Microphone Array Driven Speech Recognition: Influence of Localization on the Word Error Rate

Matthias Wölfel, Kai Nickel and John McDonough

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
email: {wolfel, nickel, jmcd}@ira.uka.de

Abstract. Interest within the *automatic speech recognition* (ASR) research community has recently focused on the recognition of speech captured with one or more microphones located in the far field, rather than being mounted on a headset and positioned next to the speaker's mouth. Far field ASR is a natural application for beamforming techniques using an array of microphones. A prerequisite for applying such techniques, however, is a reliable means of speaker localization. In this work, we compare the accuracy of source localization systems based on only audio features, only video features, as well as a combination of audio and video features using speech data collected during seminars held by actual speakers. We also investigate the influence of source localization accuracy on the *word error rate* (WER) of a far field ASR system, comparing the WERs obtained with position estimates from several automatic source localizers with those obtained from true speaker positions. Our results reveal that accurate speaker localization is crucial for minimizing the error rate of a far field ASR system.

1 Introduction

Interest within the *automatic speech recognition* (ASR) research community has recently focused on the recognition of speech captured with one or more microphones located in the far field, rather than being mounted on a headset and positioned next to the speaker's mouth. Far field ASR is a natural application for beamforming techniques using an array of microphones, which has been shown to provide superior sound capture capability with respect to a single microphone both in terms of *signal-to-noise ratio* (SNR) and *word error rate* (WER). A prerequisite for applying such techniques, however, is a reliable means of speaker localization. In prior work, we used an extended Kalman filter to directly update position estimates in an audio only speaker localization system based on time delay of arrival [1]. In other work, we enhanced our Kalman filter-based audio localizer with video information to obtain more accurate position estimates [2]. We have also proposed an audio-video source localizer based on a particle filter [3], which for some applications has several advantages as compared to the

[0] This work was sponsored by the European Union under the integrated project CHIL, *Computers in the Human Interaction Loop*, contract number 506909.

conventional Kalman filter. In this work, we compare the accuracy of source localization systems using only audio features, only video features, as well as a combination of audio and video features. We also investigate the influence of source localization accuracy on the WER of a far field ASR system, comparing the WERs obtained with position estimates from several automatic source localizers with those obtained from true speaker positions. To provide a baseline, we also compare the performance of our far field ASR system with the performance from a *close-talking microphone* (CTM).

The speech material used in our empirical studies was collected as part of the European Union integrated project CHIL [4], *Computers in the Human Interaction Loop*, which aims to make significant advances in the fields of speaker localization and tracking, speech activity detection and far field ASR. The corpus is comprised of seminars and oral presentations collected with both near and far field microphones. In addition to the audio sensors, the seminars were recorded by calibrated video cameras. This simultaneous audio-visual data capture enables the realistic evaluation of component technologies as was never possible with earlier data bases. One of the long term goals of the CHIL project is the reliable recognition of speech in a real reverberant environments, without any constraint on the number of simultaneously active sound sources. This problem is surpassingly difficult, given that speech recorded with far field microphones is generally degraded by both background noise and reverberation. Moreover, our speech material is challenging for other reasons: The style of the speech varies greatly, from spontaneous to read, and contains many of the artifacts seen in spontaneous speech, such as filled pauses, restarts and hyper articulation. Although the seminars were held in English, many of the speakers are non-native and hence speak with pronounced European accents. In addition, the seminars are most often concerned with automatic speech recognition and related topics, which implies that recognition vocabularies and language models built with the standard corpora of training text are poorly matched to this recognition task.

The remainder of this article is organized as follows. Section 2 describes the development of a baseline system at the Universität Karlsruhe (TH) including data collection and labeling, speaker localization, beamforming, language model training and acoustic training and adaptation. Section 3 presents a variety of source localization and ASR experiments using different types of acoustic source localization schemes. Finally, Section 4 concludes the presented work and give plans for future work.

