IDIAP RESEARCH REPORT

Speech Acquisition in Meetings with an Audio-Visual Sensor Array

Iain McCowan *   Maganti Hari Krishna *
Daniel Gatica-Perez *   Darren Moore *
Sileye Ba *

IDIAP–RR 05-03

January, 2005

to appear in
IEEE International Conference on Multimedia and Expo 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 05-03

Speech Acquisition in Meetings with an Audio-Visual Sensor Array

Iain McCowan Maganti Hari Krishna Daniel Gatica-Perez
Darren Moore Sileye Ba

January, 2005

to appear in
IEEE International Conference on Multimedia and Expo 2005

Abstract. Close-talk headset microphones have been traditionally used for speech acquisition in a number of applications, as they naturally provide a higher signal-to-noise ratio -needed for recognition tasks- than single distant microphones. However, in multi-party conversational settings like meetings, microphone arrays represent an important alternative to close-talking microphones, as they allow for localisation and tracking of speakers and signal-independent enhancement, while providing a non-intrusive, hands-free operation mode. In this article, we investigate the joint use of a small table-top microphone array and a camera array for speaker tracking and speech enhancement in meetings. Our methodology first fuses audio and video for person tracking, and then integrates the output of the tracker with a beamformer for speech enhancement. We compare and discuss the features of the resulting speech signal with respect to that obtained from single close-talking and table-top microphones.

1 Introduction

A significant trend in computing research is towards pervasive computing, where the computer becomes an integral part of the user's environment; observing, analysing and influencing behaviour through an array of multimodal sensors. Applications include instrumented meeting rooms or lecture halls, where the goal is to enhance (co-located or remote) collaboration, both in real-time and through recorded multimedia archives.

In this article, we investigate the use of an audio-visual sensor array for speech acquisition in a meeting room. Audio is captured using a circular, table-top array of 8 microphones, and visual information is captured from 3 different camera views. Both audio and visual information are first used to find and track the location of each speaker in the meeting room. Microphone array beamforming techniques are then applied, providing hands-free (untethered) speech acquisition from each tracked location.

Multiple microphones and video cameras have been recently used for tracking speakers in video-conferencing [1] and meeting analysis [2]. These works did not study the use of tracking for speech enhancement. The work closest to ours is perhaps [3], where a microphone array and two cameras were used to enhance and recognize speech. The specific algorithms used for localization, tracking, and enhancement are however substantially different to ours.

While some stationary background noise exists in meeting rooms due to devices such as laptops and data projectors, the most significant source of 'noise' (with respect to a given person's speech) is the concurrent occurrence of speech from other people. In [4] it was identified that around 10-15% of words, or 50% of speech segments, in a meeting contain a degree of overlapping speech. These overlapped speech segments are problematic for speaker segmentation, and speech and speaker recognition. In previous work to recognise overlapping speech in meetings [5], we used a superdirective beamformer [6] followed by a post-filtering stage. In that case, the post-filter was based on one which assumed the predominant noise was diffuse [7]. In the current paper we instead propose a new post-filter that more effectively removes overlapping speech.

The paper is organised as follows. Section 2 presents the sensor array configuration and discusses inter-modality calibration issues. Section 3 details our technique for audio-visual speaker tracking, while our microphone array speech enhancement approach is described in Section 4. Section 5 presents and discusses experimental results of the integrated tracking and speech enhancement system. Conclusions and future work plans are given in Section 6.

2 Audio-visual sensor array

2.1 Sensor configuration

The sensor array is deployed in a 8.2m×3.6m×2.4m meeting room containing a 4.8m×1.2m rectangular table [8]. The audio sensors are configured as an eight-element, circular equi-spaced microphone array centered on the table, with diameter 20cm, and composed of high quality miniature electret microphones. Additionally, lapel and headset microphones are used for each speaker. The video sensors include seven CCTV cameras . 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 more cameras are co-located with the microphone array on the table for close views of meeting participants, but are not used for the reported experiments. Sample images from the room and the sensor array can be seen in Figure 1. All sensors are connected to fully synchronized capture devices.

