IDIAP RESEARCH REPORT

Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array

Hari Krishna Maganti * Daniel Gatica-Perez *
Iain McCowan **

IDIAP–RR 06-24

April, 2006

* IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL)
** eHealth Research Centre, Brisbane, Australia

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 06-24

Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array

Hari Krishna Maganti   Daniel Gatica-Perez   Iain McCowan

April, 2006

Abstract. We address the problem of distant speech acquisition in multi-party meetings, using multiple microphones and cameras. Microphone array beamforming techniques present a potential alternative to close-talking microphones by providing speech enhancement through spatial filtering and directional discrimination. Beamforming techniques rely on the knowledge of a speaker location. In this paper, we present an integrated approach, in which an audio-visual multi-person tracker is used to track active speakers with high accuracy. Speech enhancement is then achieved using microphone array beamforming followed by a novel post-filtering stage. Finally, speech recognition is performed to evaluate the quality of the enhanced speech signal. The approach is evaluated on the data recorded in a real meeting room for stationary speaker, moving speaker and overlapping speech scenarios. The results show that the speech enhancement and recognition performance, achieved using our approach are significantly better than single table-top microphone and comparable to lapel microphone for all the scenarios. The results also indicate that the audio-visual based system performs significantly better than audio-only system, both in terms of enhancement and recognition. This reveals that the accurate speaker tracking, provided by the audio-visual sensor array proved beneficial to improve the recognition performance in a microphone array based speech recognition system.

1 Introduction

With the advent of ubiquitous computing, a significant trend in human-computer interaction is the use of a range of multimodal sensors and processing technologies to observe the user's environment. These allow users to communicate and interact naturally, both with computers or with other users. Example applications include advanced computing environments [1], instrumented meeting rooms [30, 40] and seminar halls [7] facilitating remote collaboration. The current article examines the use of multimodal sensor arrays in the context of instrumented meeting rooms. Meetings consist of natural, complex interaction between multiple participants, and so automatic analysis of meetings is a rich research area, which has been studied actively as a motivating application for a range of multi-disciplinary research [30, 40].

Speech is the predominant communication mode in meetings. Speech acquisition, processing, and recognition in meetings are complex tasks, due to the non-ideal acoustic conditions (e.g. reverberation, noise from presentation devices and computers usually present in meeting rooms) as well as the unconstrained nature of group conversation in which speakers often move around and talk concurrently. A key goal of speech processing and recognition systems in meetings is the acquisition of high-quality speech without constraining users with tethered or close-talking microphones. Microphone arrays provide a means of achieving this through the use of beamforming techniques [20], [33].

A key component of any practical microphone array speech acquisition system is the robust localization and tracking of speakers. Tracking speakers solely based on audio is a difficult task due to a number of factors: human speech is an intermittent signal, speech contains significant energy in the low-frequency range, where spatial discrimination is imprecise, and location estimates are adversely affected by noise and room reverberations. For these reasons, a body of recent work has investigated an audio-visual approach to speaker tracking in conversational settings such as video-conferences [19] and meetings [6]. To date, speaker tracking research has been largely decoupled from microphone array speech recognition research. With the increasing maturity of approaches, it is timely to properly investigate the combination of tracking and recognition systems in real environments, and to validate the potential advantages that the use of multimodal sensors can bring for the enhancement and recognition tasks.

The present work investigates an integrated system for hands-free speech recognition in meetings based on an audio-visual sensor array, including a multimodal approach for multi-person tracking, and speech enhancement and recognition modules. 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 used to track the location of all active speakers in the meeting room. Speech enhancement is then achieved using microphone array beamforming followed by a novel post-filtering stage. The enhanced speech is finally input into a standard HMM recognizer system to evaluate the quality of the speech signal. Experiments consider three scenarios common in real meetings: a single seated active speaker, a moving active speaker, and overlapping speech from concurrent speakers. The speech recognition performance achieved using our approach is compared to that achieved using headset microphones, lapel microphones, and a single table-top microphone. To quantify the advantages of a multimodal approach to tracking, results are also presented using a comparable audio-only system. The results show that the audio-visual tracking based microphone array speech enhancement and recognition system performs significantly better than single table-top microphone and comparable to lapel microphone for all the scenarios. The results also indicate that the audio-visual based system performs significantly better than audio-only system in terms of signal-to-noise ratio enhancement (SNRE) and word error rate (WER). This demonstrates that the accurate speaker tracking, provided by the audio-visual sensor array proved beneficial to improve speech enhancement, which resulted in enhanced speech recognition performance.

The paper is organized as follows: Section 2 discusses the related work. Section 3 gives an overview of the proposed approach. Section 4 describes the sensor array configuration and inter-modality calibration issues. Section 5 details the audio-visual person tracking technique. Section 6 presents the speech enhancement module, while speech recognition is described in Section 7. Section 8 presents the the data, the experiments and their discussion, and finally conclusions are given in Section 9.

2 Related work

Most state-of-the-art speech processing systems rely on close-talking microphones for speech acquisition, as they naturally provide best performance. However, in multiparty conversational settings like meetings this mode of acquisition is not suitable, as they are intrusive and constrain natural behavior of a speaker. For such scenarios, microphone arrays present a potential solution by offering distant, hands-free, and high quality speech acquisition through beamforming techniques [38].

