FP6-506811

AMI

Augmented Multiparty Interaction

Integrated Project
Information Society Technologies

D4.6 State-of-the-art reports emerging from WP4

Due date: 31/12/2006 **Submission date:** 11/02/2007
Project start date: 1/1/2004 **Duration:** 36 months
LEAD CONTRACTOR: TUM **Revision:** 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

AMI

AUGMENTED MULTI-PARTY
INTERACTION

D4.6 State-of-the-art reports emerging from WP4

Abstract:

This document, which updates previous AMI deliverables D4.3, 4.4, and 4.5, describes the state-of-the-art in AMI areas related to audio and vido processing. The fields covered are conversational speech recognition with remote microphones, localization and tracking, and recognizing attentional cues.

Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

State-of-the-art overview

Conversational multi-party speech recognition using remote microphones

1 Introduction

The focus in large vocabulary automatic speech recognition research has been devoted to the transcription of speech found in natural environments for quite some time. The recorded speech is rarely planned but spontaneous or even conversational which contributes to relatively poor performance on these tasks. More recently more attention was devoted to the automatic transcription of conference room meetings. This interest is partly driven by the direct demand for transcripts of meetings. Moreover these transcripts can form the basis for higher level processing such as content analysis, summarisation, analysis of dialogue structure etc. This increased interest is manifest in yearly evaluations of speech recognition systems by the U.S. Institute for Standards and Technology (NIST) (e.g. [1]) or the existence of large scale projects such as AMI or CHIL [1]. Initial work on meeting transcription was facilitated by the collection of the ICSI meeting corpus and the NIST meeting transcription evaluations in 2002. Further meeting resources were made available by NIST [24] and Interactive System Labs (ISL) prior to the 2004 NIST RT04s Meeting evaluations[1]. In these evaluations the meeting domain was considered to cover so-called conference style meetings only, i.e. meetings with participants sitting around a meeting table. In the RT05s evaluations lecture-style meetings were added. Here a person presents material and answers questions from the audience.

As work in this domain is new many questions relating to fundamental properties of the data are yet unanswered. It is evident that the data varies greatly with the acoustic environment, the recording conditions and the content. A variety of recording configurations using speaker associated or remote microphones poses additional challenges. Overlapped speech or reverberation in the meeting room are a further cause degradation in recognition performance. This is especially present in lecture type scenarios where rooms may be large and the distance of the speaker to recording equipment is far greater than in conferences rooms.

1.1 Remote microphone recordings

Meetings typically take place in rooms with non-ideal acoustic conditions in the presence of significant background noise, and may contain large sections of overlapping speech. In such circumstances, headset microphones have, to date, provided the best recognition performance, however they have a number of disadvantages in terms of cost and ease of use. The alternative is to acquire the speech from one or more distant microphones, however, such ‘remote microphone recordings’ generally result in reduced ASR performance. A large body of research is concerned with techniques to enhance recordings from distant microphones with the goal of improving ASR performance. We describe a number of these techniques and subdivide them based on the amount of prior knowledge we have about the microphones.

1.2 Structure

This overview of state of the art in conversational multi-party speech recognition using remote microphones has a strong focus on developments made in the AMI and M4 projects [2] and hence does not

[1] Computers in the Human Interaction Loop, an EC IP project

[2] At the time of writing the AMI project has participated in the NIST evaluation in 2005 and 2006 [26, 28].

claim completeness in all associated areas. However the authors aim to show the range of topics and techniques as well as highlight the issues that are important for work in this domain.

In the rest of the paper we discuss the state-of-the-art for front-end processing, acoustic and language modelling as well as general system architecture. Automatic speech recognition systems are very complex and consist of many components to achieve competitive performance. Many of the methods used for meeting transcription are generic or work on general conversational speech. Hence the focus of the remaining sections is on meeting specific components and thus naturally front-ends are discussed in greater detail.

2 Front-end Processing

On of the main problems in this domain is the robust acquisition of the speech signal given the adverse conditions (in terms of ASR performance) in which most meetings are held. Meeting rooms are often reverberant (e.g., the instrumented meeting room at the University of Edinburgh has a reverberation time in the region of 0.7s); they suffer from significant background noise (e.g., from projectors and computers within the room) and activities outside the room; and meetings often contain periods in which several people are speaking concurrently. Close-talking microphones alleviate many of these problems and give the highest accuracy from current ASR systems, however they have a number of practical limitations concerning their use. Advanced processing techniques for multiple distant microphones, such as microphone array processing, offer an increasingly viable alternative which overcome many of the disadvantages of close talking microphones. In this section we first describe the problems associated with the capture of speech in meetings, even when using close talking microphones. We then describe the practical limitations of headset microphones, and present a number of distant microphone systems which can overcome these limitations.

2.1 Close talking microphones

Recordings made using close talking microphones have the advantage of high signal-to-noise ratio and implicit knowledge of the number of speakers as a single speaker is associated with each channel. Despite these advantages, there still remain significant challenges when carrying out ASR on these recordings in realistic environments, such as are encountered in meeting room scenarios. These challenges are often not dissimilar to that encountered in the far field.

The most serious problems typically encountered are the presence of cross-talk and the poor reliability of speech end-point detection. Cross-talk occurs when speech from neighbouring individuals is captured. While this is predominantly a problem for lapel based-recordings, it can also occur with head mounted microphones. When cross-talk occurs in the absence of speech activity from the target speaker, the effect can be reliably suppressed by a comparison between channels using cross-correlation and/or energy based analysis [46, 60, 36]. A more complex situation arises when cross-talk is overlapping with target speaker activity. Detection of such cases is possible through the use of extended statistics [60], but this does not deal with the fundamental problem that overlapping segments are likely to result in lower speech recognition performance. Speech end-point detection is a trivial problem in ideal recording conditions, but in meeting room recordings this becomes a challenging task, also because of the great variability in recording conditions and the presence of high-energy, non-speech sounds in the recording. These sounds are often produced by the target speaker (the most prevalent source of such noise is breathing onto a headset microphone worn too close to the mouth). Previous work in this domain has looked at statistical approaches for speech activity detection using HMM/GMM based classifiers with additional components to control cross-talk between channels [46, 60, 36].

Partners in the AMI project have undertaken similar approaches, demonstrating significant performance improvements over previous efforts [9, 20]. In comparing results between the AMI system submissions for the NIST Rich Transcription 2005 and 2006 Spring evaluations, the increase in WER due to automatic speech segmentation was reduced from 6.4% (20.9% relative) to 3.1% (12.8 % relative) [26, 28]. Thus, while there is still room for further improvement, a large component of the problem has been addressed in the AMI project.

Aside from the issues concerning ASR on close talking microphones, there are also more practical limitations that arise in realistic meeting scenarios it is impractical to provide every participant in a meeting with a headset microphone since the cost of such devices is prohibitive. Participants also find them obtrusive and feel self-conscious wearing them, and unless radio microphones are used, participants are effectively tethered to one location, unable to act or move naturally. The multiple distant microphone processing techniques described below address these problems since they remove the need for individual participant microphones.

2.2 Microphone arrays

Microphone arrays offer a principled approach to recovering a particular person's speech from a mixture of distant microphone signals [45]. A microphone array consists of multiple omni-directional microphones arranged in purposeful geometries in a room. Microphone arrays filter the received signals according to the spatial configuration of speech sources, noise sources and microphones, and are thus able to focus on sound originating from a particular location. The capabilities of such microphone arrays include location of sources in reverberant enclosures, identification and separation of the sources, enhancement of speech signals from desired sources, and separation of speech from non-speech audio signals [55]. A body of previous work, e.g. [45, 42], has shown that arrays can be an effective alternative to close-talking microphones for single speaker ASR in noisy environments. In addition, in a multi-speaker environment, the directional nature of the array allows discrimination between speakers leading to improved ASR performance for overlapping speech [44].

Microphone array speech enhancement generally involves *beamforming*, which consists of filtering and combining the individual microphone signals in such a way as to enhance signals coming from a particular location. The simplest beamforming technique is delay-sum beamforming, in which a delay filter is applied to each microphone channel before summing them to give a single enhanced output channel. Each channel delay (with respect to some reference channel) is calculated to align the speech signal arriving from a particular source location, ensuring constructive in-phase addition of the desired signal during the summation. As the noise components in the signal are combined out of phase, this procedure leads to a relative increase in the signal level (i.e. speech from the desired direction) with respect to the noise level.

Other more sophisticated beamforming techniques exist which calculate the channel filters to optimise a particular criterion - such as gain with respect to an isotropic noise field or a set of particular noise locations. These techniques can be broadly categorised as being fixed (data-independent) or adaptive (data-dependent) beamformers. In general, fixed beamformers have the advantage of providing less distortion to the desired speech signal, while adaptive beamformers tend to yield greater reduction of the noise level. In the robust speech recognition literature the most commonly used fixed beamforming techniques are delay-sum and superdirective beamforming [17, 16], while adaptive techniques have generally been variations of the Generalised Sidelobe Canceller (GSC) [25].

In practise, the beamformer seldom exhibits the level of improvement that the theory promises and further enhancement is desirable. One method of improving the system performance is to add a post-filter to the output of the beamformer. The use of a post-filter has been shown to improve the broadband noise reduction of the array [56], and lead to better performance in speech recognition applications [42]. Most approaches are based upon the post-filter proposed by Zelinski [62], which uses the input channel auto- and cross-spectral densities to estimate a Wiener post-filter to be applied to the beamformer output. The use of such a post-filter with a standard sub-array beamforming microphone array was thoroughly investigated by Marro et al [39], and has been used successfully in a number of speech enhancement and robust speech recognition applications. Other post-filter formulations better suited to more complex diffuse or non-stationary noise environments have been proposed in e.g. [40, 15].

With the increasing interest in using microphone arrays for speech recognition, an emerging research direction has been closer integration of the beamforming stage with statistical speech models. The motivation behind such approaches is the fact that traditional array processing is formulated to maximise the signal-to-noise ratio (SNR), rather than necessarily minimise the error rate of the eventual speech recognition. In fact traditional microphone array processing techniques enhance the signal based purely on geometrical information rather than any knowledge of the speech spectrum. Recent techniques that attempt to incorporate some form of speech model in the enhancement include, e.g., a likelihood-maximising beamformer (LIMABEAM) [53], a range of new speech-specific source separation algorithms [49, 52], and a beamformer based on a dual excitation speech model [10].

All of the above techniques assume the location of the desired speaker is known. In some situations, such as known seating configurations around a table, this assumption may be realistic. More generally, however, beamforming should be preceded by a step that locates (and potentially tracks) each speaker, e.g. [19]. Recent research has started to investigate the integration of speaker tracking with beamforming for speech recognition [58, 41, 7].

