics, Speech, and Sig. Processing (ICASSP)*, HongKong, Apr., 2003.

[13] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. "A joint particle filter for audio-visual speaker tracking," *Proc. of the IEEE Conf. on Multimedia Interfaces (ICMI)*, Trento, Oct., 2005.

[14] T.Robinson, J. Fransen, D.Pye, J.Foote, and S. Renals. "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," *Proc. of Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, April, 1995.

[15] E. Shriberg, A. Stolcke, and D. Baron. "Observations on overlap: findings and implications for automatic processing of multi-party conversation," *Proc. of the 7th Eurospeech Conf. on Speech Communication and Technology (Eurospeech-2001)* , Aalbourg, Sep., 2001.

[16] M. Wolfel, K. Nickel, and J. McDonough. "Microphone Array Driven Speech Recognition: Influence of Localization on the Word Error Rate," *Proc. of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, July, 2005.