Beamforming consists of filtering and discriminating active speech sources from various noise sources based on location. The simplest beamforming technique is delay-sum, in which a delay filter is applied to each microphone channel before summing them to give a single enhanced output [8]. The extension of delay-sum is superdirective beamforming, where filters are calculated to maximize the array gain for the particular direction [9]. The post filtering of beamformer output significantly improves desired signal enhancement by reducing background noise [27]. Microphone array speech recognition, i.e, the integration of beamformer with automatic speech recognition for meeting rooms has been investigated in [31], [36]. In the same context, at National Institute of Standards and Technology (NIST) meeting recognition evaluations, recent techniques were evaluated which include speech recognition based on multiple distant microphones where the system can handle varying number of microphones, unknown microphone placements, and unknown number of speakers [32].

The localization and tracking of multiple active speakers are crucial for optimal performance of microphone-array based speech acquisition systems. Many computer vision systems [10], [5] have been studied to detect and track people, but are affected by occlusion and illumination effects. Acoustic source localization algorithms can operate in different lighting conditions and localize in spite of visual occlusions. Most acoustic source localization algorithms are based on time-difference of arrival (TDOA) approach, which estimate the time delay of sound signals between the microphones in an array. The generalized cross-correlation phase transform (GCC-PHAT) method [21] is based on estimating the maximum GCC between the delayed signals and is robust to reverberations. The steered response power (SRP) method [22] is based on summing the delayed signals to estimate the power of output signal and is robust to background noise. The advantages of both the methods i.e, robustness to reverberations and background noise are combined in the SRP-PHAT method [11]. To enhance the accuracy of TDOA estimates and handle multi-speaker cases, Kalman filter smoothing was studied in [37] and combination of TDOA with particle filter approach has been investigated in [41]. However, due to the discreteness and vulnerability to noise sources and strong room reverberations, tracking based exclusively on audio estimates is an arduous task. To account for these limitations, multimodal approaches combining acoustic and visual processing have been pursued recently for single-speaker [2, 3, 13, 14, 39, 45] and multi-speaker [4, 6, 19] tracking. The goal of fusion is to make use of complementary advantages: initialization and recovery from failures can be addressed with audio and precise object localization with visual processing.

Being major research topics, speaker tracking and microphone array speech recognition have recently reached levels of performance where they can start being integrated and deployed in real environments. Recently Asano et al. presented a framework where a Bayesian network is used to detect speech events by the fusion of sound localization from a small microphone array and vision tracking based on background subtraction from two cameras [2]. The detected speech event information was used to vary beamformer filters for enhancement and also to separate desired speech segments from noise in the enhanced speech which was then used as input to the recognizer. In other recent work, particle filter data fusion with audio from multiple large microphone arrays, and video from multiple calibrated cameras was used in the context of seminar rooms [28]. The audio features were based on time delay of arrival estimation. For the video features, dynamic foreground segmentation based on adaptive background modelling as primary feature along with foreground detectors were used. The system assumes that the lecturer is the person standing and moving while the audience are sitting and moving less and that there is essentially one main speaker (the lecturer). As we describe in the remainder of the paper, our work substantially differs from previous works in the specific algorithms used for localization, tracking and speech enhancement. Our work is focussed on robust speech acquisition in meetings, and specifically has two advantages over [2] and [28]. First our tracking module can track multiple speakers irrespective of the state of the speakers, e.g. seated, standing, fixed or moving. Secondly, in the enhancement module, the beamformer is followed by a post-filter which helps in broadband noise reduction of the array, leading to better performance in speech recognition. Finally, our sensor setup aims at dealing with small group discussions and relies on a small microphone array, unlike [28] which relies on large arrays.

3 Overview of our approach

A schematic description of our approach is shown in Figure 1. The goal of the blocks on the bottom left part of the figure (Audio Localization, Calibration, and Audio-Visual Tracker) is to accurately estimate, at each time-step, the 3-D locations of each of the people present in a meeting, $\hat{Z}_t = (\hat{Z}_{i,t})$, $i \in \mathcal{I}_t$, where $\mathcal{I}_t$ is the set of person identifiers, $\hat{Z}_{i,t}$ denotes the location for person i, $m_t = |\mathcal{I}_t|$ denotes the number of people in the scene.The estimation of location is done with a multimodal approach, where the information captured by the audio-visual sensors is processed to exploit the complementarity of the two modalities. Human speech is discontinuous in nature. This represents a fundamental challenge for tracking location based solely on audio, as silence periods imply, in practice, lack of observations: people might silently change their location in a meeting (e.g. moving from a seat to the whiteboard) without providing any audio cues that allow for either tracking in the silent periods or re-identification. In contrast, video information is continuous, and people location can in principle be continuously inferred through visual tracking. On the other hand, audio cues are useful, whenever available, to robustly reinitialize a tracker, and to keep a tracker in place when visual clutter is high.