2.3 Table-top microphones

The simplest alternative to close-talking microphones is to use individual omnidirectional microphones located on the meeting table, each in front of one or more participants. Although table-top microphones remove the need for individual microphones, their performance for ASR is significantly worse, primarily as a result of the decay of sound energy with distance. As described above, close-talking microphones capture speech from the wearer at a higher level than other sound sources from the environment (other speakers, background noise sources). For distant microphones however, the differences in distance travelled from each source to the microphone are not as substantial. The received signal contains a variable mixture of all sources, and background noise, room reverberation and crosstalk also severely effect the quality of the received signal. Recognition experiments carried out on the ICSI meeting corpus [22] have shown that the word error rates (WER) for individual table-top microphones were double those of the close-talking microphones.

The performance of table-top microphones can be improved by employing well known noise reduction (e.g. those using Wiener filtering) and echo cancellation techniques (e.g. those using adaptive filtering) that attempt to recover the original speech from the noisy signal. In addition, if multiple table microphones are available, then the beamforming techniques described above may be used to enhance the output and perform localisation of speakers, even if the microphone locations are unknown. Such techniques have been widely used in speech recognition systems developed for recent NIST evaluations [1, 2]. For example, the following processing steps were used for multiple distant microphone processing in the AMI system [27, **?**]:

- First, gain calibration was performed by normalising the maximum amplitude level of each of the individual microphone channels

- Wiener filter was applied to each channel to remove the stationary background noise
- Delay vectors between each channel pair were calculated for every frame using the normalised cross correlation between channels
- Relative scaling vectors were measured corresponding to the ratio of frame energies between each channel and the reference channel
- The delay and scaling vectors were then used to calculate beam-forming filters for each frame using the standard super-directive technique
- The beamformed output was used as input to the ASR system

The above processing steps significantly improved the recognition performance and reduced the gap between close-talking microphones and table-top microphones. In 2005 the AMI system achieved word error rates of 30.6% for close talking microphones and 42.0% for the above system in the rt05s evaluation.

Table top arrays of microphones currently provide the best compromise between ASR performance and ease of use, since they do not require dedicated calibrated arrays within the room. [21] also shows that table-top arrays consisting of inexpensive conventional electret microphones can achieve similar recognition results to that of a single expensive sensor and as such, these arrays provide a cost effective alternative to calibrated arrays.

In NIST evaluations (e.g. [3]) recordings from many different meeting rooms are used and each room has its own configuration specifics in terms of number of microphones, their location in relation to the speakers, the room geometry etc. However, so far no particpant has worked with the specific room geometry. In the case of very low number of microphones and wide spacing none of the above techniques was found to be robust and simple energy based selection of microphones proved to be more efficient and the performance gap between close talking and table-top microphone array recordings could be narrowed substantially[28].

2.4 Dynamic microphone networks

While microphone array techniques, including those based on uncalibrated arrays of table top sensors provide enhanced output compared to the output of individual distant microphones, they have several strong requirements constraining their application: they generally assume a fixed number of microphones, strictly simultaneous sampling between channels, calibrated microphone gain levels, and a known, static microphone placement. Such stringent requirements cannot be guaranteed in most practical situations.

With the increasing prevalence of networked devices containing microphones an alternative to fixed arrays, based on the concept of distributed sensor networks [14, 4] is becoming available. So called ’dynamic’ or ’ad-hoc’ microphone networks comprise a group of individual devices such as PDAs or mobile telephones which, communicating via wireless network, act as elements in a microphone array. Requiring little or no pre-installed infrastructure and capable of using readily available sensors, such a system would allow high quality speech acquisition for ASR from groups of people at low cost, without the need for close talking microphones.

Such systems present a number of challenges compared to fixed arrays many of which are currently being addressed. Strict synchronisation between channels cannot be guaranteed in ad-hoc arrays. This is being addressed using a number of techniques based on the transmission of a global clock signal to each device [38, 34, 8]. The location of the microphones must be determined automatically in the case of ad-hoc arrays. Work on such 'self locating' microphone arrays has recently been reported in the literature [57, 48, 51, 50, 13], however these algorithms present some limitations, such as requiring a calibration signal to be played, or that close initial estimates of the microphone locations be provided. Differing devices will also have variable channel gains and this will also need to be addressed for such arrays to be used effectively as elements in an array.

While research on ad-hoc arrays is still in its infancy, and a fully functional audio acquisition system providing beamformed output from a dynamic array is still some way off, such a system has clear benefits over a conventional array.

2.5 Speaker Diarisation

Diarisation is the task to find out *who spoke when* in a multiple speaker scenario. This relates to a combination of speech activity detection and speaker clustering where the relevance of the former was outlined above. Speaker clustering is important for systems that adapt to speakers. Diarisation is a difficult task requiring complex systems for optimal performance (e.g. [6]). Experience reported at meeting workshops [1, 2] so far indicates that optimisation of diarisation criteria does not coincide with optimal ASR performance.

In 2006 NIST introduced the scoring of overlapped speech, i.e. words spoken by multiple speakers at the same time. Hence, ideally a diarisation system is capable of handling such overlap. In the upcoming RT'07 evaluations a joint speech recognition/diarisation performance will be measured. This new metric is called “speaker attributed word error rate” and will count correctly recognised words as wrong if the associated speaker label is incorrect.

2.6 Feature extraction

Most techniques mentioned above are enhancement based, i.e. the objective is to improve the audio quality prior to recognition. This has the advantage that later stages in the speech recognition process are allowed to operate in a standard way. Hence systems in meeting transcription make use of standard features such as Mel Frequency Cepstral Coefficients (MFCC) [18] or Perceptual Linear Prediction (PLP) coefficients [32] or derivatives thereof (e.g. [27, **?**, 43]).

Recently there is increased interest in feature space representations that cover a long time span. Many systems now make use of so called posterior based augmentations of the above feature vector. The AMI 2006 system for example includes features based on phone state posterior probability as computed by an MLP[**?**]. The LCRC features are derived from Mel frequency log filterbank (FB) coefficients where 23 FB coefficients are extracted every 10ms. 15 vectors of left context are then used to find the LC state level phone posterior estimates. The same procedure is performed with the right context. These posteriors are then combined with a third MLP network and after logarithmic compression the 135D feature vector is reduced to dimension 70 using principal component analysis. Final dimensionality reduction using heteroscedastic linear discriminant analysis (HLDA)[35] to 25 feature components is performed and the vector is appended to the standard 39D vector. These techniques work well for both close talking and far field microphone processing [**?**].

3 Acoustic Modelling

Acoustic modelling techniques proposed for meeting transcription in general do not differ greatly from general acoustic modelling techniques used for transcription of conversational telephone speech. An important reason for this is the enhancement based front-end that try to eliminate the additional acoustic variability.

3.1 Resources

As is normal for large vocabulary ASR, in-domain training data is vital for good performance. By now several corpora of meeting recordings are available amounting to between 150 to 200 hours of speech.

The ICSI Meeting corpus [33] was originally the largest meeting resource available consisting of 70 technical meetings at ICSI with a total of 73 hours of speech. The number of participants is variable and data is recorded from head-mounted and a total of four table-top microphones. Further meeting corpora were collected by NIST [24] and ISL [12], with 13 and 10 hours respectively.Both NIST and ISL meetings have free content (e.g. people playing games or discussing sales issues) and number of participants. As part of the AMI project a major collection and annotation effort of the AMI meeting corpus[5] was undertaken and has finished in June 2006. Data was collected from three different model meeting rooms in Europe (mostly Edinburgh and IDIAP at the moment). Overall more than 100 hours of transcribed speech are now available for free download. The meeting language is English. Each meeting normally has four participants and the corpus is split into a *scenario* portion and individual meetings. The scenario portion involves the same participants over multiple meetings on one specific task. Further small sets of meeting recordings for testing purpose have been made available in the context of NIST evaluations.

3.2 Model training procedures

Overall the amount of data available from meeting recordings is minimal compared to other domains such as Broadcast News (BN) or conversational telephone speech (CTS) where multiple 1000s of hours of speech are now transcribed. Hence it is not surprising that system developers decided to make use of these background resources. In systems such as [54, 59, 27] the comparatively large Switchboard corpora as well as CallHome Corpus where used[3] to bootstrap or train models for meeting transcription, on the basis that the target is conversational speech. However, these resources are recorded over the telephone and hence have different bandwith to that usually available in meetings. The problem was addressed by downsampling in the case of [54, 59, **?**] while in [27] a adaptation technique was used to map between different bandwidths. In [43] instead the use of BN data was suggested which was verified in [**?**]. In both cases the use of the additional background material allowed substantial improvement in word error rates.

For training on multiple remote microphone data again different strategies have been developed. In [59] training on all microphone channel recordings simultaneously was found to outperform training on single channels, e.g. by picking the central microphone or prior enhancement (e.g. as in [27]). The AMI 2006 system [**?**] has expanded this technique. When using a SAT style training on each microphone channel (CHAT), i.e. one set of CMLLR transforms per channel, a performance gain of 1% WER absolute can be observed [29].

[3] Available from the Linguistic Data Consortium

Apart from these data issues standard acoustic models are based on decision tree state-clustered Hidden Markov Models (e.g. [61]) or equivalent forms. Maximum likelihood training schemes are generally replaced by discriminative training using discriminative criteria such as the minimum phone error(MPE) criterion [47]. Front-end feature transforms such as heteroscedastic linear discriminant analysis [35] allow a more effective construction of feature spaces while speaker adaptive training techniques such as vocal tract length normalisation (VTLN) show similar performance gains to those obtained with the same techniques on CTS data[30] (and in contrast to performance on BN). Purely test-adaptive techniques such as maximum likelihood linear regression [37, 23] are used mostly for adapting to speakers rather than the environment.

4 Language Modelling and Vocabulary

Similarly to acoustic modelling in-domain data availability is a major issue in language modelling and vocabulary selection. Vocabularies are normally selected by using the most frequent in-domain words, and if necessary, augmenting the list with the most frequent words from other sources, for example BN text corpora. Even though meetings can be held on a wide range of topics the approach appears to yield sufficient coverage [31].

Language model training data for conversational speech is sparse. Hence models are constructed from other sources such as BN data and interpolated. This is true for both CTS and meeting data. Hence most systems use interpolated language models from a variety of sources, including data collected from the web [11] specifically for the task [**?**, 27].

The use of web-datafor building domain specific language models (LMs) has proven highly effective (e.g. [31]). Such data is collected by querying search engines with *n*-grams representative of the target domain. The choice of queries has a significant impact on how well the retrieved web-data matches the domain. Traditionally, queries were deemed representative of the target domain solely by examining the *n*-gram counts of a sample of in-domain text (T). In more recent work[**?**] on search models the queries were selected by selecting the most frequent *n*-grams that occurred in the target domain but did not occur in the background data (B).