2 Baseline System

The CHIL seminar data present significant challenges to both modeling components used in ASR, namely the language and acoustic models. With respect to the former, the currently available CHIL data primarily concentrates on technical topics with focus on speech research. This is a very specialized task that contains many acronyms and therefore is quite mismatched to typical language models currently used in the ASR literature. Furthermore, large portions of the

data contain spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. On the acoustic modeling side, and in addition to the latter difficulty, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech. The above problems are compounded by the fact that, at this early stage of the CHIL project, not enough data is available for training new language and acoustic models matched to this seminar task, and thus one has to rely on adapting existing models that exhibit gross mismatch to the CHIL data. Clearly, these challenges present themselves in both close-talking microphone data, as well as the far-field data captured using the *microphone array* (MA).

2.1 Data Collection and Labeling

The data used for the experiments described in this work was collected during a series of seminars held by students and visitors at the Universität Karlsruhe (TH), Germany, since Fall 2003. The students and visitors spoke English, but mainly with German or other European accents, and with varying degrees of fluency. This data collection was done in a very natural setting, as the students were far more concerned with the content of their seminars, their presentation in a foreign language and the questions from the audience than with the recordings themselves. Moreover, the seminar room is a common work space used by other students who are not seminar participants. Hence, there are many "real world" events heard in the recordings, such as door slams, printers, ventilation fans, typing, background chatter, and the like.

The seminar speakers were recorded with a Sennheiser CTM, a 64-channel Mark III MA developed at NIST (National Institute of Standards and Technologies) mounted on the wall, four T-shaped MAs with four elements mounted on the four walls of the seminar room and three Shure Microflex table-top microphones located on the work table where the position was not fixed. A diagram of the seminar room is shown in Figure 1. All audio files have been recorded at 44.1 kHz with 24 bits per sample. The high sample rate is preferable to permit more accurate position estimations, while the higher bit depth is necessary to accommodate the large dynamic range of the far field speech data. For the recognition process the speech data was down-sampled to 16 kHz with 16 bits per sample. In addition to the audio data capture, the seminars were simultaneously recorded with four calibrated video cameras that are placed at a height of 2.7 m in the room corners. Their joint field of view covers almost the entire room. The images are captured at a resolution of 640x480 pixels and a framerate of 15 frames per second, and stored as jpg-files for offline processing.

The data from the CTM was manually segmented and transcribed. The data from the far distance microphones was labeled with speech and non-speech regions. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.7 second. Based on this marks the true position of the speaker's head (ground truth) in three dimensions could be calculated within an accuracy of approximately 10 cm [5].

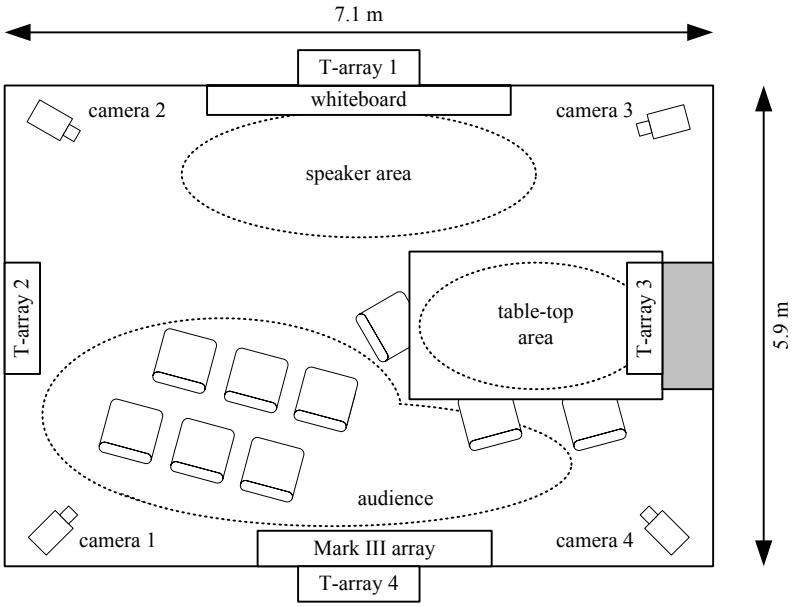

Fig. 1. The CHIL seminar room layout at the Universität Karlsruhe (TH).