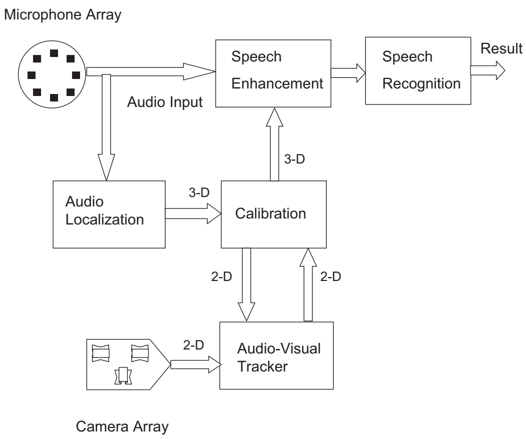


Figure 1: *System block diagram. The microphone array provides audio inputs to the speech enhancement and audio localization modules. 3-D audio localization estimates are generated by the audio localization, which are mapped onto the corresponding 2-D image plane by the calibration module. The audio-visual tracker processes this 2-D information along with the visual information from the camera array to track the active speakers. The 3-D estimates are reconstructed by the calibration module from the two different camera views which are then input to the speech enhancement module. The enhanced speech from the speech enhancement module, which comprises of beamformer followed by a post-filter, is used as input to the speech recognition module.*

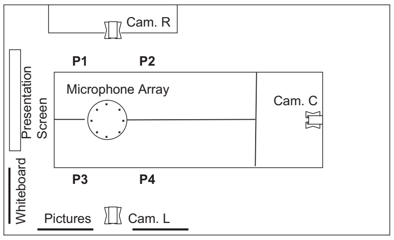
Our approach uses data captured by a fully calibrated audio-visual sensor array consisting of three cameras and a small microphone array, which covers the meeting workspace with pair-wise overlapping views, so that each area of the workspace is viewed by two cameras. The sensor array configuration and calibration are further discussed in Section 4. In our methodology, the 2-D location of each person visible in each camera plane is continuously estimated using a Bayesian multi-person state-space approach. For multi-person state configurations in camera plane k, $X_t^k = (X_{i,t}^k)$, $i \in \mathcal{I}_t$, where $\mathcal{I}_t$ is the set of person identifiers described in the previous paragraph, and $X_{i,t}^k$ denotes the configuration of person i, and for audio-visual observations $Y_t^k = (Y_t^{k,a}, Y_t^{k,v})$, where the vector components $Y_t^{k,a}$ and $Y_t^{k,v}$ denote the audio and visual observations, respectively, the filtering distribution of states given observations $p(X_t^k|Y_{1:t}^k)$ is recursively approximated using a Markov Chain Monte Carlo (MCMC) particle filter. This algorithm is described in Section 5. For this, a set of 3-D audio observations, $\{Y_t^{a3}\}$ are derived at each time-step using a robust source localization algorithm based on the SRP-PHAT measure [23]. Using the sensor calibration method described in Section 4, these observations are mapped onto the two corresponding camera image planes by a mapping function $f_\Phi : \mathbb{R}^3 \rightarrow (\{0, 1, 2\} \times \mathbb{R}^2)^2$, where Φ indicates the camera calibration parameters, which associates a 3-D position with a six-dimensional vector containing the camera index k_t^j and the 2-D image position $Y_t^{k_t^j,a}$ for the corresponding pair of camera planes, $j \in \{1, 2\}$. Visual observations are extracted from the corresponding image planes. Finally, for each person, the locations estimated by the trackers, $\hat{X}_{i,t}^{k_t^1}$, $\hat{X}_{i,t}^{k_t^2}$ for the corresponding camera pair, k_t^1, k_t^2, are merged. The corresponding 3-D location estimate is obtained using the inverse mapping $\hat{Z_{i,t}} = f_\Phi^{-1}(k_t^1, \hat{X}_{i,t}^{k_t^1}, k_t^2, \hat{X}_{i,t}^{k_t^2})$.

The 3-D estimated locations for each person are integrated with the beamformer as described in Section VI. 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 recalculated. For further speech signal enhancement, the beamformer is followed by a post-filtering stage. After speech enhancement, speech recognition is performed on the enhanced signal. This is discussed in Section VII. In summary, a baseline speech recognition system is first trained using the headset microphone data from the original Wall Street Journal corpus. A number of adaptation techniques, including Maximum Likelihood Linear Regression (MLLR) and Maximum a Posteriori (MAP) are used to compensate for the channel mismatch between the training and test conditions. Finally, to fully compare the effects of audio vs. audio-visual estimation of location in speech enhancement and recognition, the audio-only location estimates directly computed from the speaker localization algorithm are also fed into the enhancement and recognition blocks of our approach.

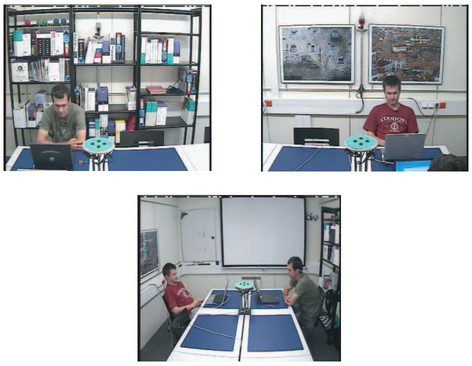
4 Audio-visual sensor array

4.1 Sensor configuration

All the data used for experiments is recorded at the instrumented meeting room. The meeting room is a 5.6m×3.5m containing a 4.8m×1.2m rectangular table. Figure 2.a shows the layout, positions of microphone array, video cameras and typical speaker positions in the meeting room. The sample images of the three views from the meeting room are as shown in Figure 2.b. 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 three wide-angle cameras (center, left, and right) giving complete view of the room. 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. The meeting room allows capture of fully synchronized audio and video data.