5 Speech decoding

A major component in the development of any speech recognition system is the decoder. As task complexities and, consequently, system complexities have continued to increase the decoding problem has become an increasingly significant component in the overall speech recognition system development effort, with efficient decoder design contributing to significantly improve the trade-off between decoding time and search errors. One approach that has seen considerable interest in recent years is the Weighted Finite State Transducer (WFST) based decoder. Pioneered by Mohri and others at AT&T [**?**], the key advantage behind the use of WFSTs for speech decoding is that it enables the integration and optimisation of all knowledge sources within the same generic representation. While the use of static networks in speech decoding is far from being a new idea, the explicit use of weighted finite-state transducers is relatively recent, providing a more efficient framework for carrying out speech recognition and also enabling simpler decoder design and greater flexibility in the integration of new knowledge sources in various stages of the system hierarchy.

6 System architectures

Systems for meeting transcription are not yet in wide-spread use. At this stage the system architectures mostly follow patterns that have been developed elsewhere (e.g. in CTS transcriptions). State-of-the-art systems operate in multiple passes where each pass normally outputs both a first-best results and a word-graph. The latter is often used to constrain the search space for subsequent stages, or more recently, to allow for output combination of complementary systems, i.e. systems that are trained to yield similar performance with different error types. Depending on the allowance in terms of real time the system complexity is usually increased by the number of passes, with decreasing gains in word error rate in later passes. So far only few experiments are published that try match system architecture to the type of input data (i.e. dependent on the microphone configuration). In the case of multiple remote microphone data experiments with recognition on each microphone channel have not yielded superior performance. In recent NIST evaluations the practical benefit of a unified system structure regardless of the input data was noted by all participants. Integration with other systems, such as those for diarisation or source localisation systems was not yet shown to yield clear advantages.

7 Conclusions

In this paper we tried to give a brief overview of state-of-the-art in speech recognition of conversational speech with remote microphones. Although the discussion is clearly incomplete at this stage we highlighted properties and main issues of current systems and discussed several different approaches to fundamental problems. From the results so far (word error rates of 20–35% and still a large difference to results with close-talking data) and the fact that current systems have not yet touched on the full complexity of the domain, we infer that there is room for substantial improvement. Front-ends that are on the one side fully integrated into ASR systems, but are at the same time aware of the physical conditions, are likely to yield substantial improvements. Distributed system architectures will allow effective implementation of such systems. Several important questions are only just being addressed, for example the issue of time-overlapped speech, or effects of reverberation.

References

[1] Spring 2004 (RT04S) rich transcription meeting recognition evaluation plan. 2004.

[2] Spring 2005 (RT05S) rich transcription meeting recognition evaluation plan. 2005.

[3] Spring 2006 (RT06S) rich transcription meeting recognition evaluation plan. 2005.

[4] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks, 2002.

[5] J. C. andS. Ashby, S. Bourban, M. G. M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma. The AMI meeting corpus. In *Proc. MLMI'05*, 2005.

[6] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo. Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 diarization system. In *Proc. NIST MLMI Meeting Recognition Workshop*, 2005.

[7] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh. Detection and separation of speech event using audio and video information fusion. *Journal of Applied Signal Processing*, 11:1727–1738, 2004.

[8] P. Bergamo, S. Asgari, H. Wang, D. Maniezzo, L. Yip, R. E. Hudson, K. Yao, and D. Estrin. Collaborative sensor networking towards real-time acoustical beamforming in free-space and limited reverberance. In *IEEE Transactions on Mobile Computing*, volume 3, pages 211–224, 2004.

[9] K. Boakye and A. Stolcke. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In *Proc. Interspeech (ICSLP)*, 2006.

[10] M. Brandstein and S. Griebel. Explicit speech modeling for microphone array speech acquisition. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 133–151. Springer, 2001.

[11] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc HLT'03*, 2003.

[12] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Proc. ICSLP'02*, 2002.

[13] J. Chen, R. Hudson, and K. Yao. Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. In *IEEE Transactions on Signal Processing*, volume 50, 2002.

[14] C.-Y. Chong and S. Kumar. Sensor networks: Evolution, opportunities, and challenges. In *Procedings of the IEEE*, volume 91, pages 1247–1256, 2003.

[15] I. Cohen and B. Berdugo. Microphone array post-filtering for non-stationary noise suppression. In *Proceedings of IEEE ICASSP*, 2002.

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

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

[18] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, 28(4):357–366, Aug. 1980.

[19] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 157–178. Springer, 2001.

[20] J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proc. Interspeech (ICSLP)*, 2006.

[21] L. Docio-Fernandez, D. Gelbart, and N. Morgan. Far-field asr on inexpensive microphones. In *Proc. of Eurospeech*, 2003.

[22] M. N. et al. Meetings about meetings: Research at ICSI on speech in multiparty conversations. In *Proc. of ICASSP*, 2003.

[23] M. J. F. Gales and P. C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, 1996.

[24] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi. The nist meeting room pilot corpus. In *Proc. LREC'04*, 2004.

[25] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30:27–34, January 1982.

[26] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proceedings of NIST Rich Transcription 2005 Spring Evaluation Workshop*, Edinburgh, UK, 2005.

[27] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proc. NIST MLMI Meeting Recognition Workshop*, 2005.

[28] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The ami meeting transcription system : Progress and performance. In *Proc. NIST RT'06 Workshop*, Springer LNCS, 2006.

[29] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, V. Wan, and J. Vepa. The AMI system for the transcription of speech in meetings. In *Proc. ICASSP 2007*, 2007.

[30] T. Hain, L. Burget, J. Dines, I. McCowan, G. Garau, M. Karafiat, M. Lincoln, D. Moore, V. Wan, R. Ordelman, and S. Renals. The development of the AMI system for the transcription of speech in meetings. In *Proc. MLMI'05*, 2005.

[31] T. Hain, J. Dines, G. Gaurau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: an investigation. In *Proc.Interspeech'05*, Lisbon, Portugal, 2005.

[32] H. Hermansky. Perceptual linear prediction (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, Apr. 1990.

[33] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi meeting corpus. In *Proc. ICASSP 2003*, 2003.

[34] I. Kozintsev, R. Lienhart, D. Budnikov, I. Chikalov, and S. Egorychev. Providing common i/o clock for wireless distributed platforms. In *Proc. ICASSP 2004*, volume 3, pages 909–912, 2004.

[35] N. Kumar. *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, 1997.

[36] K. Laskowski, Q. Jin, and T. Schultz. Crosscorrelation-based multispeaker speech activity detection. In *Proceedings of ICSLP*, Jeju Island, Korea, 2004.

[37] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9(2):171–186, 1995.

[38] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung. On the importance of exact synchronization for distributed audio signal processing. In *Proc. ICASSP 2003*, volume 4, pages 840–843, 2003.

[39] C. Marro, Y. Mahieux, and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259, May 1998.

[40] 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.

[41] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba. Speech acquisition in meetings with an audio-visual sensor array. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, July 2005.

[42] I. McCowan, C. Marro, and L. Mauuary. Robust speech recognition using nearfield superdirective beamforming with postfiltering. In *Proc. ICASSP 2000*, volume 3, pages 1723–1726, 2000.

[43] F. Metze, C. F?gen, Y. Pan, T. Schultz, and H. Yu. The isl rt-04s meeting transcription system. In *Proc. NIST Meeting Recognition Workshop*, 2004.

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

[45] M. Omologo, M. Matassoni, and P. Svaizer. Speech recognition with microphone arrays. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 331–353. Springer, 2001.

[46] T. Pfau, D. P. W. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. *Proceedings of ASRU*, 2001.

[47] D. Povey and P. C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *In Proc. ICASSP'02*, 2002.

[48] V. Raykar, I. Kozintsev, and R. Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. In *IEEE Transactions on Speech and Audio Processing*, volume 13, 2005.

[49] M. Reyes-Gomez, B. Raj, and D. Ellis. Multi-channel source separation by factorial HMMs. In *Proceedings of ICASSP-03*, volume 1, pages 664–667, April 2003.

[50] Y. Rockah and P. Schultheiss. Array shape calibration using sources in unknown locations - part i: Far-field sources. In *IEEE Transactions on Acoustics,Speech and Signal Processing*, volume 35, pages 286–299, 1987.

[51] Y. Rockah and P. Schultheiss. Array shape calibration using sources in unknown locations - part ii: Near-field sources and estimator implementation. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 35, pages 724–735, 1987.

[52] S. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proceedings of Eurospeech03*, pages 1009–1012, 2003.

[53] M. Seltzer, B. Raj, and R. Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12(5), September 2004.

[54] A. Stolcke, R. Gadde, A. Venkataraman, D. Vergyri, and J. Zheng. The SRI-RT02 Speech-To-Text System. In *Proc. NIST Rich Transcription Workshop*, 2002.

[55] K. Uwe-Simmer, J. Bitzer, and C. Marro. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[56] K. Uwe-Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, pages 39–57. Springer, 2001.

[57] A. Weiss and B. Friedlander. Array shape calibration using sources in unknown locations-a maxilmum-likelihood approach. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 37, pages 1958–1966, 1989.

[58] M. Wolfel, K. Nickel, and J. McDonough. Microphone array driven speech recognition: Influence of localization on the word error rate. In *Proceedings of MLMI*, May 2005.

[59] C. Wooters, N. Mirghafori, A. Stolcke, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, , and M. Ostendorf. The 2004 ICSI-SRI-UW meeting recognition system. 3361:196–208, January 2005.

[60] S. Wrigley, G. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91, January 2005.

[61] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. 1994 ARPA Human Language Technology Workshop*, pages 307–312. Morgan Kaufmann, 1994.

[62] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of ICASSP-88*, volume 5, pages 2578–2581, 1988.

Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

State-of-the-art overview

Recognition of Attentional Cues in Meetings

Updated version 23.01.2007

Information Society
Technologies

Sixth Framework Programme

1 Introduction

Attention refers to the cognitive process of selectively concentrating on one thing while ignoring other things. This is an everyday wording of a definition given by one of the first great psychologists, William James:

> Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought... It implies withdrawal from some things in order to deal effectively with others.
>
> (Principles of Psychology, 1890)

Attention is basic for all forms of perception; for outer perception via our senses, vision, hearing, taste and smell, as well as for inner perception, the perception of feelings, emotions, thinking. Attention processes allow us to direct attention to certain aspects in the environment with special care. An impressive example of the capacity to direct attention is the *cocktail party effect*: at a party, in the chaos of noise and voices, we are able to focus and concentrate on voices further away and thus follow a conversation, despite the other acoustic signals that continuously reach our ear. We are also especially good at selectively directing our attention to one or another aspect of an object, such as its color, shape or movement.