2.2 Speaker Localization: Audio Features

The lecturer is the person that is normally speaking, therefore we can use audio features using multiple microphones to detect the speaker position.

Consider the j-th pair of microphones, and let $\mathbf{m}_{j1}$ and $\mathbf{m}_{j2}$ respectively be the positions of the first and second microphones in the pair. Let $\mathbf{x}$ denote the position of the speaker in a three dimentional space. Then the *time delay of arrival* (TDOA) between the two microphones of the pair can be expressed as

$$T_j(\mathbf{x}) = T(\mathbf{m}_{j1}, \mathbf{m}_{j2}, \mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_{j1}\| - \|\mathbf{x} - \mathbf{m}_{j2}\|}{c} \tag{1}$$

where c is the speed of sound. To estimate the TDOAs a variety of well-known techniques [6, 7] exist. Perhaps the most popular method is the *phase transform* (PHAT), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} \, e^{j\omega\tau} \, d\omega \tag{2}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals of a microphone pair in a microphone array. Normally one would search for the highest peak in the PHAT to estimate the TDOA. As we are using a particle filter framework, described in Section 2.4, we need to calculate the probability of the acoustic observation A given that the state of the system is characterized by the current

particle s_i. We decompose the acoustic observation A into m pairs of microphones (in our case 12), such that the total probability is given by

$$p(A|\mathbf{s}_i) = \frac{1}{m}\sum_{j=1}^{m} p(A^j|\mathbf{s}_i) \tag{3}$$

In order to obtain a (pseudo) probability score for each microphone pair j, we consider the PHAT value at the time delay position $T_j(\mathbf{x} = s_i)$ given a particular particle s_i. As the values returned by the PHAT can be negative, but probability density functions must be strictly nonnegative, we found that setting all negative values of the PHAT to zero yielded the best results.

$$p(A^j|\mathbf{s}_i) = \max(R_j(T_j(\mathbf{x} = s_i)), 0) \tag{4}$$

To get the final probability distribution which tells us how likely the acoustic observation A is produced by a particle s_i, we must normalize over all particles:

$$\overline{p}(A|\mathbf{s}_i) = \frac{p(A|\mathbf{s}_i)}{\sum_i p(A|\mathbf{s}_i)} \tag{5}$$

2.3 Speaker Localization: Video Features

For the task of person tracking in video sequences, there is a variety of features to choose from. In our lecture scenario, the problem comprises both locating the lecturer and disambiguating the lecturer from the people in the audience. As lecturer and audience cannot be separated reliably by means of fixed spatial constraints as, e.g., a dedicated speaker area, we have to look for features that are more specific for the lecturer than for the audience.

Intuitively, the lecturer is the person that is standing and moving (walking, gesticulating) most, while people from the audience are generally sitting and moving less. In order to exploit this specific behavior, we decided to use dynamic foreground segmentation based on adaptive background modeling as primary feature, a detailed explanation can be found in [3]. In order to support the track indicated by foreground segments, we use detectors for face and upper body, also described in [3]. Both features (foreground F and detectors D) are linearly combined using a mixing weight β (for our experiments β was fixed to 0.7, this value was optimized on a development set), so that the particle weights for view j are given by

$$p(V^j|s_i) = \beta \cdot p(D^j|s_i) + (1-\beta) \cdot p(F^j|s_i) \tag{6}$$

To combine the different views, we sum over the weights from the v different cameras in order to obtain the total weight of the visual observation of the particular particle:

$$p(V|s_i) = \frac{1}{v}\sum_{j=1}^{v} p(V^j|s_i) \tag{7}$$

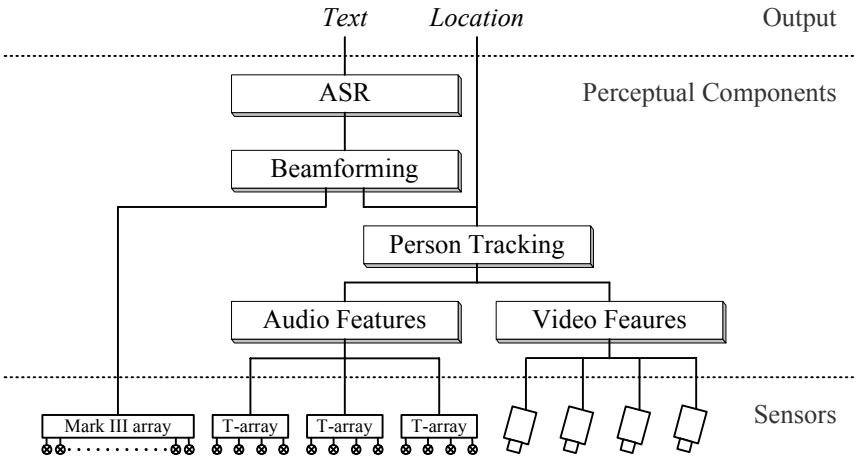

Fig. 2. Components of the system.

To get a (pseudo) probability value which tells us how well the particle s_i explains the visual observation V we have to normalize over all values:

$$\overline{p}(V|s_i) = \frac{p(V|s_i)}{\sum_i p(V|s_i)} \tag{8}$$

2.4 Data Fusion with Particle Filter

Particle filters [8] represent a generally unknown probability density function by a set of random samples $\{s_i\}$. Each of these particles is a vector in state space and is associated with an individual weight π_i. The evolution of the particle set is a two-stage process which is guided by the observation and the motion model:

1. *The prediction step:* From the set of particles from the previous time instance, an equal number of new particles is generated. In order to generate a new particle, a particle of the old set is selected randomly in consideration of its weight, and then propagated by applying the motion model. In the simplest case, this can be additive Gaussian noise, but higher order motion models can also be used.
2. *The measurement step:* In this step, the weights of the new particles are adjusted with respect to the current observation. This means, the probability $p(z_t|s_i)$ of the observation z_t needs to be computed, given that the state of particle s_i is the true state of the system.

As we want to track the lecturer's head centroid, each particle $s_i = (x, y, z)$ represents a coordinate in space. The ground plane is spanned by x and y, the height is represented by z. The particles are propagated by simple Gaussian diffusion, thus representing a coarse motion model:

$$s_i' = s_i \cdot (N_{\sigma=0.2m}, N_{\sigma=0.2m}, N_{\sigma=0.1m}) \tag{9}$$

Using the features as described before in Sections 2.2 and 2.3, we can calculate a weight π_i for each particle at the current time instance by combining the probability of the current acoustical observation A and the visual observation V using a weighting factor α:

$$\pi_i = \alpha \cdot \overline{p}(A|s_i) + (1 - \alpha) \cdot \overline{p}(V|s_i) \tag{10}$$

The weighting factor α was set by

$$\alpha = \frac{m_0}{m} \cdot 0.6 \tag{11}$$

where m is the total number of microphone pairs and m_0 the number of values above 0. The average value of α was approximately 0.4. Therefore, more weight was given to the video features.