2.2 Sensor calibration

To relate points in the 3-D camera reference with 2-D image points, we calibrate the three cameras of our meeting room to a single 3-D external reference, using a standard camera calibration procedure [9] with the software available in [10]. The method estimates the different camera parameters with a given number of image planes. The microphone array has its own external reference, so in order to map a 3-D point in the microphone array reference to an image point, we also need to define a transform for basis change. Finally, to complete the audio-video mapping we need to find the correspondence between image points and 3-D microphone array points. From stereovision, the 3-D reconstruction of a point can be done with the image coordinates of the same point in two different camera views. Each point in each camera view defines a ray in the 3-D space. Optimization methods can be used to find the intersection of the two rays, which correspond to the reconstructed 3-D point [11]. This last step is used to map the results of the tracking algorithm, (i.e. the people location in the image planes) back to 3-D points, as input to the beamformer.

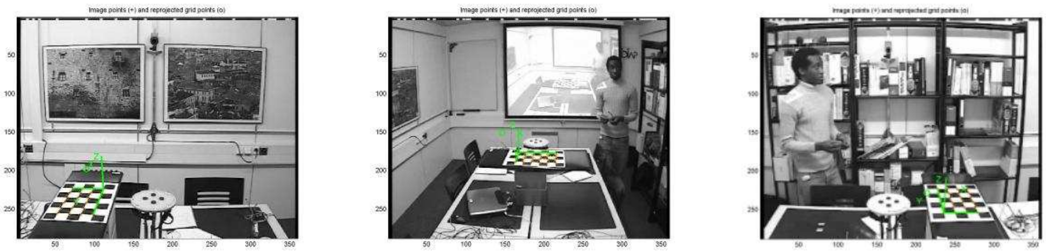

Figure 1: Camera 3-D external reference

3 Audio-visual Person Tracking

We address the problem of tracking multiple people as one of approximate inference in a graphical model, using sequential Monte Carlo (SMC) methods [12]. We use a multi-object state space formulation, which in addition to being mathematically rigorous, allows for the explicit definition of object interaction models. For multi-object configurations $X_t = (X_{1,t}, ..., X_{N_O,t})$, and audio-visual observations Y_t, the filtering distribution $p(X_t|Y_{1:t})$ is recursively approximated by a weighted set of samples or particles $\{X_t^{(n)}, w_t^{(n)}\}_{n=1}^{N_s}$, and updated as new observations become available, via importance sampling. Given a multi-object dynamical model $p(X_t|X_{t-1})$, a multi-object observation likelihood $p(Y_t|X_t)$, and the particle set at the previous time step, a set of candidate configurations at the current time step are drawn from a proposal distribution $q(X_t) = \sum_n w_{t-1}^{(n)} p(X_t|X_{t-1}^{(n)})$. The weights are then computed as $w_t^{(n)} \propto p(Y_t|X_t^{(n)})$. A person is represented by the silhouette of her head in the image plane. A mixed state-space is defined over joint multi-object configurations, where in addition to a set of continuous variables modeling head motion, discrete variables are included to model each participant's speaking status.

Observation models are derived from audio and video. Audio observations are derived from an speaker localization algorithm as follows. First, a sector-based source localization algorithm is used to generate candidate 3-D locations of people when they speak [13]. Given the higher sampling rate for audio, multiple audio localization estimates are merged for each video frame. We then use the sensor calibration procedure in the previous section to project the 3-D audio estimates on the corresponding 2-D image planes. Finally, the audio observation likelihood relates the distance between the audio localization estimates and the candidate configurations on the image plane. Visual observations are based on shape and spatial structure of human heads. The shape observation model is derived from edge features computed over a number of perpendicular lines to a proposed head configuration. The

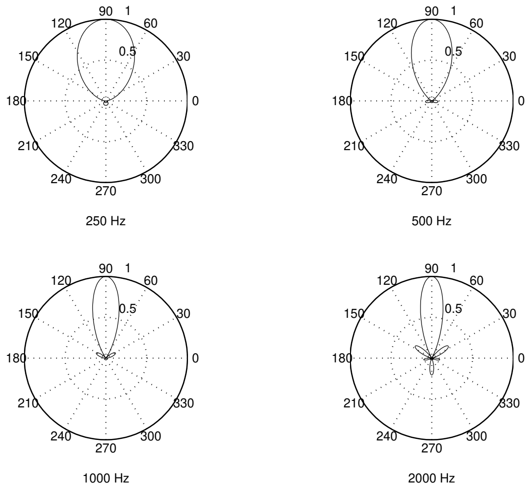

Figure 2: Horizontal polar plot of the directivity pattern of the superdirective beamformer for an 8-element circular array of radius 10cm

spatial structure observations are derived from skin-color blob features.

Inference with a simple particle filter on the high-dimensional space defined by several objects being tracked is computationally infeasible. We have dealt with this issue in [12] by using Markov Chain Monte Carlo (MCMC) sampling, which efficiently places samples as close as possible to regions of high likelihood.