The amount of incoming information to the primate visual system is much greater than that which can be fully processed. Only part of this information is processed in full detail while the remainder is left relatively unprocessed. ([19]). There are two prime mechanisms in attention control: our sense system is constantly aware of its environment and changes ask for our immediate attention, second there is the process of directing our attention and selecting the information channel in order to better do what our task requires us to do. The first, bottom-up attentional selection process is a fast, and often compulsory, stimulus-driven mechanism. The other mechanism, top-down attentional selection, is a slower, goal-directed mechanism where the observer's expectations or intentions influence the allocation of attention. Observers can volitionally select regions of space or individual objects to attend. (see [19].)

Although attention is an innate activity of the mind, it shows itself in observable phenomena of body movements, facial expressions and the way people (or animals) act or behave. So it is tacitly assumed that eye gaze, body posture, arm and head movements are phenomena that go with changes in the attentional state of a person. This implies that these phenomena are possible *cues* pointing at the attentional state of a person or a change in his attentional state.

Attention is required for every form of sensual and cognitive perception, and thus for obtaining information. This is the very reason why *eye-gaze*, and head movements are *cues* for identifying the attentional state and focus of attention of someone. Focussing our attention to something or someone in our environment is done by directing our sense organs towards the object that we want to perceive.

Gaze direction is thus a natural sign for focussing attention to something in our visual environment. Laying our hand behind the ear and directing our head in the direction of a sound source is a natural sign of focussing our auditive attention to something that we hear. Becoming aware of this natural function, gaze becomes also a signal that is used to express attentiveness. Gaze patterns, as well as body postures thus become codes, by means of which the interactors show whether they are or are not attentive, whether they are willing to hear, or to interact. Backchannels ([32]) are codes used by hearers to show their attentiveness and signals for the speaker that he can go on. Deictic pointing is a way of referring to something and attracting someone's attention to that something that we want to highlight as being of special interest in a specific situation. Deictic pointing gestures are often accompanied with verbal gestures especially the use of referring expressions that help to identify the referred object. Gazing at something is the third way in which people try to focus other peoples attention.

Since our perceptions are mostly related to organized activities, which are being performed to achieve specific aims, the attentional cues do not occur unrelated. They show specific patterns. Identification of these patterns of attentional cues and how they correlate with specific activities is important for recognition of these activities from the observable cues.

The term *focus of attention* has different meanings in different research areas. The primary use of the term focus of attention refers to a perceptual variable indicating the object or person someone is attending to ([4]), a description of someone's focus of attention during an activity. At a semantical level the description of someone's focus of attention during an activity involves describing which actions, objects or people someone is attending to. At a syntactical level this could involve describing the spatial and temporal properties of someone's (visual) attention. As such this is not a directly observable category.

As people often orient themselves towards the physical objects or persons they are attending to, the notion of focus of attention has a derived meaning referring to the physically observable behavior of orientation towards an object by means of posture, head orientation and/or gaze. This could be called the *visual focus of attention*. Psychological attention and physically observable attention do not necessarily coincide but are correlated highly. This is a generally held assumption.

The term *focus of attention* is also used within a (computational) linguistic context. In a theory of discourse structure, developed by Grosz and Sidner, three components of discourse structure are distinguished: linguistic structure, intentional structure, and attentional state. The *attentional state* is considered as an abstraction of the discourse participants focus of attention. This state records the objects, properties, and relations that are salient at a given point in the discourse. (see [5])

Dialogic attention involves listening to a person (auditory mode of dialogic attention) or speaking to one or more persons (articulatory mode of dialogic attention). The focus of dialogic attention identifies these persons, the extent of dialogic attention describes the number of persons within this focus. ([28])

Research in focus of attention and in techniques for recognition of attentional cues has the interest of several research areas, including human interaction, conversational analyses, human machine interaction, user interfaces and attentive systems, embodied conversational agents and virtual environments.

2 Recognition of Attention in Meetings

Meetings are coordinated multi party activities where people collaborate in a shared task. Joint attention is required for effective collaboration. Most of the activities are conversational, participants verbally exchange ideas, have discussions. In order to understand what is going on in a meeting, to be able to write a summary of activities, discussions and decisions made, we need to understand the basic processes that underly these activities and also how these processes show in observable phenomena. Communication is a joint activity, the process of sharing ideas. In conversations a speaker produces meaningful signals by which he tries to direct the attention of the listeners mind to the ideas that he wants to communicate. The speaker selects those codes that he expects are most effective in directing the attention of the interlocutors in the way he wants.

By pointing at something people try to attract someone's attention to something. (spatio deictic referring). For example someone is looking in the direction of the window, then points at it and says:

"Look there is Socrates".

Several issues are involved:

a) that someone is making a movement with his arm, hand, that he moves his body to the one who's attention he tries to attract, his eye-gaze and head-movements, his speech, tone, and wording.

b) that this movement made is a pointing gesture, that he points at something to attract the attention of someone, who is in the same field of perception.

c) how this pointing gesture is related to other simultaneous activities by the same or other actors in the environment

d) how this pointing gesture fits in a sequence of activities over time by the same or by other actors or things in the environment

e) what the target is that the actor refers to

f) what the intention is that the actor has with pointing at this target

For automatic recognition of such a conversational gesture as deictic pointing we need computer vision technology as well as knowledge about the situation and the course in time of the conversational activities in which the act of pointing is embedded. Several modalities play a role in the process of recognition and interpretation.
Since the interpretation of the actions, hand and head movements made by participants in meetings can only be inferred when considering them in synchronized and parallel sequences of actions often Dynamic Bayesian Networks or (multi-layered) Hidden Markov Models are used for the prediction (classification) of the movements in terms of signals that have some particular pragmatic meaning in the context of the meeting.

People in meetings coordinate their actions in collaboration. Research has been performed to recognize meeting activities and segment meetings into sequences of meeting activities such as, note taking, discussion, presentation. These activities show in particular patterns of joint attention or *group focus of attention* and requires modeling simultaneous activities of all participants in a meeting. A meeting location has a number of distinguished locations of interest that participants refer to or focus their attention at in a meeting: white-board, laptop, documents, participants, someone entering the room, and in the AMI design project meetings the prototype of the remote control is a recurrent topic of interest. Head movements is one of the features (together with speech, linguistic and paralinguistic features, body postures and arm movements) that are used for recognition of sequences of meeting activities. (see [1], [7], [16], [21], [20],[33], [2].)

3 Gazing Behavior in Face to Face Communication

A speaker can use gaze to indicate that the party being gazed at is an addressee of his utterance. ([3]) Listeners gaze and direct themselves towards speakers in order to hear and see better what the speaker is saying. At the same time the speakers monitors who is listening and how his speech is received by the hearers. Gaze behavior is also of importance in turn-taking management. There are specific patterns in gaze behavior and hence in head movements related to conversational behavior. Identifying these patterns and recognizing this patterns is part of understanding what is going on in the meeting in terms of who is being addressed, who is trying to take the floor, who is anticipating floor changes, who is participating in a conversation.

Speakers gazing practices often demonstrate explicitly to co-participants that an initiating action is being directed to a particular party, thus selecting that party to speak next. This shows the gazed-at participant that he or she is the intended recipient, and it shows the participants not gazed at that they are not the intended recipient. For this method to work, then, an intended recipient must see the gaze. Others may also need to see it to grasp that someone (else) has been selected. (see:[15]). As has been shown already in the sixties and seventies, eye gaze serves an important role in guiding the conversation, both at the side of the speaker and the listener (e.g. [11]). At the speaker's side, looking at the listener may serve the function of monitoring the attention level and the processing status of the incoming speech, and help to regulate the flow of conversation. At the listener's side, looking at the speaker serves both the function of providing feedback for the speaker's monitoring activity, to inspect the speaker's behavior (facial expression, posture etc) for information about the speakers attitude and emotion, and to monitor for nonverbal cues for turn-taking. Most of these early findings concerning the use of nonverbal cues in communication are based on the analysis of dyadic conversations. Later research on triadic and multi party conversations has confirmed and extended the early findings. [31] provide evidence that gaze behavior is a reliable predictor of addressee-hood.

The main focus of [27] is the exploration of behavioral cues that could potentially be used for classification of the addressee of utterances in a situation where two users interact with each other and with a service system. Facial orientation, utterance length and reactions on system events are considered important cues for focus of attention. In order to evaluate the potential integration of these features, Naive Bayes classifier is used. Classes correspond with attentional states of the users. Both participants look at the system (class A), The speaker looks at the system but the partner looks at the speaker (B),The speaker looks at the partner but the partner looks at the system(C) and Both participants look at each other (D). Results show that facial orientation together with systems events (dialogical context) together are reliable features for predicting the attentional state of the interaction.

Recent work on addressee identification in four-participants face-to-face meetings presented in [9] have shown that using utterance, contextual and speaker's gaze features addressee can be predicted with an accuracy of 82.57%. In comparison, classification that used solely gaze directional cues achieved a significantly lower accuracy of 66.45%. This indicates that gaze directional cues when used alone are unreliable features for predicting addressee in meeting discussions. It was also found that listeners' gaze direction provides useful information for addressee identification only in the situation where gaze features are used alone. In all these cases, classification was performed by means of Bayesian Networks.

4 Research on Head Orientation and Gaze

Most existing work in detecting a user's visual focus of attention makes use of camera-based head pose tracking ([25], [24]) or eye tracking ([30]). The eye-based gaze detectors require a robust eye-tracker, and then typically extract gaze from the position of the pupils relative to the userŠs head. The head-based techniques estimate the focus of attention by determining the orientation of the userŠs head and assuming that the user is looking in the direction in which their head is pointing. These techniques can be accurate, but the fact that they are camera-based means that they are typically not mobile, and can encounter difficulties when the lighting of the scene changes. (see also: [17].)

Head position and eye gaze together and interactively determine whether an observer looking at a picture of a face judges whether the person on the pictures is looking at the viewer or next to him.

Gaze direction is constituted by head orientation and eye orientation ([13]) and can be used as a deictic signal, indicating the current focus of interest ([14]). We have reason to believe that we can use head orientations as a valid substitute for gaze when determining the focus of interest (c.f. [18]). In an experiment with a four-person setting, it was found that in 87.0 % of all cases, the participants rotated their heads and eyes in the same direction

Eye gaze is hard to monitor in a non-intrusive way. Therefore most systems that want to detect visual focus of attention of a person use head orientation. How good can we predict eye gaze from head orientation?

