

FAR-FIELD MULTIMODAL SPEECH PROCESSING AND CONVERSATIONAL INTERACTION IN SMART SPACES

Gerasimos Potamianos,¹ Jing Huang,¹ Etienne Marcheret,¹ Vit Libal,¹ Rajesh Balchandran,¹ Mark Epstein,¹ Ladislav Seredi,² Martin Labsky,² Lubos Ures,² Matthew Black,* Patrick Lucey*

¹ IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

² IBM Czech Republic – The Park, V Parku 2294/4, Prague-Chodov 14800, CZ

Emails: ¹ {gpotam, jghg, etiennem, libalvit, rajeshb, meps}@us.ibm.com ² {ladislav_seredi, martin.labsky, lubos_ures}@cz.ibm.com

ABSTRACT

Robust speech processing constitutes a crucial component in the development of usable and natural conversational interfaces. In this paper we are particularly interested in human-computer interaction taking place in “smart” spaces – equipped with a number of far-field, unobtrusive microphones and camera sensors. Their availability allows multi-sensory and multi-modal processing, thus improving robustness of speech-based perception technologies in a number of scenarios of interest, for example lectures and meetings held inside smart conference rooms, or interaction with domotic devices in smart homes. In this paper, we overview recent work at IBM Research in developing state-of-the-art speech technology in smart spaces. In particular we discuss acoustic scene analysis, speech activity detection, speaker diarization, and speech recognition, emphasizing multi-sensory or multi-modal processing. The resulting technology is envisaged to allow far-field conversational interaction in smart spaces based on dialog management and natural language understanding of user requests.

Index Terms— Acoustic scene analysis, speech activity detection, speaker diarization, speech recognition, audio-visual speech recognition, dialog systems, fusion, smart rooms.

1. INTRODUCTION

There has been increasing interest in recent years in advancing the human-computer interaction paradigm from its current stage, where the focus lies on the computers and humans are forced to adapt own behavior to their limitations, to a paradigm where computers are enabled to detect, understand, learn, and adapt to human activity and requests, while fading into the background. Crucial to this effort is the development of robust technology to process the most natural human communication modality, i.e. speech, allowing the development of usable, natural, and unobtrusive conversational human-machine interfaces.

In this paper, we overview recent efforts at IBM Research towards achieving this goal inside smart spaces equipped with a multitude of audio and visual sensors. Their availability allows multi-sensory and multi-modal processing, thus improving robustness of speech-based perception technologies in a number of scenarios of interest. Here, we primarily concentrate on lectures and meetings tak-

ing place in smart conference rooms. The problem has recently attracted much interest, being the focus of a number of research efforts and international projects, for example CHIL [1], AMI / AMIDA [2], and the U.S. National Institute of Standards and Technology (NIST) Smartspace effort [3]. Developed perception technologies in this scenario have been the subject of vigorous evaluation campaign series, for example the Rich Transcription (RT) Meeting Recognition Evaluation [4] and the campaign on the Classification of Events, Activities and Relationships [5]. Needless to say, these technologies can be appropriately adapted to additional ambient intelligence scenarios, for example smart homes. This subject is of interest to the NETCARITY project [6] that focuses on elderly home occupants, and the DICIT project, where far-field speech technology is being developed to facilitate conversational interaction with the television set-top box [7].

In this paper, we focus on speech technology in smart spaces. Following an overview of relevant scenarios (Section 2), we first consider the problem of *acoustic scene analysis* in smart rooms (Section 3). Detection of acoustic events is crucial in monitoring, providing rich activity information, and being beneficial to speech processing – allowing for example better modeling of non-speech events. Of main interest of course is the problem of *automatic speech recognition* (ASR) or *speech-to-text* (STT), and its complementary technologies, *speech activity detection* (SAD) and *speaker diarization* (SPKR). All three partially address the “what”, “when”, and “who” of human interaction, and are important drivers of additional technologies, for example speaker localization, speaker recognition, summarization, and question answering. SAD, SPKR, and STT are discussed in Sections 4, 5, and 6, respectively. All above technologies can become more robust by taking advantage of the presence of multiple sensors, as discussed in some of the following sections. They can also benefit from the presence of visual information, giving rise to multimodal processing. A prime example is *audio-visual automatic speech recognition* (AVASR), presented in Section 7. Following this, *conversational interaction* is discussed in Section 8, and finally the paper concludes with a summary in Section 9.

2. SMART SPACES AND THE CHIL DATABASE

Smart spaces equipped with multiple audio-visual sensors have recently attracted significant research interest in a number of scenarios. For example, in the NETCARITY project, smart homes are envisaged that help the elderly to improve their wellbeing, independence, safety, and health [6]. For example, automatic audio-visual analysis of the home environment based on camera and microphone input (see also Fig. 1) can help identify health-critical situations – for example, detect a person falling – and monitor the activities of

This work has been partially supported by the European Commission through FP6 ICT projects CHIL, NETCARITY, and DICIT.