A particle's weight is set to 0 if the particle leaves the lecture room[1] or if its z-coordinate leaves the valid range for a standing person ($1.2m < z < 2.1m$). The final hypothesis about the lecturer's location over the whole particle set $1 \ldots q$ (in our case $q = 300$) can be derived by a weighted summation over the individual particle locations $s_{i,t}$ at time t:

$$\Lambda_t = \frac{1}{q} \sum_{i=1}^{q} \pi_{i,t} \cdot s_{i,t} \tag{12}$$

Sampled Projection Instead of Triangulation

A common way to obtain the 3D position of an object from multiple views is to locate the object in each of the views and then to calculate the 3D position by using triangulation. This approach, however, has several weak points: Firstly, the object must be detected in at least two different views at the same time. Secondly, the quality of triangulation depends on the points of the object's images that are chosen as starting points for the lines-of-sight; if they do not represent the same point of the physical object, there will be a high triangulation error. Thirdly, searching for the object in each of the views separately—without incorporating geometry information—results in an unnecessarily large search space.

In the method proposed here, we avoid the aforementioned problems by not using triangulation at all. Instead, we make use of the particle filter's capacity to predict the object's location as a well-distributed set of hypotheses; i.e., many particles cluster around likely object locations, and fewer particles populate the space between. As the particle set represents a probability distribution of the predicted object's location, we can use it to narrow down the search space. So instead of searching a neighborhood exhaustively, we only look for the object at the particles' positions.

When comparing the proposed method to Kalman filter-based tracking, the following advantage becomes apparent: A particle filter is capable of modeling

[1] We restrict the particles to be within the full width of the room's ground plane ($0 < y < 7.1m$) and half of the depth ($0 < x < 3m$).

multi-modal distributions. That means in particular, that no single measurement has to be provided and no information is lost by suppressing all but the strongest measurement, as it is the case for Kalman filter. Furthermore, there is no data-association problem such as would be encountered when trying to match object candidates from different views in order to perform explicit triangulation.

2.5 Beamforming

In this work, we used a simple delay and sum beamformer implemented in the subband domain. Subband analysis and resynthesis was performed with a cosine modulated filter bank [9, §8]. In the complex subband domain, beamforming is equivalent to a simple inner product

$$y(\omega_k) = \mathbf{v}^H(\omega_k)\mathbf{X}(\omega_k)$$

where ω_k is the center frequency of the k^{th} subband, $\mathbf{X}(\omega_k)$ is the vector of subband inputs from all channels of the array, and $y(\omega_k)$ is the beamformed subband output. The speaker position comes into play through the *array manifold vector* [10, §2]

$$\mathbf{v}^H(\omega_k) = \left[e^{j\omega_k\tau_0(\mathbf{X})}\ e^{j\omega_k\tau_1(\mathbf{x})}\ \ldots\ e^{j\omega_k\tau_{N-1}(\mathbf{x})}\right]$$

where $\tau_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{m}_i\|/s$ is the propagation delay for the i-th microphone located at $\mathbf{m}_i$.

2.6 Language Model Training

To train *language models* (LM) for LM interpolation we used corpora consisting of broadcast news (160M words), *proceedings* (17M words) of conferences such as ICSLP, Eurospeech, ICASSP or ASRU and *talks* (60k words) by the Translanguage English Database. Our final LM was generated by interpolating a 3-gram LM based on broadcast news and proceedings, a class based 5-gram LM based on broadcast news and proceedings and a 3-gram LM based on the talks. The perplexity is 144 and the vocabulary contains 25,000 words plus multi-words and pronunciation variants.

2.7 Acoustic Model Training

As relatively little supervised data is available for acoustic modeling of the recordings the acoustic model has been trained on *Broadcast News* [11] and merged with the close talking channel of meeting corpora [12, 13] summing up to a total of 300 hours of training material.

The speech data was sampled at 16kHz. Speech frames were calculated using a 10 ms Hamming window. For each frame, 13 *Mel-Minimum Variance Distortionless Response* (Mel-MVDR) cepstral coefficients were obtained through a discrete cosine transform from the Mel-MVDR spectral envelope [14]. Thereafter, linear discriminant analysis was used to reduce the utterance based cepstral

mean normalized features plus 7 adjacent to a final feature number of 42. Our baseline model consisted of 300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks.