In 88.7% of the time, the focus of interest could be determined solely by the head orientation. Based on the fact that the head orientation component of gaze is so prevalent, it is expected that we see the same systematics that occur in gaze behavior when looking at head orientations alone. This is validated by analyzing a corpus consisting of head orientations and speaker data. ([24]).

How good are outside observers of meeting, i.e. people looking at a meeting through video, in telling what the focus of attention of a participants is? How precise are they in deciding where the head of the participants is directed to? The research described in [22] shows that differences in gaze behavior between speakers and listeners in a multi-party setting also exist when we look at their head orientations. By analyzing a corpus of four-person meetings it appeared that speakers are generally being looked at by more persons than listeners are. In an experiment, conducted in a virtual environment, it was found that observers apply these systematics when asked to identify the speaker when shown only the head orientations of the meeting participants. The virtual environment proved to be a suitable tool for research in perception of human behavior since it allows for good stimulus control.

In [24] an overview of work on tracking focus of attention in meeting situations is presented. A system has been developed that is capable of estimating participants focus of attention from multiple cues. The system employs an omni-directional camera to simultaneously track the faces of participants sitting around a meeting table and use neural networks to estimate their head poses. In addition, it uses microphones to detect who is speaking. The system predicts participants focus of attention from acoustic and visual information separately, and then combines the output of the audio- and video-based focus of attention predictors. The work reports recent experimental results. In order to determine how well we can predict a subject's focus of attention solely on the basis of his or her head orientation, we have conducted Experiment in which head and eye orientations of participants in a meeting were recorded using special tracking equipment show that head orientation was a sufficient indicator of the subject's focus target in 89% of the time. The paper also discusses how the neural networks used to estimate head orientation can be adapted to work in new locations and under new illumination conditions.

5 Annotating Focus of Attention

In order to train and evaluate the quality of attention recognition techniques hand annotated video corpora are made, in which the focus of attention of each of the participants is labelled continuously. A fixed list of possible targets is used to identify the focus. In annotating focus of attention we stick to the visual focus of attention of individuals, defined by the head orientation or eye gaze. So if someone is looking at a person but thinking about his upcoming holiday we will only label where he is or she is looking at.

For research in addressing behavior in face to face meetings focus of attention of participants in scenario based recorded meetings (collected in the M4 and AMI projects) was annotated. The coding is based on observations of the gaze and head turning of participants. The target set of interests for this research on addressing are meeting participants. Reliability of marking the changes in the gazed target (segmentation) was about 80%, and reliability of target labeling showed a kappa value of 0.95. Experiments with Bayesian Network models show that information about the focus of attention of participants, speakers as well as listeners contributes to the reliability of addressee prediction. ([8, 10])

Telling in what direction or what the participants are looking at is not only relevant for reliability of coding of the focus of attention in meeting behavior but also for technology mediated meeting participation, where remote participants rely on similar video and audio technology as annotators.

6 Applications

Modeling and tracking a person's focus of attention is useful for many applications: Intelligent supportive computer applications could use information about a userŠs focus of attention to infer the user's mental status, his/her goals and cognitive load and adjust their own responses to the user accordingly. For multi-modal human computer interaction, the user's focus of attention can be used to determine his/her message target. For example, in interactive intelligent rooms, focus of attention could be used to determine whether the user is to control the refrigerator, the TV set, or he/she is talking to another person in the room. ([23])

Recognition of the attentional state of communicating and collaborating agents is a requirement for attentive systems, which observe user activity to anticipate user needs as well as for Attentive User Interfaces, user interfaces that manage the user attention deciding when to interrupt the user, the kind of warnings, and the level of detail of the messages presented to the user ([29]).

Systems are build that integrate perceptual attention into multi-party, multiconversational dialogue layers [26]. A computational model of the dynamics of attentional state should model the perceptional aspects and the way the senses react to the perceived signals, as well as the activities, goals and intentions in which the actors are involved. In [12] a computational model of controlling the focus of perceptual attention for embodied agents is proposed. It provides the potential to support multi-party dialogues in a virtual world. It demonstrates that embodied agents can respond dynamically to events that are not even relevant to the tasks and shift their attention among objects in the environment.

Finally, Horvitz et al. present an overview of principles and methodologies in research on integrating models of attention into human-computer interaction. ([6]).

7 Conclusion

We provided a short, and necessarily incomplete, overview of research in the area related to the recognition of attentional cues in meetings. The analyses concerns (a) the behaviors of participants in face to face conversations, as well as in face to face meetings, related to various meeting activities (b) the analyses of this behavior by outside observers of meetings, notably annotators of meetings, and (c) the impact for remote meeting participation.

The recognition of attentional cues by machines requires recognition and tracking of human bodies and body parts, head movements, arm and hand positions, and thus builds on state of the art technology in computer vision. We do not give an overview of this research area.

We have only reviewed the main lines of research related to the issue of recognition of attentional cues. Insights gained by this research can help our understanding of what is going on in meetings. The study of meeting behavior, the roles that the various communication channels play in conversations is of prime importance for specifications of the requirements that technology mediated meetings should satisfy. The locations of participants in the meeting, the positioning of video and audio recorders, screens and sound boxes in meeting rooms, they all influence the way (remote) participants and outside observers perceive various modes of interactions that occur in the meeting. The identification of patterns of attentional behavior, head movements, gestures, eye gaze, can only be done after detailed and very careful observations of people in situations that are as realistic as possible. Conversational analysts have been doing this type of research already for many decennia. Often this research is based on observations by researchers who did real-time annotations. The multi-modal meeting corpus recorded in the AMI project makes it possible to gain more insight in what is going on in meetings. We have pointed at the relevance of the issues and research reviewed in this overview for the understanding of outside observers, annotators and remote participants, who observe and annotate meetings by means of video and audio technology.

Many issues arise that need further investigation. To name one issue: are perceived changes in the audio field a cue for changes in focus of attention of participants in the meeting? What are the consequences for the perception and experience of meeting participation of the reconstructed audio field for remote participants.

A lot of experiments have been performed to study the importance of gaze and mutual gaze in remote meetings, and the impact of the lack of a visual channel on conversations and meetings. Often these experiments are necessarily performed in controlled situations in order to be able to exclude influences that may interfere with the conditions under study. Making it often quite risky to infer results from this experiments to the real situations in which the knowledge should be applied. It is a great advantage that we now have the opportunity to study meetings in settings as realistic as possible in which the technology to support meetings is being used. This is by far the best way to get more insight in the impact of these technologies and to further the development of this technology and to get it tuned to the practice of meetings in the best possible way.

References

[1] Marc Al-Hames, Alfred Dieleman, Daniel Gatica-Perze, Stephan Reiter, Steve Renals, Gerhard Rigoll, and Dong Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proceedings MLMI'05, Edingburgh, Scotland*, 2005.

[2] A. Dielmann and S. Renals. Multistream dynamic Bayesian network for meeting segmentation. *Lecture Notes in Computer Science*, 3361:76–86, 2005.

[3] Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York, 1981.

[4] D. Gopher. *The Blackwell dictionary of Cognitive Psychology, chapter Attention*. Basil Blackwell Inc., 1990.

[5] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12:175–204, 1986.

[6] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication. from principles to applications. 2004.

[7] McCowan Ian, Daniel Gatica-Perez, Bengio Samy, Moore Darren, and Bourlard Herve. Towards computer understanding of human interactions. *Proceedings of EUSAI 2003, LNCS 2875, (E. Aarts et al. ed.)*, pages 235–251, 2003.

[8] N. Jovanovic and R. op den Akker. Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92, 2004.

[9] N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, 2006.

[10] N. Jovanovic, R. op den Akker, and N. Nijholt. A corpus for studying addressing behavior in face-to-face meetings. In *6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal*, 2005.

[11] Adam Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[12] Youngjun Kim, Randall W. Hill, and David R. Traum. Controlling the focus of perceptual attention in embodied conversational agents. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 1097–1098. ACM Press, New York, NY, USA, 2005.

[13] Chris L. Kleinke. Gaze and eye contact: a research review. *Psychological Bulletin*, 100(1):78–100, 1986.

[14] Stephen R.H. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology*, 53A(3):825–845, 2000.

[15] Gene H. Lerner. Selecting next speaker: the context-sensitive operation of a context-free organization. *Language in Society*, 32:177–201, 1998.

[16] I. McCowan, Gatica-Perez D., S. Bengio, and G. Lathoud. Automatic analysis of multimodal group actions in meetings. Technical Report RR. 03-27, IDIAP, Martigny, 2003.

[17] D. Merrill and T. Selker. The attentional mixer, internal tech report, context-aware computing group. Technical report, MIT Media Lab, 2004.

[18] Kazuhiro Otsuka, Yoshinao Takemae, Junji Yamato, and Hiroshi Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of International Conference on Multimodal Interface (ICMI'05)*, pages 191–198, Trento, Italy, 2005.

[19] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

[20] Stephan Reiter and Gerhard Rigoll. Multimodal meeting event recognition fusing three different types of recognition techiques. Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), Martigny, June 2004.

[21] Stephan Reiter and Gerhard Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[22] Rutger Rienks, Ronald Poppe, and Dirk Heylen. Differences in head orientation between speakers and listeners: experiments in a virtual environment. *IJHCS*, 2005.

[23] R. Stiefelhagen, J. Yang, and A Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, Vol.13, No. 4, 2002.

[24] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02)*, pages 273–280, Pittsburgh, PA, 2002.

[25] Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *Extended abstracts on Human factors in computing systems (CHI'02)*, pages 858–859, Minneapolis, MN, 2002.

[26] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings AAMASŠ02*, pages 15–19, 2002.

[27] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of ICMI*, 2005.

[28] R. Vertegaal. *Look who's talking to whom. Mediating Joint Attention in Multiparty Communication and Collaboration.* PhD thesis, University of Twente, 1998.

[29] R. Vertegaal. Attentive user interfaces. *Communications of the ACM*, Vol. 46(3):33–36, 2003.

[30] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Why conversational agents should catch the eye. In *Extended abstracts on Human factors in computing systems (CHI'00)*, pages 257–258, The Hague, The Netherlands, 2000.

[31] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the conference on Human factors in computing systems (CHI'02)*, pages 301–308, Seattle, WA, 2002.

[32] V. H. Yngve. On getting a word in edgewise. *Papers from the sixth regional meeting of the Chicago Linguistics Society, Chicago: Chicago Linguistics Society.*, 1970.

[33] Dong Zhang, Daniel Gatica-Perez, Samy Begio, Iain McCowan, and Guillaume Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*, Washington DC, July 2004.