a



b

Figure 2: *Schematic diagram of the meeting room and the left, right and center sample images. Cam.,C,L,R denotes camera, center, left, and right. P1, P2, P3, P4 indicate the typical speaker positions. Meeting room contains visual clutter due to bookshelves, skin-colored posters as shown in the above pictures. Audio clutter is caused from the laptops and other computers in the room. Speakers act naturally with no constraints on speaking styles or accents.*

4.2 Sensor calibration

To relate points in the 3-D camera reference with 2-D image points, we calibrate the three cameras (center, left, and right) of the meeting room to a single 3-D external reference, using a standard camera calibration procedure [44]. This method, with a given number of image planes represented by checkerboard at various orientations, estimates the different camera parameters which define an affine transformation relating the camera reference and the 3-D external reference. 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 define a transformation for basis change between microphone array reference and the 3-D external reference. Finally, to complete the audio-video mapping we 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 are used to find the intersection of the two rays, which correspond to the reconstructed 3-D point [18]. This last step is used to map the results of audio-visual tracker, (i.e, the speaker location in the image plane) back to the 3-D points, as input to the speech enhancement module.

5 Person tracking

To jointly track multiple people in each image plane, we use the probabilistic multimodal multi-speaker tracking method recently proposed in [15], consisting of a Dynamic Bayesian Network in which approximate inference is performed by an MCMC particle filter [25, 12]. In the following, the notation is simplified with respect to Section 3 by dropping the camera index symbol, so multi-person configurations $X_t^k = (X_{i,t}^k)$ are denoted by $X_t = (X_{i,t})$, observations Y_t^k by Y_t, etc.

The multi-person state-space formulation, in which configurations represent the joint state of all people in a scene, is both mathematically rigorous and allows for the explicit definition of object interaction models. Given a multi-person dynamical model $p(X_t|X_{t-1})$, and a multi-person observation likelihood model $p(Y_t|X_t)$, a particle filter recursively approximates the filtering distribution $p(X_t|Y_{1:t})$ by a weighted set of N_s particles $\{X_t^{(n)}, w_t^{(n)}\}_{n=1}^{N_s}$, using the particle set at the previous time step, $\{X_{t-1}^{(n)}, w_{t-1}^{(n)}\}$, and the new observations,

$$p(X_t|Y_{1:t}) \approx \mathcal{Z}^{-1} p(Y_t|X_t) \sum_{n=1}^{N_s} w_{t-1}^{(n)} p(X_t|X_{t-1}^{(n)}). \tag{1}$$

Given the multi-person state-space, a mixed state-space is further defined for single-person configurations $X_{i,t}$, where both the geometric transformations of a person's head model template (in our work, an elliptical silhouette) in the image plane and the speaking activity are tracked. A single-person state is composed of a continuous vector of transformations comprising 2-D translation and scaling, and a discrete binary variable modelling the person speaking activity.

As can be seen from Equation 1, the three key elements of the approach are the dynamical model, the observation likelihood model, and the sampling mechanism. Firstly, the dynamical model includes two factors: one describing independent single-person dynamics, and another one explicitly modelling pairwise interactions, constraining the dynamics of each person based on the state of the others, via a pairwise Markov Random Field (MRF) prior. Secondly, the observation model is derived from both audio and video. On one hand, audio observations are derived from a speaker localization algorithm as follows. A sector-based source localization algorithm based on SRP-PHAT measure is used to generate candidate 3-D locations of people when they speak [23]. Given the higher sampling rate for audio, multiple audio localization estimates are merged for each video frame. We then use the sensor calibration procedure described in the previous section to project the 3-D audio estimates on the corresponding 2-D image planes. The audio observation likelihood relates the distance between the audio localization estimates and the candidate configurations on the image plane. As stated in Section 3, the audio observations increase the robustness of the tracker, and are specially effective at maintaining adequate tracking in cases of high visual clutter. On the other hand, 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 spatial structure observations are based on a parametric representation of the overlap between skin-color blobs and head configurations. These two visual cues complement each other, as the first one is edge-oriented while the second one is region-oriented. Finally, the approximation of Eq. 1 in the (possibly) high-dimensional space defined by multiple people is done with MCMC techniques, more specifically designing a Metropolis-Hastings sampler at each time-step in order to efficiently place samples as close as possible to regions of high likelihood. For this purpose, a proposal distribution in which the configuration of one single object is modified at each step of the chain is defined, and each move in the chain is accepted or rejected according to the Metropolis-Hastings algorithm. After discarding an initial set of samples (so-called burn-in period), the generated MCMC samples will approximate the target filtering distribution [25]. More details can be found in [15].

The output of the multi-person tracker is represented by the mean of the filtering distribution at each time-step. From here, the 2-D locations of each person's head center for the specific camera pair where such person appears, which correspond to the translation components of the mean configuration in each camera and are denoted (slightly abusing the notation) by $\hat{X}_{i,t}^{k1}$, $\hat{X}_{i,t}^{k2}$, can be extracted and triangulated to obtain the corresponding 3-D locations $\hat{Z_{i,t}}$. These 3-D points are finally used as input to the speech enhancement module, as described in the following section.

6 Speech enhancement

The microphone array speech enhancement system is the same as that used in [29]. It includes a filter-sum beamformer followed by a post-filtering stage, as described in the following.