* Work performed during summer internships with the IBM T.J. Watson Research Center. *Matthew Black* is with the Speech Analysis and Interpretation Laboratory at the University of Southern California, Los Angeles, CA 90089, USA; email: matthew.black@usc.edu. *Patrick Lucey* is with the Speech, Audio, Image and Video Technology Laboratory, Queensland University of Technology, Brisbane, Australia; email: p.lucey@qut.edu.au.

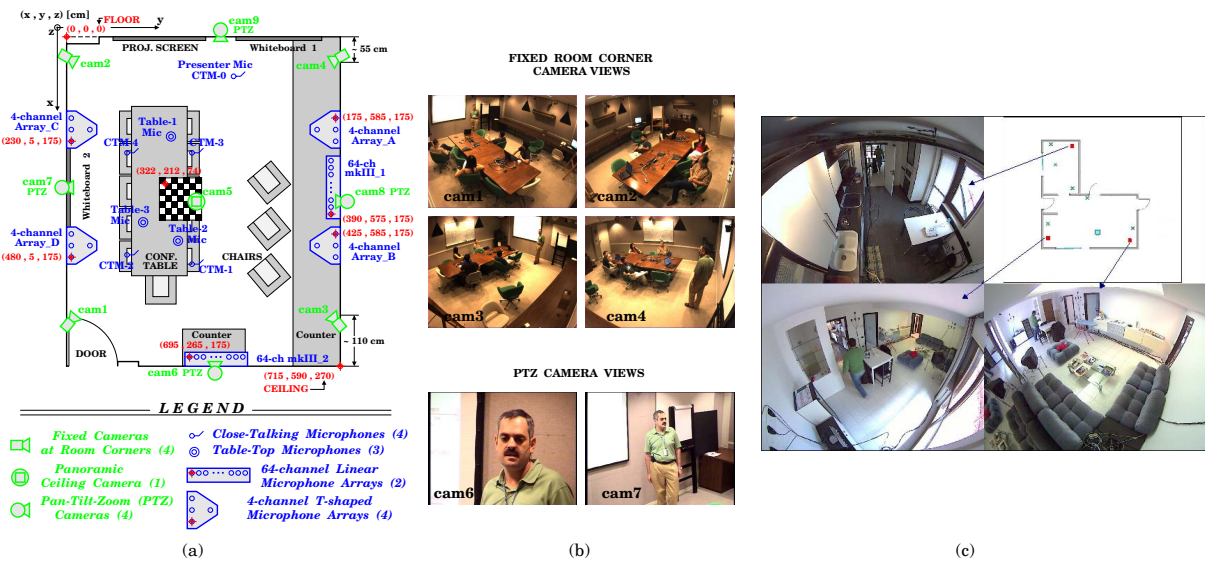


Fig. 1. Examples of smart spaces equipped with multiple audio-visual sensors: (a) Schematic diagram of the IBM smart conference room, developed in the CHIL project; (b) Example of video frames recorded during a lecture in this room; (c) Schematic diagram and recorded video frame examples in the smart home used by NETCARITY project partner FBK.

daily living of the elderly, issuing reminders, identifying deviations, and keeping distant loved ones in the loop. Robust perception technologies required to achieve such goals become possible by multi-sensory and multi-modal processing, exploiting the availability of multiple data streams. Speech processing technologies constitute a prime example, assisted for example by the presence of microphone arrays [8]. As a result, far-field conversational interaction with domestic devices becomes within reach, without the need of obtrusive close-talking microphones. One interesting such interaction scenario is far-field voice control of interactive TV / set-top box. This constitutes the main focus of the DICIT project [7].

Another case of interest is this of smart conference rooms, where lectures or meetings take place. These play a significant role in human collaboration activity in the workplace, with speech constituting the primary mode of communication. The main emphasis of this paper lies on speech perception technologies developed for this scenario as part of project CHIL [1]. There, five CHIL partner sites have created state-of-the-art smart rooms that contain similar sensory setups. These include multiple linear and T-shaped microphone arrays located on the room walls to allow source localization and beamforming [8, 9], as well as a number of table-top and close-talking microphones, the latter in an effort to benchmark degradation of ASR performance in the far-field. A number of visual sensors are also available, namely five fixed calibrated cameras with highly overlapping fields-of-view to allow person tracking [9, 10], and a number of pan-tilt-zoom (PTZ) cameras that provide close-ups of meeting participants (see also Fig. 1).

A large audio-visual database of lectures and interactive seminars has been generated as part of CHIL, collected in the five available smart rooms [11]. This corpus has provided development and test data for a multitude of perception technologies, rigorously evaluated as part of the RT and CLEAR campaigns [4, 5]. Here, we utilize the acoustic part of this database for acoustic scene analysis and speech processing. Notice that for the latter, additional publicly available corpora of similar nature are also employed [2, 12], in particular to allow better training of the ASR system (see also [13]). Additional datasets recorded in the CHIL smart rooms are also utilized to allow specific experiments in acoustic scene analysis [14]

and audio-visual ASR [15], as discussed in more detail in the corresponding sections.