FP6-506811

AMI

Augmented Multiparty Interaction

Integrated Project
Information Society Technologies

D4.6 State-of-the-art: localization and tracking of multiple interlocutors with multiple sensors

Due date: 31/12/2006 **Submission date:** 31/12/2006
Project start date: 1/1/2004 **Duration:** 36 months
LEAD CONTRACTOR: TUM **Revision:** 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
	Dissemination Level	
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

AMI

AUGMENTED MULTI-PARTY
INTERACTION

D4.6 State-of-the-art: localization and tracking of multiple interlocutors with multiple sensors

1 Introduction

The automatic analysis of meetings recorded in multi-sensor rooms is an emerging research field in various domains, including audio and speech processing, computer vision, human-computer interaction, and information retrieval [45, 86, 64, 81, 21, 61, 92]. Analyzing meetings poses a diversity of technical challenges, and opens doors to a number of relevant applications, including automatic structuring and indexing of meeting collections, and facilitation of remote meetings.

In the context of meetings, detecting, localizing, and tracking people and their speaking activity are crucial for enriching the interactive experience between participants of a meeting, on a single site as well as on multiple sites, e.g. videoconferencing. The localization and tracking tasks play fundamental roles in two areas. The first one is media processing: speaker location is useful to select or steer a camera as part of a visualization or production model, to enhance the audio stream via microphone-array beamforming for speech recognition, to provide accumulated information for person identification, and to recognize location-based events (e.g. a presentation). The second one is human interaction analysis: social psychology has highlighted the role of non-verbal behavior (e.g. gaze and facial expressions) in interactions, and the correlation between speaker turn patterns and aspects of the behavior of a group [63]. Extracting cues to identify such multimodal behaviors requires reliable speaker localization and tracking.

Although the above tasks are facilitated in meetings by the constraints of the physical space and the expected type of human activities, they still pose several challenges. The interactive nature of meetings involves spontaneous multi-party speech, which contains highly dynamic patterns of speaker turns, including short speech utterances, non-linear human motion, partial and total visual occlusion, and multiple sources (multiple overlapping speakers and/or background noise sources). The meeting room environment is thus highly dynamical, and in order to develop useful applications for enhanced user experience *during* meetings (e.g. online and realtime) and *after* meetings (e.g. automatic summarization or queries from a user), robust and computationally tractable methods that can cope with the multisource aspect as well as the highly dynamical aspect are necessary. Ideally, such methods should address, in principled ways, the need for fusion of perceptual data (e.g. multiple audio and visual sources), in order to exploit the modalities' redundancy and complementarity, and the need for accurate descriptions of the interactive, multiperson processes that meetings contain (e.g. representing the dynamic status of each individual, while accounting for the constraints introduced by their interaction).

This report presents an overview of existing work on localization and tracking of multiple interlocutors with multiple sensors. Rather than being exhaustive, the report attempts to provide a non-expert reader with pointers to what the authors regard as representative work in the domain, and to present a succint discussion of the advantages and limitations of such methods, including the ones that have been developed as part of the AMI project. The *multisource* context will be the focus throughout. Additionally, given that the emphasis of the review is on localization and tracking of talking people, we limit the review to audio-only and audio-visual (AV) methods.

The report is organized as follows. Section 2 reviews existing work on localization and tracking with audio sensors. Section 3 does so for work using audio-visual sensors. Finally, Section 4 discusses available resources (i.e., data and related annotations) in this domain.

2 Localization and tracking with audio sensors

The speed of sound in the air being finite and relatively low in an indoor environment (around 342 m/s), most practical audio localization/tracking applications rely on time asynchrony between the waves arriving at multiple microphones in multiple locations, called microphone arrays. A three-fold inverse problem then arises: that of inferring, from the slight differences between the recorded signals, the number of active speakers at any given time (detection), their instantaneous positions (localization) and their spatio-temporal trajectories over time (tracking).

These "slight differences" are tightly linked to the geometrical placement of the microphones. More precisely, the differences are usually measured from three non-exclusive viewpoints:

1. Time asynchrony: the time of flight of the acoustic waves from mouth to microphone is different for different microphones, due to their different placements. Audio localization/tracking methods relying on time asynchrony usually require precise knowledge of the microphone array's geometry. However, they do not require any particular knowledge about the room, and are thus the most developed in terms of practical applications. Omnidirectional microphones are used in most cases. Typical geometries include Uniform Linear Arrays (ULA) and Uniform Circular Arrays (UCA): a finite number of microphones equally spaced along a line or a circle, respectively. However, a solution particularly designed for meeting rooms is the Huge Microphone Array (HMA) including a large number of microphones on a wall [76].

2. Impulse response: for a given mouth location, the path travelled by the sound to the various microphones will vary. Hence, the impulse response will vary as well. Assuming the impulse response characteristics of the room to be perfectly known beforehand (calibration), it is possible to deduce the position of the speaker. However, the tedious calibration step is often undesirable in practical application (e.g. portable videoconferencing systems), so the impulse responses need to be estimated in an online, automatic fashion. This task is also called blind Multiple Inputs Multiple Outputs (MIMO) channel identification [16, 11], where geometrical knowledge of the microphone array is not required. The task is tightly linked to Blind Source

Separation approaches [12]. Solving this problem amounts to retrieve a complete model of the meeting room, which would permit not only to locate but also to separate the various signals at the same time. This problem is still difficult and open to research. A preliminary test of such a method can be found in [13].

3. Microphone channel: recently, it was proposed to use several directional microphones placed at the same location, but oriented towards different directions [60]. The direction-dependent transfer function of each microphone is assumed to be known, so that the speaker location can be reconstructed. This solution can also be combined with the first group of approaches [72].

In the following we focus on the first group of solutions, for which [9, 44] provide comprehensive introductions. These methods are typically linked, directly or indirectly, to the following observation: if two signals $x_1(t)$ and $x_2(t) = x_1(t - \tau_{12})$ are received by two microphones, with a relative delay τ_{12}, the cross-correlation function $f(\tau) = \int_t x_1(t) x_2(t - \tau) dt$ will have a maximum at $\tau = \tau_{12}$. We first briefly mention the detection issue within the context of source localization, then examine the various instantaneous source localization methods. Finally, tracking of speaker trajectories over time is presented.

2.1 Detection

Since the speech of a given person is sporadic, it is needed *not* to estimate any speaker location during silence. Traditional Voice Activity Detectors (VADs) typically use single channel features, such as energy and zero-crossing rate. Although well adapted to single channel tasks such as automatic speech recognition, they are suboptimal in terms of localization precision [49]. Thus, they are not necessarily adapted to the task of acoustic source localization. Alternative methods that rely on the cross-correlation between channels can be found in [17, 51, 52].

2.2 Localization

The usual meaning of "localization" is to identify the location of the various active speech sources in physical space, from a short time frame on which speech is considered as stationary (typically 20 to 30 ms). As mentioned above, methods based on the asynchrony between the various microphones rely on the cross-correlation between those signals. They can be divided into two types: time delay of arrival (TDOA), and direct methods.

The TDOA methods consist in first estimating the time delays between each pair of microphones, and then deriving the location of the sources from geometrical considerations, e.g. using the Linear Intersection algorithm [8]. The main bottleneck is the time delay estimation (TDE) step, which may be affected by reverberations. A practical improvement can be obtained by modifying the cross-correlation function, e.g. using the generalized cross correlation phase transform (GCC-PHAT) [43]. In the case of multiple sources and one microphone pair, an alternative method for the estimation of multiple time-delays is the Adaptive Eigenvalue Decomposition Algorithm [6]. However, in the general case of multiple microphone pairs, multiple sources, and multiple sound paths (reverberations), it is not obvious how to pair the various time-delays observed at the various pairs of microphones in order to deduce the exact location of the acoustic sources.

The direct methods avoid this bottleneck by directly inferring the source(s) locations from the measured signal. They can be divided into two groups: Coherent Signal Subspace Processing (CSSP), and beamforming. CSSP methods [87, 19] are extensions of narrowband methods originated in the fields of radar and communications [73]. Although they allow in theory to estimate jointly the number of sources and their locations, they suffer from sensitivity to reverberant environments and/or need sufficient amounts of data.

Beamforming localization methods, also known as Steered Response Power (SRP) methods, are a reasonable alternative that does not rely on strong room knowledge/modelling assumptions. The idea is to estimate the power at any location in space by compensating for the corresponding delays between the signals ("steering" the array) [44]. Multiple simultaneous sources will be reflected by multiple maxima across the search space. However, reverberations will also appear as maxima ("virtual sources"). A partial solution to this issue is to combine the flexibility of SRP methods with the robustness to reverberations of PHAT: this is known as SRP-PHAT [22]. In general, the main drawback of SRP methods is that the search space can be large (e.g. the whole room in the case of meetings). Recently, an approach was proposed that discretizes the search space into volumes, and reduces the search to "active" volumes only [25, 26]. In such a framework, localization amounts to determine whether there is an acoustic source present within each of a predetermined, finite number of volumes. However, spatial resolution and interference between the signals of the different speakers may be an issue. Indeed, using the whole spectrum to locate multiple sources leads to unnecessary noise in the location estimation. In other words, spectral data from all sources is used to locate a given source, thus inducing an unnecessary bias in the location estimate. A possible way to address these issues is the "sparsity assumption" in the frequency domain [51, 52]. This assumption is derived from statistical observations on human speech [71], and simply means that within a frequency bin, only one speech source is dominant in terms of magnitude, while all other sources can be neglected.

In practice, although CSSP methods and, more recently, MIMO/BSS methods have seen very promising developments, they are still not as practically effective as the SRP methods. One drawback of SRP methods is that number of active sources and reverberations are not determined automatically. However, this issue can be efficiently addressed by clustering/tracking methods.

2.3 Tracking

Tracking can be viewed as the task of filtering instantaneous location estimates provided by the methods mentioned above. The Kalman filter [39, 90] assumes dynamics to be linear and Gaussian. These assumptions become an issue when dealing with human motion (non linearities such as sharp turns). Furthermore, in spontaneous speech, utterances are short (typically less than a second), speaker changes often, and overlaps represent a non-negligible portion of speech [75].

The Extended Kalman Filter (EKF) was proposed to accomodate non-linear dynamics through a linearization step [79], however it is known to be practically difficult to tune its parameters [37]. More recently the Unscented Kalman Filter (UKF) was proposed to avoid this linearization step and accomodate non-Gaussian measurement noise sources [38, 37, 53]. For a recent application of the UKF to acoustic source localization, see [24]. However, these approaches may encounter difficulties when dealing with spontaneous speech, which is both highly changing in space (speaker changes) and sporadic over time (short utterances).