2.8 Acoustic Adaptation: Close Talk Speech

The adaptation of the close talking acoustic model was done in consecutive steps:

1. A supervised Viterbi training of the CHIL adaptation speakers followed by a *maximum a posteriori* (MAP) combination of this model with the acoustic model of the original system: To find the best mixing weight, a grid search over different mixing weights was performed. The weight, which reached the best likelihood on the hypotheses of the first pass of the unadapted speech recognition system, was chosen as the final mixing weight.
2. A supervised *maximum likelihood linear regression* (MLLR) in combination with *feature space adaptation* (FSA) and *vocal track length normalization* (VTLN) on the close talking CHIL development set: This step adapts to the speaking style of the lecturer and the channel. In the case of non-native speakers the adaptation should also help to cover some 'non nativeness'.
3. A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

2.9 Acoustic Adaptation: Far Distance Speech

The adaptation of the far distance acoustic model was done in consecutive steps:

1. Four iterations of Viterbi training on far distance data from NIST [15] and ICSI [16] over all channels on top of the acoustic trained models to better adjust the acoustic models to far distance.
2. A supervised MLLR in combination with FSA and VTLN on the far distance (single distance or MA processed) CHIL development set: This step adapts to the speaking style of the lecturer and the channel (in particular to the room reverberation). In the case of non-native speakers the adaptation should also help to cover some non-native speech.
3. A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

3 Experiments

In order to evaluate the performance of the described system, we ran experiments on recordings as described before on five seminars/speakers providing a total of approximately 130 minutes speech material with 16.395 words.

Tracking mode	Average error (cm)	
	all frames	*speech frames*
Audio only	46.1	41.7
Video only	36.3	36.5
Video & Audio	30.5	29.1

Table 1. Averaged error in 3D head position estimation of a lecturer over all frames (approximately 130 Minutes) and frames where speech was present (approximately 105 Minutes).

3.1 Source Localization

The error measure used for source localization is the average Euclidean distance between the hypothesized head coordinates and the labeled ones.

It can be seen, Table 1, that even though the video only tracker performs considerably better than the audio only tracker, the performance can still be significantly increased by combining both modalities. This effect is particularly distinctive during one recording in which the lecturer is standing most of the time in one dark corner of the room, thus being hard to find using solely video features (116cm mean error). While the video only tracker has the same performance for all frames and speech only frames, the precision of the audio only and the combined tracker is higher for the frames where speech is present compared to the precision over all frames.

3.2 Speech Recognition

The speech recognition experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which was developed and is maintained jointly by the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, USA. All tests used the language and acoustic models described above for decoding.

Tracking mode	WER
Close Talking Microphone	34.0%
Microphone Array	
single microhone	66.5%
estimated position (Audio only)	59.8%
estimated position (Video only)	59.1%
estimated position (Audio & Video)	58.4%
labeled position	55.8%

Table 2. *Word error rates* (WER)s for a close talking microphone and a single microphone of the array and the microphone array with different position estimates.

As mentioned before the big advantage of the MA is the big gain in WER over a single channel, compare the numbers in Table 2. Infact using a MA with an estimated speaker position over a single far distance channel we gain back 26.9% of the accuracy compared to the CTM.

3.3 Source Localization vs. Speech Recognition

Figure 3 compares the average position error of the source localization to the WER. If the error of the labeled position to the ground truth is around 15 cm (our calculatinon of the accuracy is approximately 10 cm), then a linear relationship can be seen.