For speech acquisition, the multi-object person tracker is run in each video stream independently, outputting the 2-D location of each person's head center for each view. The video-audio mapping described in Section 2 is used to reconstruct 3-D locations from two 2-D points. Such 3-D points are used as input to the beamformer, as described in the following section.

4 Speech Enhancement in Meetings

Similar to the system presented in [5], the microphone array speech enhancement system includes a filter-sum beamformer followed by a post-filtering stage.

4.1 Beamformer

For the beamformer, we use the superdirective technique to calculate the channel filters maximising the array gain, while maintaining a minimum constraint on the white noise gain. This technique is fully described in [6].

Figure 2 shows the polar directivity pattern of this superdirective beamformer at several frequencies for the array used in our experiments. We see that this geometry gives reasonable discrimination between speakers separated by at least 45^o, making it suitable for small group meetings of up to 8 participants (assuming a relatively uniform angular distribution of participants).

For the experiments in this paper we integrated the tracker output with the beamformer in a straightforward manner. Any time the distance between the tracked speaker location and the beamformer's focus location exceeded 5cm, the beamfomer channel filters were recalculated. Steering errors less than this give a negligible degradation in signal gain.

4.2 Post-filter for Overlapping Speech

The use of a post-filter following the beamformer has been shown to improve the broadband noise reduction of the array [14], and lead to better performance in speech recognition applications [5]. Much of this previous work has been based on the use of the (time-delayed) microphone auto- and cross- spectral densities to estimate a Wiener transfer function. While this approach has shown good performance in a number of applications, its formulation is based on the assumption of low correlation between the noise on different microphones. This assumption clearly does not hold when the predominant 'noise' source is coherent, such as overlapping speech. In the following we propose a new post-filter better suited for this case.

Assume that we have S beamformers concurrently tracking S different people within a room, with (frequency-domain) outputs b_s, $s = 1 : S$. We further assume that in each b_s, the energy of speech from person s (when active) is higher than the energy level of all other people. It has been observed (see [15] for a discussion) that the spectrum of the additive combination of two speech signals can be well approximated by taking the maximum of the two individual spectra in each frequency bin, at each time. This is essentially due to the sparse and varying nature of speech energy across frequency and time, which makes it highly unlikely that two concurrent speech signals will carry significant energy in the same frequency bin at the same time. This property was exploited in [15] to develop a single-channel speaker separation system.

We apply the above property over the S frequency-domain beamformer outputs to calculate S simple masking post-filters, $h_s(f)$, $s = 1 : S$,

$$h_s = \begin{cases} 1 & \text{if } s = \underset{s'}{\arg\max} \; |b_{s'}|^2, \; s' = 1 : S \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

Each post-filter is then applied to the corresponding beamformer output to give the final enhanced output for person s as $z_s = h_s b_s$. We note that when only one person is actively speaking, the other beamformers will essentially be providing an estimate of the background noise level, and so the post-filter should also function to reduce background noise. This post-filter also has the benefit of low computational cost compared to other formulations which require the calculation of channel auto- and cross-spectral densities.

5 Experiments and Results

5.1 Data collection

A corpus consisting of several single- and two-person sequences was recorded in the multi-sensor room. People were asked to read a number of sentences from the Wall Street Journal (WSJ) speech corpus, displayed both on the presentation screen and on a laptop placed on the opposite extreme of the meeting table. In the single-person sequences, the person occupies a number of typical positions in the room, moving naturally across locations. In the two-person sequences, people remain seated, each reading a different sentence simultaneously. In some sequences, people were asked to deliberately look away from the microphone array while speaking. Participants were predominately non-native english speakers. In all sequences, participants wore headset and lapel microphones. For the experiments reported here, we use three sequences (two single-person (*seq-1s-A*, *seq-1s-B*), one two-person (*seq-2s-A*) with duration 80, 124, and 36 seconds, respectively.