As an alternative, Sequential Monte-Carlo (SMC) methods, also known as Particle Filtering (PF), approximate the optimal Bayesian filter by representing probability distributions through a finite set of particles [33, 23]. For a state-space model, a PF recursively approximates the filtering distribution of states given observations using a dynamical model, an observation model, and sampling techniques, by predicting candidate configurations and measuring their likelihood, in a process that amounts to random search in the configuration space. Applications to single acoustic source localization and tracking can be found in [83, 88, 89], and a comprehensive review in [54]. However, the fast-changing speaker turns encountered in spontaneous multi-party speech requires either specific multisource models [46] or adapting the single-source model to "switching between speakers" situations [55]. Estimating the number of active speech sources is still an issue, tightly linked to the data association issue. Although Particle Filters can model multiple objects via multi-modal distributions, deciding which modes are significant and which objects they belong to is an open issue. Moreover, when the number of active objects varies very often along time, complex birth/death rules are needed.

Recently, alternative approaches were proposed where the number of active speech sources need not be known [47]. An unsupervised, online approach called "short-term clustering" was proposed, that automatically groups location estimates that are close to each other in space and time, and separate those that are not. [48] demonstrates on real data that it can be directly applied to the task of spontaneous speech segmentation, without requiring any constraint from the participants. The resulting speech segmentation is very precise, even when multiple participants talk at the same time. In addition, "short-term clustering" is shown to allow for detection and solving of trajectory crossing issues.

3 Localization and tracking with audio-visual sensors

Localizing and tracking speakers in enclosed spaces using AV information has increasingly attracted attention in signal processing and computer vision [69, 34, 20, 67, 27, 84, 95, 2, 5, 18, 15], given the complementary characteristics of each modality. Broadly speaking, the differences among existing works arise from the overall goal (tracking single vs. multiple speakers), the specific detection/tracking framework, and the AV sensor configuration. Much work has concentrated on the single-speaker case, assuming either single-person scenes [20, 67, 2], or multiperson scenes where only the location of the current speaker needs to be tracked [69, 34, 27, 84, 95, 5]. Many of these works used simple sensor configurations, i.e., one camera and a microphone pair [20, 67, 84, 5]. Other works have addressed the data fusion problem using multiple microphones and a single camera [28], and others have used multiple cameras, either non-calibrated [29] or fully calibrated [95, 65], given the fact that, while single cameras are useful for remote conferencing applications, multiperson conversational settings like meetings often call for the use of multiple sensors to cover the entire workspace (table, whiteboards, etc.). Among the existing techniques, probabilistic generative models based on exact [67] or approximate inference methods, both variational [5] and sampling-based [84, 95, 28, 29, 65], appear to be the most promising, given their principled formulation and demonstrated performance.

None of the above works, however, can handle the problem of continuously inferring, from audio and video data, the location and speaking status for multiple people in a realistic conversational setting. In fact, although audio-based multispeaker tracking and vision-based multiobject tracking have been studied for a few years as separate problems in signal processing [82, 70, 85, 51] and computer vision [36, 68, 93, 94], respectively, the AV multispeaker tracking problem has been studied only relatively recently, making use of more complex sensor configurations [21, 40, 77, 14, 15, 18, 4, 30]. Each of these works is briefly discussed in the following. For presentation purposes, we categorize the existing work as being either system-oriented or model-oriented, where the emphasis in the first case is on module integration, while in the second case is on a unifying mathematical formulation.

Regarding the first category, the work in [21] described a system based on a device that integrates a small circular microphone array and several calibrated cameras, whose views are merged into a panorama, The system, in which each person is tracked independently, consists of three modules: AV auto-initialization, using either a standard acoustic source localization algorithm or visual cues, visual tracking using a Hidden Markov Model (HMM), and tracking verification. The work in [40] described a non-probabilistic multispeaker detection algorithm using an omnidirectional camera (which has limitations of resolution) and a microphone array, calibrated with respect to each other. At each video frame, the method extracts skin-color blobs by traditional techniques, and then detects a sound source using standard beamforming on the small set of directions indicated by the skin-blob locations. The work in [77] described an AV multispeaker system, based on a stereo camera and a linear microphone array, consisting of three separate modules: stereo-based visual tracking of 3-D head location and pose for each person independently, estimation of the direction of arrival of the audio signal with the microphone array, and estimation of audio-visual synchronous activity. Two hypothesis tests are used to make independent decisions about the speaking activity and visual focus of the speakers, based on simple statistical models defined on the observations derived from each module. The work in [14] uses a number of standard techniques in separate modules that are later integrated into a system that estimates the location and identity of the meeting participants, and detects the current speaker, using a setup including four calibrated cameras (an omnidirectional camera located on the center of the meeting table, and four cameras located in the corners of the meeting room), and a 16-microphone array, located on one end of the table.

For the second category, a number of probabilistic generative models have been recently proposed for the task of simultaneously inferring location and speaking activity of multiple interlocutors. All of them are based on PF techniques [15, 18, 4, 30, 31], but differ in the choices of state space, dynamical model, observation model, and sampling technique. The work in [15] used two calibrated cameras and four linear sub-microphone arrays on a wall, and was based on the model first proposed in [36], which defines a multi-person state-space where the number of people can vary over time. A full-body multi-person observation model was defined by two terms: one for video, derived from a pixelwise background substraction model, and one for audio, derived from a set of short-time Fourier transforms computed on each microphone's signal. The PF relied on basic importance sampling (IS), and so is likely to become rapidly inefficient as the number of people increases. The work in [18] used the same calibrated sensor setup as [21], and tracked multiple speakers with a set of independent PFs, one for each person. For sampling, each PF uses a mixture proposal distribution, in which the mixture components are derived from the output of single-cue trackers (based on audio, color, or shape information). This proposal distribution increases robustness in case of tracking failure in single modalities. Furthermore, each single-object observation model is assumed to be factorized over the various cues. The work in [4] uses a setup composed of a stereo camera and a circular 8-microphone array, and uses a basic PF to perform inference over a multi-person state space, assuming that the multi-object observation model can be factorized over participants. However, the approach was only applied to two-people scenes, likely due to the known limitations of the basic PF algorithm. The work in [30, 31], developed in the context of the AMI project, presents an approach in a meeting room consisting of three uncalibrated cameras covering the physical space with mostly non-overlapping fields-of-view, and a circular 8-microphone array placed on the center of the meeting table. The model uses a mixed-state, multi-object state-space, which integrates a pair-wise person occlusion model through the addition of a Markov Random Field prior in the multi-object dynamic model. To address the problems of traditional PFs in handling the high-dimensional state space defined by the joint multi-person configurations, inference in this model is performed with a Markov Chain Monte Carlo particle filter (MCMC-PF), which results in high sampling efficiency [56, 42]. The model integrates audio-visual data through an observation model where audio observations are derived from a source localization algorithm, and visual observations are based on models of the shape and spatial structure of human heads. Overall, the model in [30, 31] has two advantages over [15, 18]. First, it explicitly incorporates a pairwise person interaction prior term, which is especially useful to handle person occlusion. Second, it uses an MCMC sampling technique, which allows to track several objects in a tractable manner (effectively close to the case of independent PFs), while preserving the rigorous joint state-space formulation.

An important initiative related to evaluation of audio-visual technologies for localization and tracking is the recent Workshop on Classification of Events, Actions and Relations (CLEAR), where audio-visual approaches were evaluated in the context of seminar and conference rooms to track single presenters on common data and using a common evaluation protocol [1, 10, 41, 66, 7, 32]. As representative examples, in [41], a 3D tracking with stand-alone video and audio trackers was combined using a Kalman filter. The work in [66] proposed an algorithm based on a particle filter approach to integrate acoustic source localization, person detection, and foreground segmentations using multiple cameras and multiple pairs of microphones. It was demonstrated that the specific audio-visual formulation yields greater tracking accuracy than a filter based on individual modalities. The reader is referred to the CLEAR workshop proceedings for details about all the approaches.

Some of the recent AMI work focused on integrating, improving and evaluating a system for hands-free speech recognition in meetings [62, 59, 58] based on an audio-visual sensor array, including the multi-modal approach for multi-person tracking [30, 31], and speech enhancement and recognition modules. As mentioned before, 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. However, with a few exceptions, speaker tracking research has been largely decoupled from microphone array speech recognition research. The work in [3] 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. In the work in [91], a particle filter that fuses audio from multiple large microphone arrays and video from multiple calibrated cameras was used in the context of seminar rooms, in which there is essentially one main speaker (the lecturer).

In the system in [59], 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 [30, 31]. 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 multi-modal 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 outperforms single table-top microphones and is comparable to lapel microphone for all the scenarios, as measured by both 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 both speech enhancement and recognition. An analysis of the effects of location accuracy on the recognition of overlapping speech is presented in [58].

4 Available Data Resources

Most of the research summarized in the previous two sections has been conducted over a number of non-standardized audio or audio-visual data sets, which vary from each other with respect to the specific sensor setup, the type of recorded situations, the structure of the data set, the type of existing annotations, and their degree of availability to others for research purposes. The community in this domain, however, has already acknowledged the considerable effort involved in collecting such data, and the need to rely in common evaluation procedures.

In the context of the AMI project, an audio-visual corpus, called AV16.3, was recorded and annotated, and reported in [50]. This corpus includes a high variety of scenarios, ranging from static, constrained cases, to dynamic and natural ones, with multiple seated or moving speakers in a meeting room. The sensors include two eight-microphone circular arrays on a table, and three cameras around the room. The calibration of the cameras allowed to reconstruct the ground-truth location of the mouth of each person with a 3-D error inferior to 1.2 cm. Overall, this data set should be interesting for both the audio and the vision communities, and is publicly available. A second corpus, called AV16.7, was recorded to evaluate the multi-person tracking task [78], and contains sequences including up to four people conversing and moving in the meeting room. Finally, an audio-visual corpus for speech recognition, called the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus, was also recorded. The specification and structure of the full corpus are detailed in [57]. The corpus includes cases of single stationary speakers, single moving speakers, and stationary overlapping speakers. In the first scenario, the speaker reads out sentences from different positions within the meeting room. In the second one, the speaker moves between different positions while reading the sentences. Finally, in the third scenario, two speakers simultaneously read sentences from different positions within the room. Much of the data comprises non-native English speakers with different speaking styles and accents. The corpus is therefore suitable for research on both tracking and speech recognition.

Another important data resource is the one coordinated by NIST and the CHIL (Computers in the Human Interaction Loop) european project through the CLEAR initiative, where data collected and annotated in the CHIL meeting and lecture rooms become available for purposes of common evaluation [80].

References

[1] A. Abad, C. Canton-Ferrer, C. Segura, J. L. Landabaso, D. Macho, J.R. Casas, J. Hernando, M. Pardas, C. Nadeu, ”UPC Audio, Video and Multimodal Person Tracking Systems in the CLEAR Evaluation Campaign,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR),* Southampton, Apr. 2006.