6.1 Beamformer

For the beamformer, we use the superdirective technique to calculate the channel filters maximizing the array gain, while maintaining a minimum constraint on the white noise gain. This technique is fully described in [8].

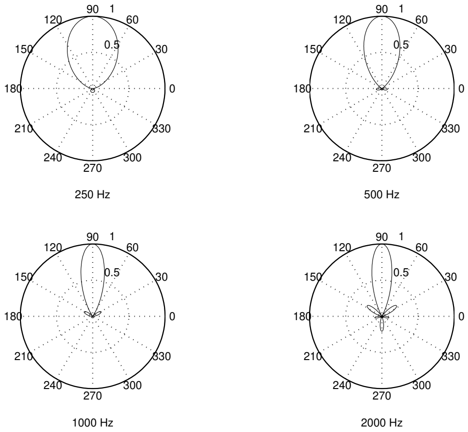


Figure 3: *Horizontal polar plot of the directivity pattern of the superdirective beamformer for an 8-element circular array of radius 10cm*

Figure 3 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 2cm, the beamfomer channel filters were recalculated.

6.2 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 [27], and lead to better performance in speech recognition applications [31]. 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 [35] 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 [35] 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, $h_s(f)$, $s = 1 : S$,

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

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. To achieve this effect, in the single speaker scenarios in this paper a second beamformer was oriented to the opposite side of the table for use in the above post-filter. This post-filter also has the benefit of low computational cost compared to other formulations, such as those based on [43], which require the calculation of channel auto- and cross-spectral densities.

7 Speech Recognition

With the ultimate goal of automatic speech recognition, speech recognition tests are performed for the stationary, moving speaker, and overlapping speech scenarios. This is also important to quantify the distortion to the desired speech signal. For the baseline, a full HTK based recognition system, trained on the original Wall Street Journal database (WSJCAM0) is used [34]. 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. 52-element feature vectors were used, comprising of 13 MFCCs (including the 0th cepstral coefficient) with their first, second, and third order derivatives. 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 [17], 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 si_dt5a 5000-word task and 20.44% WER on the si_dt20a task 20000 word task, which roughly corresponds to the results reported in the SQALE evaluation using the WSJCAM0 database [42].

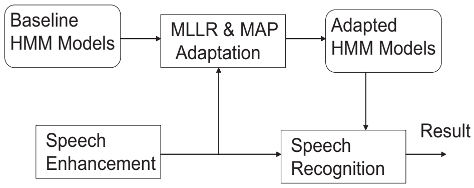


Figure 4: *Speech recognition adaptation. The baseline HMM models are adapted using MLLR and MAP techniques. The acoustics of the enhanced speech signal from speech enhancement block are adjusted to improve the speech recognition performance.*

To reduce the channel mismatch between the training and test conditions, the baseline HMM models are adapted using a maximum likelihood linear regression (MLLR) [24] and maximum-a-posteriori (MAP) [16] adaptation as shown in figure 4. 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).

8 Experiments and results

The following sections describe the database specification, tracking, speech enhancement and recognition results. The results, along with the additional meeting room data results for single speaker, specifically switching seats and overlap speech from two side-by-side simultaneous speakers can be viewed at the companion website: `www.idiap.ch/ ~ hakri/avsensorarray/avdemos.htm`

8.1 Database specification

All the experiments are conducted on a subset of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus. The specification and structure of the full corpus are detailed in [26]. We used a part of *Single speaker stationary*, *Single speaker moving*, and *Stationary overlapping speech* data. In single speaker stationary, the speaker reads out sentences from different positions within the meeting room and in single speaker moving, the speaker is moving between the different positions while reading the sentences. In overlapping speech, two speakers simultaneously read sentences from different positions within the room. Most of the data comprised of non-native English speakers with different speaking styles and accents. The data is divided into development (DEV) and evaluation (EVAL) sets with no common speakers in both the sets. Table 1 describes the data used for the experiments.

Table 1: Data description

Scenario	No. of sentences	Total time(min.)	No. of speakers
Stationary	160	22	6
Moving	78	12	6
Overlap	70	11	4

8.2 Tracking Experiments

The multi-person tracking algorithm was applied to the data set described in the previous section, for each of the three-scenarios (stationary single-person , moving single-person, and two-person overlap). In the tracker, all models that require a learning phase (e.g. spatial structure head model), and all parameters that are manually set (e.g. dynamic model parameters), were learned or set using a separate data set, originally described in [15], and kept fixed for all experiments. Regarding the number of particles, experiments were done for 500, 800, and 500 particles for the stationary, moving, and overlap cases, respectively. In all cases, 30% of the particles were discarded during the burn-in period of the MCMC sampler, and the rest was kept for representing the filtering distribution. It is important to notice that the number of particles was not tuned but simply set to an sensible fixed value, following the choices made in [15]. While the system could have performed adequately with less particles, the dependence on the number of particles was not investigated here. All reported results are computed from a single run of the tracker.

The algorithm was objectively evaluated the following procedure. The 3-D euclidean distance between a ground truth location of the speakers mouth and the automatically estimated location was used as performance measure. 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 the inverse mapping. The ground truth was produced at a rate of 1 frame/sec, i.e., every 25 video frames. The 3-D Euclidean distance is then averaged over all frames in the data set.