5.2 Tracking

For all sequences described earlier in this section, the audio-visual tracker was initialized by hand in the first frame of each view. Regarding head location, all objects were sucessfully tracked, and with a good accuracy. Sample frames are shown in Figure 3. Regarding speaker activity, a double

	SNRE (dB)			
	Single-speaker			Two-speaker
Signal	***at***	***away***	***all***	***crosstalk***
headset	20.36	22.76	22.05	16.34
lapel	15.44	14.60	15.28	10.63
beamformer	6.47	6.32	6.43	3.84
post-filter	20.59	19.96	20.63	11.27

Table 1: SNRE (signal-to-noise ratio enhancement) results for single-speaker and two-speaker sequences. All results are in dB, and are calculated with respect to the level on a single table-top microphone.

ellipse is drawn over those people for which the model has inferred speaking activity. The 3-D audio localization estimates mapped onto each image plane are denoted by green ×. In general, when people talk naturally (e.g. addressing the other speakers at the meeting table), our algorithms detect and infer the source location and speaking status reasonably well. When people clearly face away from the array, the audio estimates degrade, and so does the inference of the speaking status. For the two-person sequence, the algorithm infers correctly those segments where only one speaker takes the turn, and often infers simultaneous activity during overlapped speech periods. However, one can still observe a "dominant speaker" effect. This issue requires further research.

5.3 Speech Enhancement

To evaluate the effectiveness of the microphone array in acquiring a clean speech signal of each person, we calculated the average segmental signal to noise ratio (SNR) for the following cases:

at: When the speaker is looking at the microphone array.
away: When the speaker is looking away from the array.
all: The entire sequence (weighted mean of *at* and *away*).
crosstalk: The SNR is calculated taking frames with only the desired speaker as signal, and frames with only the other speaker as noise.

The first three of these measures were calculated over both the single-speaker sequences (*seq-1s-A*, *seq-1s-B*), and the *crosstalk* measure was only calculated on the two-speaker sequence (*seq-2s-A*). To normalise for different levels of individual speakers, all results are quoted with respect to the input on a single table-top microphone, and so in fact represent the SNR enhancement (SNRE). These results are shown in Table 1.

These results show a number of encouraging trends. First, by comparing the *at* and *away* measures, we see that the microphone array enhancement techniques, and implicitly the tracker, are robust when the speaker is not looking directly at the array. Any level difference can be explained by the differences in input speech level over those segments, as seen in the differences between 'headset' *at* and *away*. A second observation is that the final output of the proposed tracking microphone array technique ('post-filter') is comparable to the headset microphone in reducing background noise level. While the SNRE levels for 'beamformer' are lower than those for the lapel, it still provides 6.4 dB of improvement over a single table-top microphone, which is a good result for an 8-element array. It is clear that the proposed post-filter for reducing crosstalk speech is effective in significantly improving on the beamformer output. From the *crosstalk* results, we see that the proposed tracking microphone array gives a similar reduction of speech crosstalk as a lapel microphone. It should be noted that this 2-speaker sequence represents the worst-case scenario for speech separation, as the 2-speakers are sitting next to each other, and are thus within the main-lobe width of each other's beamformer at low frequencies (reflected in the low *crosstalk* 'beamformer' SNRE). Ongoing experiments are investigating the speech enhancement performance across a wider range of crosstalk scenarios.