[2] P. Aarabi and S. Zaky, “Robust sound localization using multi-source audiovisual information fusion,” *Information Fusion,* vol. 3, no. 2, pp. 209–223, Sep. 2001.

[3] 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.

[4] H. Asoh, F. Asano, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, J. Ogata, and K. Yamamoto, “An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion,” in *Proc. Int. Conf. on Information Fusion (IF),* Stockholm,, Jun 2004.

[5] M. Beal, H. Attias, and N. Jojic, “Audio-video sensor fusion with probabilistic graphical models,” in *Proc. European Conf. on Computer Vision (ECCV)*, May 2002.

[6] J. Benesty, “Adaptative eigenvalue decomposition algorithm for passive acoustic source localization,” *Journal of the Acoustic Society of America*, vol. 107, no. 1, pp. 384–391, January 2000.

[7] K. Bernardin, T. Gehrig, R. Stiefelhagen, ”Multi- and Single View Multiperson Tracking for Smart Room Environments,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR),* Southampton, Apr. 2006.

[8] M. Brandstein, *A Framework for Speech Source Localization Using Sensor Arrays*, Ph.D. thesis, Brown University, 1995.

[9] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[10] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, F. Tobia, ”A Generative Approach to Audio-Visual Person Tracking,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[11] H. Buchner, R. Aichner, and W. Kellerman, “Trinicon: A versatile framework for multichannel blind signal processing,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[12] H. Buchner, R. Aichner, and W. Kellermann, “Relation between blind system identification and convolutive blind source separation,” in *Proc. HSCMA Workshop*, Piscataway, NJ, USA, Mar. 2005.

[13] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, “Simultaneous localization of multiple sound sources using blind adaptive mimo filtering,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005.

[14] C. Busso, S. Hernanz, C.-W. Chu, S.-I. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, “Smart room: Participant and speaker localization and identification,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[15] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, “Multiple person and speaker activity tracking with a particle filter,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[16] J. Chen, J. Benesty, and A. Huang, “MIMO acoustic signal processing,” Invited Talk, HSCMA Workshop, Mar. 2005.

[17] J.F. Chen and W. Ser, “Speech detection using microphone array,” *Electronic Letters*, vol. 36, no. 2, Jan. 2000.

[18] Y. Chen and Y. Rui, “Real-time speaker tracking using particle filter sensor fusion,” *Proc. of the IEEE*, vol. 92, no. 3, pp. 485–494, Mar. 2004.

[19] E. Di Claudio and R. Parisi, “Multi-source localization strategies,” in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 9, pp. 181–201. Springer, 2001.

[20] R. Cutler and L. Davis, “Look who’s talking: Speaker detection using video and audio correlation,” in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, New York, Jul. 2000.

[21] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Int. Conf. on Multimedia (MM)*, Juan les Pins, Dec. 2002.

[22] J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*, Ph.D. thesis, Brown University, Providence RI, USA, 2000.

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

[24] T.V. Dvorkind and S. Gannot, "Speaker localization using the unscented kalman filter," in *Proc. HSCMA Workshop*, Mar. 2005.

[25] R. Duraiswami, D. Zotkin, and L.S. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[26] D.N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, September 2004.

[27] J. Fisher, T. Darrell, W.T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Dec. 2000.

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

[29] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, Nice, Oct. 2003.

[30] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[31] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-Visual Probabilistic Tracking of Multiple Speakers in Meetings," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 15. No. 2. pp. 601-616, Feb. 2007.

[32] T. Gehrig, J. McDonough, ”Tracking of Multiple Speakers with Probabilistic Data Association Filters,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[33] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, “Novel approach to nonlinear/non-Gaussian bayesian state estimation,” in *IEE Proceedings*, 1993, vol. 140, pp. 107–113.

[34] J. Hershey and J. Movellan, “Audio vision: Using audio-visual synchrony to locate sounds,” in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Nov. 1999.

[35] M. Isard, *Visual Motion Analysis by Probabilistic Propagation of Conditional Density*, D.Phil. Thesis, Oxford University, 1998.

[36] M. Isard and J. MacCormick, “BRAMBLE: A Bayesian multi-blob tracker,” in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, Jul. 2001.

[37] S.J. Julier and J.K. Uhlmann, “A new extension of the Kalman filter to nonlinear systems,” in *Proc. Int. Sym. on Aerospace/Defense Sensing, Simulation and Controls (AeroSense)*. 1997.

[38] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte, “A new approach for filtering nonlinear systems,” in *Proc. American Control Conf.*, 1995, pp. 1628–1632.

[39] R.E. Kalman, “A new approach to linear filtering and prediction problems,” *Trans. of the ASME, Journal of Basic Engineering*, vol. 82, pp. 35–45, March 1960.

[40] B. Kapralos, M. Jenkin, and E. Milios, “Audio-visual localization of multiple speakers in a video teleconferencing setting,” *Int. J. Imaging Syst. and Tech.*, vol. 13, pp. 95–105, 2003.

[41] N. Katsarakis, G. Souretis, F. Talantzis, A. Pnevmatikakis, L. Polymenakos, ”3D Audiovisual Person Tracking Using Kalman Filtering and Information Theory,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[42] Z. Khan, T. Balch, and F. Dellaert, “An MCMC-based particle filter for tracking multiple interacting targets,” in *Proc. European Conf. on Computer Vision (ECCV)*, Prague, May 2004.

[43] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

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

[45] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'ready: a meeting recorder and browser," *ACM Computing Surveys*, vol. 31, no. 2es, Jun. 1999.

[46] J.R. Larocque, J.P. Reilly, and W. Ng, "Particle filters for tracking an unknown number of sources," *IEEE Trans. on Signal Processing*, vol. 50, no. 12, December 2002.

[47] G. Lathoud, I.A. McCowan, and J.M. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *Proc. NIST ICASSP Meeting Recognition Workshop*, 2004.

[48] G. Lathoud, J.M. Odobez, and I.A. McCowan, "Short-term spatio-temporal clustering of sporadic and concurrent events," IDIAP-RR 04-14, IDIAP, 2004.

[49] G. Lathoud and I.A. McCowan, "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays," in *Proc. SAPA 2004*, Oct. 2004.

[50] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Martigny, Jun. 2004.

[51] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[52] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-Based Detection for Hands-Free Speech Enhancement in Cars," *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*, 2006.

[53] J. LaViola, "A comparison of Unscented and Extended Kalman Filtering for estimating quaternion motion," in *Proc. American Control Conf.*. June 2003, pp. 2435–2440, IEEE Press.

[54] E. Lehmann, *Particle Filtering Methods for Acoustic Source Localisation and Tracking*, Ph.D. thesis, Australian National University, July 2004.

[55] E. Lehmann, "Importance sampling particle filter for robust acoustic source localisation and tracking in reverberant environments," in *Proc. HSCMA Workshop*, Piscataway, NJ, USA, March 2005.

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

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

[58] H.-K. Maganti and D. Gatica-Perez, “Speaker Localization for Microphone-Array-Based ASR: the Effects of Accuracy on Overlapping Speech,” in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Banff, Nov. 2006.

[59] H.-K. Maganti, D. Gatica-Perez, and I. McCowan, “Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array,” IDIAP Research Report IDIAP-RR-06-24,, submitted to IEEE. Trans. on Audio, Speech, and Language Processing, Apr. 2006

[60] M. Matsumoto and S. Hashimoto, “Multiple signal classification by aggregated microphones,” *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, July 2005.

[61] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, “Automatic analysis of multimodal group actions in meetings,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[62] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, “Speech acquisition in meetings with an audio-visual sensor array,” in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, Amsterdam, Jul. 2005.

[63] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[64] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, “The meeting project at ICSI,” in *Proc. Human Language Technology Conf. (HLT)*, San Diego, CA, March 2001.

[65] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, “A joint particle filter for audio-visual speaker tracking,” in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Trento, Oct. 2005.

[66] K. Nickel, T. Gehrig, H.K. Ekenel, J. McDonough, R. Stiefelhagen. ”An Audio-visual Particle Filter for Speaker Tracking on the CLEAR06 Evaluation Dataset,” in *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, Apr. 2006.

[67] V. Pavlovic, A. Garg, and J. Rehg, “Multimodal speaker detection using error feedback dynamic bayesian networks,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, SC, 2000.

[68] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, “Color-based Probabilistic Tracking,” in *Proc. European Conf. on Computer Vision (ECCV)*, Copenhagen, May 2002.

[69] G.S Pingali, G. Tunali, and I. Carlbom, “Audio-visual tracking for natural interactivity,” in *Proc. ACM Int. Conf. on Multimedia (MM)*, Orlando, Oct. 1999.

[70] I. Potamitis, H. Chen, and G. Tremoulis, “Tracking of multiple moving speakers with multiple microphone arrays,” *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, Sep. 2004.

[71] S.T. Roweis, “Factorial Models and Refiltering for Speech Separation and Denoising,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2003.

[72] Y. Rui, D. Florencio, W. Lam, and J. Su, “Sound source localization for circular arrays of directional microphones,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[73] R.O. Schmidt, “Multiple Emitter Location and Signal Parameter Estimation,” *IEEE Trans. on Antennas and Propagation*, vol. AP-34, pp. 276–280, March 1986.

[74] E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: findings and implications for automatic processing of multi-party conversation,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Aalborg, Sep. 2001.

[75] E. Shriberg, A. Stolcke, and D. Baron, “Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech,” in *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (Prosody)*, 2001.

[76] H. F. Silverman, Ying Yu, J. M. Sachar, and W. R. Patterson III, “Performance of real-time source-location estimators for a large-aperture microphone array,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[77] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell, “A multi-modal approach for determining speaker location and focus,” in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Vancouver, 2003.

[78] K. Smith, S. Schreiber, I. Potucek, V. Beran, G. Rigoll, D. Gatica-Perez, ”2D Multi-Person Tracking: A Comparative Study in AMI Meetings,” in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington DC, May 2006.

[79] H. Sorenson, *Kalman Filtering: Theory and Application*, IEEE Press, 1985.

[80] R. Stiefelhagen and J. Garofolo (organizers), CLEAR Evaluation Workshop, Southampton, Apr. 2006.

[81] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[82] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone array measurements," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Apr. 1997.

[83] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[84] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, July 2001.

[85] B. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers with random sets," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[86] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001.

[87] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, August 1985.

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

[89] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 6, November 2003.

[90] G. Welch and G. Bishop, "An introduction to the kalman filter," TR 95-041, Dept. of Computer Sc., Uni. of NC at Chapel Hill, 2004.

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

[92] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[93] T. Yu and Y. Wu, "Collaborative tracking of multiple targets," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, Jun. 2004.

[94] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington DC, Jun. 2004.

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