The results are presented in Table 2, Figure 5, and the companion website. Table 2 summarizes the overall results, Figure 5 illustrates typical results for two minutes of data for each of the scenarios. Selected frames from such videos are presented in Figures 6, and the corresponding videos can be seen in the companion website. In the images and videos, the tracker output is displayed for each person as an ellipse of distinct color. Inferred speaking activity is shown as a double ellipse with contrasting tones.

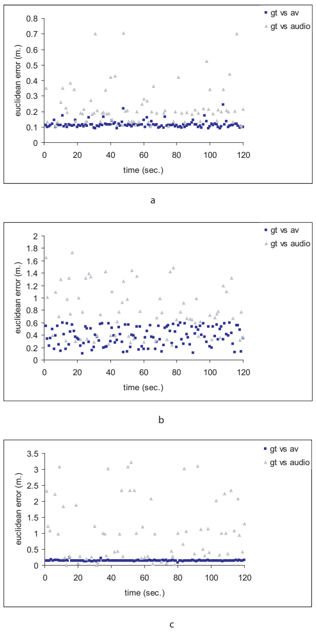


Figure 5: *Tracking results for (a) stationary, (b) moving speaker, and (c) overlapping speech, for 120 seconds of video for each scenario. For audio, the 'average' error is computed (see text). Audio estimates are discontinuous and available around 60 % of the times. The audio-visual estimates are continuous and more stable.*

From Figures 5(a) and (c), we can observe that the continuous estimation of 3-D location is quite stable in cases where speakers are seated, and the average error remains low (on average 12 cm for stationary, and 22 cm for overlap, as seen in Table 2). These errors are partially due to the fact that the tracker estimate in each camera view corresponds to the center of a person's head, which introduces errors because, in strict terms, the two head centers do not correspond to the 3-D same physical point, and also because they do not correspond to the mouth center. The overlap case is clearly more challenging than the stationary one. For the moving case, illustrated in Figure 5(b), we observe an increased although still acceptable error (38 cm in average), which can be explained at least partially due to the inaccuracy of the dynamical model, e.g. when the speaker stops and reverses the direction of motion, the tracker needs some time to adjust. This is evident in the corresponding video.

To validate our approach with respect to an audio-only algorithm, we also evaluated the results using directly the 3-D output of the speaker localization algorithm. Results are also shown in Table 2, Figures 5, and 6, and the website videos. In images and videos, the audio-only estimates are represented by '+' symbols. Remember that the audio localization algorithm outputs between zero and three audio estimates per video frame. Using this information, we compute two types of errors. The first one uses the *average* euclidean distance between the ground truth and the available audio

Figure 6: *Tracking a single speaker in (a-b) stationary, (c-d) moving, and two speakers in (e-f) overlap speech scenarios. The speaker is tracked in each view (red ellipse). A double red and yellow ellipse (b and d) indicates when the speaker is active. In two speaker case, a double green and yellow indicates the first active speaker and double red and yellow indicates the other active speaker (e).*

estimates. The second one uses the *minimum* euclidean distance between ground truth and automatic results, which explicitly considers the best (a posteriori) available estimate. While the first case can be seen as a fair, blind evaluation, the second case can be seen as a best-case scenario, in which a form of data association has been done. As shown in Figure 5, 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. The results show that, in all cases, the performance obtained with audio-only information is consistently worse than the one obtained with the multimodal approach, regarding both means and standard deviation. When the average euclidean distance is used, performance degrades by almost 100% for stationary, and even more severely for moving and overlap. Furthermore, while the best-case scenario results (minimum euclidean distance) clearly reduce the errors for audio, due to the a posteriori data association, they nevertheless remain consistently worse than the obtained with the audio-visual approach, which is clearly more accurate. Very importantly, compared to the audio-visual case, the reliability of the audio estimates (for both average and minimum) degrades much more considerably when going from the single-speaker case to the concurrent-speakers one. This is an evident limitation of audio-only approaches, for which a

Table 2: Tracking results. 3-D error between groundtruth and automatic methods. The standard deviation is in brackets.

	Error (m)		
Scenario	Audio-visual	Audio (average)	Audio (minimum)
Stationary	0.12 (0.022)	0.23 (0.126)	0.20 (0.100)
Moving	0.38 (0.148)	0.84 (0.403)	0.70 (0.363)
Overlap	0.22 (0.014)	0.93 (0.932)	0.60 (0.468)

multimodal approach, as the Table and Figures show, brings clear benefits.

8.3 Speech Enhancement and Recognition Experiments

To assess the noise reduction and evaluate the effectiveness of the microphone array in acquiring a "clean" speech signal, the segmental signal-to-noise ratio (SNR) is calculated. To normalize for different levels of individual speakers, all results are quoted with respect to the input on a single table-top microphone, and hence represent the SNR enhancement (SNRE). These results are shown in Table 3.

Table 3: SNRE results

	SNRE (dB)		
Signal	Stationary	Moving	Overlap
Headset	24.8	23.4	17.4
Lapel	16.2	15.2	11.2
Audio Beamformer	13.1	11.8	5.2
Audio Beamformer + Post-filter	14.4	12.2	5.6
AV Beamformer	15.3	13.4	6.7
AV Beamformer + Post-filter	16.8	13.6	10.1

Table 4: Adaptation and test data description

	No. of sentences	
Scenario	Adaptation	Testing
Stationary	60	100
Moving	30	48
Overlap	24	46

Speech recognition experiments were performed to evaluate the performance of the various scenarios. The number of sentences for adaptation and test data are as shown in Table 4. Adaptation data was taken from the DEV set and test data was taken from the EVAL set. Adaptation data was matched to the corresponding testing channel condition. In MLLR adaptation, a static two-pass approach was used, where in the first pass a global transformation was performed. In the second pass a set of specific transforms for speech and silence models were calculated. The MLLR transformed means are used as the priors for the MAP adaptation. All the results are scenario specific, due to the different amounts of adaptation and test data. Table 5 shows the speech recognition results after adaptation.

