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

Updated version 23.01.2007

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 [47, 64, 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 [64], 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 [47, 64, 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 [46]. 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 [58]. A body of previous work, e.g. [46, 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 [45].

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 [59], and lead to better performance in speech recognition applications [42]. Most approaches are based upon the post-filter proposed by Zelinski [66], 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) [55], a range of new speech-specific source separation algorithms [50, 53], 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 [62, 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, 28]:

- 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 [61, 49, 52, 51, 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, 56, 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[54]. 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 [28].

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 [57, 63, 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 [57, 63, 56] 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 [56]. 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 [63] 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 [28] 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. [65]) or equivalent forms. Maximum likelihood training schemes are generally replaced by discriminative training using discriminative criteria such as the minimum phone error(MPE) criterion [48]. 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 [56, 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[60] 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 [44], 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] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2), 1997.

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

[46] 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.

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

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] P. Schwarz, P. Mat?jka, and J. Cernock? Towards lower error rates in phoneme recognition. In *Proc. of 7th Intl. Conf. on Text, Speech and Dialogue*, number ISBN 3-540-23049-1 in Springer, page 8, Brno, 2004.

[55] 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.

[56] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Manda, B. Peskin, C. Wooters, and J. Zheng. Further progress in meeting recognition: The ICSI-SRI Spring 2005 Speech-to-Text evaluation system. In *Proc. NIST RT'05 Workshop.*, 2005.

[57] 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.

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

[59] 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.

[60] V. Wan and T. Hain. Strategies for language model web-data collection. In *Proc. ICASSP*, 2006.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.