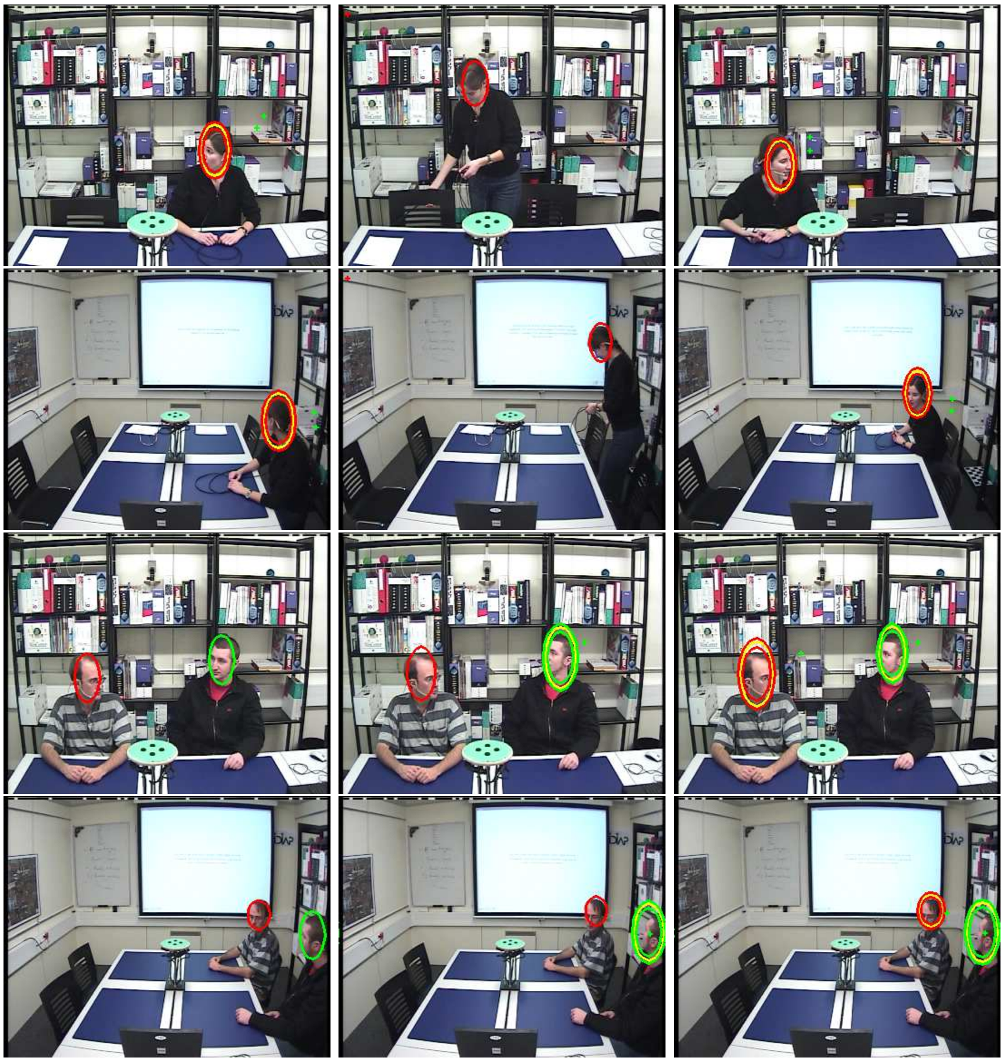

Figure 3: Audio-visual tracking results. First two rows: one-person sequence. Each column corresponds to frames 89, 1260, and 1535, respectively. Last two rows: two-person sequence. Each column corresponds to frames 1, 76, and 136, respectively.

Finally, we note that the above results do not give a complete picture of performance. As well as assessing noise level reduction, it is also necessary to quantify any distortion to the desired speech signal. Ongoing work will assess this by using the tracking microphone array as input to a speech recognition system.

6 Conclusion

We presented a framework to acquire speech in meetings, using information captured by an audio-visual sensor array. Two encouraging results were obtained. The first one confirms that the use of video for tracking helps providing a stable direction for beamforming, more accurate than the one produced by an audio-only source localization method. The second result suggests that the use of the proposed beamformer post-filter enhances speech quality to a degree comparable to that of headset mics, and better than lapels. Our study also raised a number of issues, including the improvement of the model of simultaneous speech (both in localization and tracking), and the need to study the specific effects of each functional block in the final outcome. These subjects will be studied in the future.

7 Acknowledgments

We thank Guillaume Lathoud and Jean-Marc Odobez for discussions and advice, Sebastien Bourban for technical support, and all the participants in the recordings for their time.

References

[1] B. Kapralos, M. Jenkin, and E. Milios. Audio-visual localization of multiple speakers in a video teleconferencing setting. *Int. J. Imaging Systems and Technology*, 13:95–105, 2003.

[2] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proc. of the IEEE*, 92(3):485–494, Mar. 2004.

[3] F. et al Asano. Detection and separation of speech event using audio and video information fusion. *Journal of Applied Signal Processing*, 11:1727–1738, 2004.

[4] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech 2001*, volume 2, pages 1359–1362, 2001.

[5] D. Moore and I. McCowan. Microphone array speech recognition: Experiments on overlapping speech in meetings. In *Proceedings of ICASSP*, April 2003.

[6] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, October 1987.

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

[8] D. Moore. The IDIAP smart meeting room. *IDIAP Communication 02-07*, 2002.

[9] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. ICCV*, Kerkyra, Sep. 1999.

[10] J.-Y. Bouguet. Camera Calibration Toolbox for Matlab, open-source code available at `http://www.vision.caltech.edu/bouguetj/`.

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

[12] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audio-visual tracking of multiple speakers in meetings. IDIAP-RR-04-66, Martigny, Switzerland, Dec. 2004.

[13] G. Lathoud and I. McCowan. A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays. In *Proceedings of the Workshop on Statistical and Perceptual Audio Processing SAPA'04*, October 2004.

[14] K. Uwe Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 3, pages 36–60. Springer, 2001.

[15] S. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proceedings of Eurospeech 2003*, September 2003.