In the following, we summarize the discussion regarding the speech enhancement and speech recognition experiments.

Table 5: Speech recognition results

	WER (%)		
Signal	**Stationary**	**Moving**	**Overlap**
Headset	21.3	19.3	42.9
Lapel	27.9	24.4	49.7
Distant Microphone	37.5	37.6	95.3
Audio Beamformer	31.3	33.3	81.7
Audio Beamformer + Post-filter	32.8	34.6	80.3
AV Beamformer	26.8	28.5	69.4
AV Beamformer + Post-filter	26.3	29.4	56.6

Headset, lapel and distant microphones : As can be seen from 3 and 5, as expected for all the scenarios (stationary, moving, and overlap speech) and all the testing conditions (headset, lapel, distant, audio beamformer, audio beamformer + post-filter, audio-visual (AV) beamformer, AV beamformer + post-filter), the headset speech has the highest SNRE, which in turn results in the best speech recognition performance. Note that the obtained WER corresponds to the typical recognition results with conversational speech comparable with the 20.5% obtained with the baseline system described in the previous paragraph. The headset case can thus be considered as the baseline for all the results from the other channels to be compared. The lapel microphone performance is the next result close to headset, due to its close proximity (around 8 cm.) to the mouth of the speaker. Regarding the distant microphone signal, WER obtained in this case is due to the greater susceptibility to room reverberation and low SNR, because of its distance (around 80 cm.) from the desired speaker. In all cases, the severe degradation in SNRE and WER for the overlap case compared to the single speaker case is self-evident, although obviously headset is the most robust case.

Audio-only:The audio beamformer and audio beamformer + post-filter perform better than the distant microphone for all scenarios, for both SNRE and WER. It can be observed that the post-filter helps in acquiring better speech signal than the beamformer. However, the SNR and WER performances are in all the cases inferior when compared to the headset and lapel microphone cases. This is likely due to the fact that the audio estimates are discontinuous and not available all the times, are affected by audio clutter due to laptops and computers in the meeting room, and are highly vulnerable to the room reverberation.

Audio-visual: From Tables 3 and 5, it is clear that the AV beamformer and AV beamformer + post-filter cases perform consistently better than the distant microphone and audio-only systems for both SNRE and WER. It can also be observed that the post-filter helps in acquiring better speech signal than the beamformer only. In the single stationary speaker scenario, the AV beamformer + post-filter performs better than lapel, suggesting that the post-filter helps in speech enhancement without distorting the beamformed speech signal. This is consistent with some earlier studies which have shown that recognition results from beamformed channels can be comparable or sometime better than the lapel microphones [31]. In overlap speech scenario, the post-filter is highly effective in reducing the crosstalk speech, significantly improving on the beamformer output and getting quite close to the lapel case in terms of SNRE, but slightly less so in terms of WER. Some examples of enhanced speech can be seen in the companion website.

9 Conclusion

This paper has presented an integrated framework for audio-visual tracking based microphone array speech recognition system. An audio-visual multi-person tracker is used to track the active speakers with high accuracy, which is then used as input to the superdirective beamformer. Based on the location estimates, the beamformer enhances the speech signal emanating only from the desired speaker, attenuating signals from the other competing sources. The beamformer is then followed by a novel post-filter which helps in further speech enhancement by reducing the background noise. The enhanced speech is finally input into a speech recognition module to evaluate the quality of the speech signal.

The system has been evaluated on the real data in the meeting room for stationary speaker, moving speaker and overlap speech scenarios and for all the testing conditions i.e. headset, lapel and distant microphones, and audio-only and audio-visual (AV) based systems comprising of audio beamformer, audio beamformer + post-filter, AV beamformer and AV beamformer + post-filter. The results illustrate that the audio-visual tracking based microphone array speech enhancement and recognition system performs significantly better than single table-top microphone and comparable to lapel microphone for all the scenarios. The results also show that the audio-visual based system performs significantly better than audio-only system in terms of signal-to-noise ratio enhancement (SNRE) and word error rate (WER). This demonstrates that accurate speaker tracking provided by multimodal approach proved beneficial to improve speech enhancement, which resulted in improved speech recognition performance.

Acknowledgment

This work is supported by the European Projects Augmented Multi-party Interaction (AMI, EU-IST project FP6-506811), and Detection and Identifications of Rare Audio-visual Cues (DIRAC, EU-IST project FP6-003758). We thank Darren Moore for help with the initial speech enhancement experiments, Jithendra Vepa for his support with the speech recognition system, Mike Lincoln for the collaboration in designing the MC-WSJ-AV corpus, Sileye Ba for his support with the audio-visual sensor array calibration, and Bastien Crettol for his support to collect the data. We also like to thank all the participants involved in the recording of the corpus.