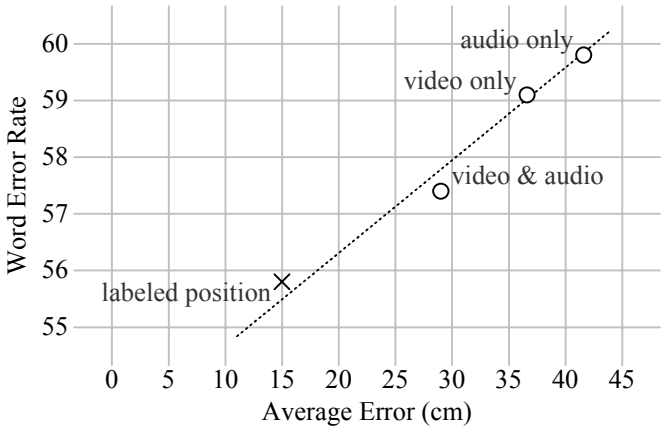

Fig. 3. Plot comparing the average position error to the word error rate.

4 Conclusions

We have compared the WER on different approached for person tracking using multiple cameras and multiple pairs of microphones. The core of the tracking algorithm is a particle filter that works with estimating the 3D location by sampled projection, thus benefiting from each single view and microphone pair. The video features used for tracking are based on adaptive foreground segmentation and the response of detectors for upper body, frontal face and profile face. The audio features are based on the TDOA between pairs of microphones, and are estimated with a PHAT function.

The tracker using audio and video input clearly outperforms both the audio- and video-only tracker on the accuracy of the estimate resulting in a decrease of WER. One reason for this is that the video and audio features described in this paper complement one another well: the comparatively coarse foreground feature along with the audio feature guide the way for the face detector, which in turn gives very precise results as long as it searches around the true head position.

Another reason for the benefit of the combination is that neither motion and face detection nor acoustic source localization responds exclusively to the lecturer and not to people from the audience – so the combination of both increases the chance of actually tracking the lecturer and therefore a decrease in WER.

In the future we want to use advanced techniques such as cepstral domain maximum likelihood beamformer [17] for the MA. For the fusion weight we want to define a criteria which depends on voice activity detection to give more weight to the audio in the case of speech and vice versa.

References

1. U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *Proc. Eurospeech*, 2005.
2. T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.
3. K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audiovisual speaker tracking," *7th Intl. Conference on Multimodal Interfaces*, 2005.
4. H. Steusslof, A. Waibel, and R. Stiefelhagen, "Computers in the human interaction loop," *http://chil.server.de*.
5. D. Focken and R. Stiefelhagen, "Towards vision-based 3-d people tracking in a smart room," *IEEE Int. Conf. Multimodal Interfaces*, 2002.
6. M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. ICASSP*, vol. II, pp. 273–6, 1994.
7. J. Chen, J. Benesty, and Y. A. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, pp. 549–57, November 2003.
8. M. Isard and A. Blake, "Condensation–conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.
9. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs: Prentice Hall, 1993.
10. H. L. Van Trees, *Optimum Array Processing*. New York: Wiley-Interscience, 2002.
11. Linguistic Data Consortium (LDC), "English broadcast news speech (Hub-4)," www.ldc.upenn.edu/Catalog/ LDC97S44.html.
12. F. Metze, C. Fügen, Y. Pan, T. Schultz, and H. Yu, "The ISL rt-04s meeting transcription system," *in Proc. ICASSP-2004 Meeting RecognitionWorkshop. Montreal; Canada: NIST*, 2004.
13. S. Burger, V. Maclaren, and H. Yu, "The isl meeting corpus: The impact of meeting type on speech style," *ICSLP*, 2002.
14. M. Wölfel, J. McDonough, and A. Waibel, "Warping and scaling of the minimum variance distortionless response," *ASRU*, 2003.
15. V. Stanford, C. Rochet, M. Michel, and J. Garofolo, "Beyond close-talk - issues in distant speech acquisition, conditioning classification, and recognition," *ICASSP 2004 Meeting Recognition Workshop*, 2004.
16. A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The icsi meeting project: Resources and research," *ICASSP 2004 Meeting Recognition Workshop*, 2004.
17. D. Raub, J. McDonough, and M. Wölfel, "A cepstral domain maximum likelihood beamformer for speech recognition," *ICSLP*, 2004.