References

[1] G. Abowd et al, “Living laboratories: The future computing environments group at the georgia institute of technology,” *Proc. of the Conference on Human Factors in Computing Systems (CHI),* , Hague, April, 2000.

[2] F. Asano et al, “Detection and Separation of Speech Event using audio and video information fusion,” *Journal of Applied Signal Processing*, vol.11, pp. 1727–1738, 2004.

[3] M. Beal, H. Attias, and N. Jojic. ”Audio-Video Sensor Fusion with Probabilistic Graphical Models,” *Proc. of the European Conf. on Computer Vision (ECCV)*, Copenhagen, May, 2002.

[4] N. Checka, K. Wilson, M. Siracusa and T. Darrell. ”Multiple Person and Speaker Activity Tracking with a Particle Filter,” *Proc. of the Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, May, 2004.

[5] R. Chellapa, C. Wilson and A. Sirohey. ”Human and machine recognition of faces: A survey,” *Proc. of the IEEE*, 83(5):705–740, 1995.

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

[7] S. M. Chu, E. Marcherat, and G. Potamianos. ”Automatic speech recognition and speech activity detection in the CHIL smart room,” *Proc. of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, July, 2005.

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

[9] H. Cox, R. Zeskind, and I.Kooij. "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(3):393–397, June, 1986.

[10] 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.

[11] J. DiBiase. "A high-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments," *Ph.D Thesis*, Brown University, Providence RI, 2000.

[12] A. Doucet, N. de Freitas and N. Gordon. *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[13] J. Fisher, T. Darrell, W. T. Freeman and P. Viola. "Learning Joint Statistical Models for Audio-Visual Fusion and Segregation," *Proc. of Neural Information Processing Systems (NIPS)*, Denver, Dec., 2000.

[14] D. Gatica-Perez, G. Lathoud, I. McCowan, J. -M. Odobez, and D. Moore. "Audio-visual speaker tracking with importance particle filters," *Proc. of the IEEE Conf. on Image Processing (ICIP)*, Barcelona, Oct., 2003.

[15] 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.

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

[17] 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.

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

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

[20] J. Kleban and Y. Gong. "HMM adaptation and microphone array processing for distant speech recognition," *Proc. of Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, June, 2000.

[21] C. Knapp and G. Carter. "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24:320–327, Aug., 1976.

[22] H. Krim and M. Viberg. "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, 13:67–94, July, 1996.

[23] 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.

[24] 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.

[25] J.S. Liu. *Monte Carlo Strategies in Scientific Computing* , Springer-Verlag, 2001.

[26] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti. "The Multi-Channel Wall Street Journal Audio-Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments ," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Dec., 2005.

[27] K. Uwe Simmer, J. Bitzer and C. Marro. "Post-filtering Techniques," *Microphone Arrays*, Springer, 3:36–60, 2001.

[28] 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.

[29] I. McCowan and H. Bourlard. "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716, Nov., 2003.

[30] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. "The Meeting Project at ICSI," *Proc. of the Human Language Technologies Conf.*, San Diego, March, 2001.

[31] D. Moore and I. McCowan. "Microphone array speech recognition: Experiments on overlapping speech in meetings," *Proc. of Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April, 2003.

[32] J.G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, C. Laprun. "The Rich Transcription 2005 Spring Meeting Recognition Evaluation," *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, July, 2005.

[33] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani. "Microphone array based speech recognition with different talker-array positions," *Proc. of Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, April, 1997.

[34] 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.

[35] S.Roweis. "Factorial Models and Refiltering for Speech Separation and Denoising," *Proc. of the Eurospeech Conf. on Speech Communication and Technology (Eurospeech-2003)* , Geneva, Sep., 2003.

[36] 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.

[37] D. Sturim, M. Brandstein and H. Silverman. "Tracking multiple talkers using microphone-array measurements," *Proc. of the Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, April, 1997.

[38] B. D. Van Veen and K. M. Buckley. "Beamforming: A versatile approach to spatial filtering ," *IEEE Acoustics, Speech, and Signal Processing Magazine*, 5(2):4–24, April, 1988.

[39] J. Vermaak, M. Gagnet, A. Blake and P. Perez. "Sequential Monte Carlo Fusion of Sound and Vision For Speaker Tracking," *Proc. of the Int.Conf. on Computer Vision (ICCV)*, Vancouver, July, 2001.

[40] A. Waibel, T. Schultz, M. Bett, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "Smart:The Smart Meeting Room Task At ISL," in *Proc. of Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* , Hong Kong, April, 2003.

[41] D. Ward and R. Williamson. "Particle filter beamforming for acoustic source localization in a reverberant environment," *Proc. of the Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, May, 2002.

[42] S.J. Young et al. "Multilingual large vocabulary speech recognition: the European SQUALE project," *Computer Speech and Language*, 11(1):73-89, 1997.

[43] R.Zelinski. "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *Proc. of the Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, April, 1988.

[44] Z. Zhang. "Flexible camera calibration by viewing a plane from unknown orientations," *Proc. of the Int.Conf. on Computer Vision (ICCV)*, Kerkyra, Sep., 1999.

[45] D. Zotkin, R. Duraiswami and L. Davis. "Multimodal 3-D tracking and event detection via the particle filter," *Proc. of the Int.Conf. on Computer Vision, Workshop on Detection and Recognition of Events in Video (ICCV-EVENT)*, Vancouver, July, 2001.