3. ACOUSTIC SCENE ANALYSIS

Acoustic scene analysis constitutes an important problem, providing crucial meta-information about the activity in the smart space, as well as helpful input to other perception components. For example, in the case of lectures and meetings in CHIL, recognition of common sounds such as clapping or laughter provides information about the meeting state or interaction, while helping to improve speech technology robustness by better modeling of noise classes. Similarly, recognition of door knocks can provide valuable information to person tracking, since the event could trigger person movement to/from the door.

In CHIL, twelve acoustic events have been identified as being of interest [14], with two variants of the acoustic scene analysis task defined: *Acoustic event classification* (AEC), where the events of interest are pre-segmented (end-pointed) and the task is to identify them within the closed set of possible events (classes), and *acoustic event detection* (AED), where events have to be localized in time and identified. The latter is of course a more challenging problem.

We have developed a statistical classification approach for both problems, based on *hidden Markov models* (HMMs). These HMMs are analogous to whole-word speech models, with a separate model used for each acoustic event class, together with three additional HMMs to model speech, silence, and all other unknown noise. The HMMs have a left-to-right topology consisting of three states for AEC, but 30-50 states for AED in order to reduce false event insertions. HMM training commences with a flat start that employs the Viterbi algorithm, using 13-dimensional *perceptual linear prediction* (PLP) features as the front-end. Following 100 iterations of alignment-based training, a *linear discriminant analysis* (LDA) projection matrix is derived that allows inclusion of dynamic information in the resulting HMM. A number of the available microphone channels is used at training.

Following training, recognition of the acoustic events employs an HMM network, similar in fashion to speech recognition decod-

ing. In order to take advantage of the multi-sensory smart room, two approaches are explored for channel combination: A simple decision fusion mechanism, based on the ROVER software for combining ASR system outputs [16], and signal-level combination (feature fusion) employing beamforming of the available channels [8].

A number of corpora are used for training and testing the resulting systems. Two are isolated acoustic event databases, recorded at CHIL partner sites UPC and FBK [14]. Each contains approximately 50-60 instances of each event of interest, recorded in a relatively quiet environment with no overlap among them. The third set is part of the CHIL corpus discussed in Section 2 [11]. This database contains realistic interactions, and as a result, overlaps between acoustic events as well as speech are widespread.

As expected, system performance is significantly better in the isolated data sets, as measured by *acoustic event error rate* (AEER), a metric akin to word error rate in ASR. For example, AEC performance on the FBK isolated corpus is good, resulting to 2.1% AEER using the ROVER approach for system combination – beamforming is significantly worse, yielding 10.3% AEER. AED is of course a much harder problem; performance stands at 13.0% AEER. The problem becomes even more intractable in the case of the realistic CHIL seminar corpus. There, the AED error climbs to 92.1%.

4. SPEECH ACTIVITY DETECTION

Speech activity detection (SAD) is a pre-requisite to both speaker diarization and ASR. After SAD, long segments of non-speech (silence or noise) are removed, and the audio is partitioned into shorter segments for fast decoding and speaker segmentation. SAD is also of interest to other technologies, for example acoustic speaker localization.

The IBM SAD system is basically an HMM-based speech/non-speech decoder; speech and non-speech segments are modeled with five-state, left-to-right HMMs. The HMM output distributions are tied across all states and modeled with a mixture of diagonal-covariance Gaussian densities. The non-speech model includes the silence phone and three noise phones. The speech model contains all speech phones. Both are obtained by applying a likelihood-based, bottom-up clustering procedure to the speaker-independent acoustic model developed for ASR (see Section 6), but MAP-adapted to the CHIL training data [17]. On CHIL data, the system results to 5.1% SAD diarization error on the development set of the 2007 RT evaluation (RT07) and 6.3% on the corresponding test set, in both cases using a single table-top microphone. Employing multiple microphones does not help much, due to the SAD system operating point chosen to ensure that very little speech is missed. For example, on the RT07 development set, SAD error becomes 5.0% using all available table-top microphones within a decision fusion framework.

This behavior differs in the IBM SAD system developed for the 2006 RT evaluation (RT06s) [18]. That system differs significantly from the one just described, and it yields inferior performance of 15.0% and 10.6% SAD error on the RT07 development set, when using a single or multiple table-top microphones, respectively.

5. SPEAKER DIARIZATION

Speaker diarization (SPKR) aims to cluster the speech segments for each speaker, addressing the “who spoke when” problem. The result can help ASR performance allowing appropriate speaker adaptation, and also is of interest to other perception technologies, for example speaker identification.

We have developed an initial speaker diarization scheme as part of the CHIL ASR system for the RT06s evaluation campaign. A simple speaker clustering procedure combines SAD segments into pseudo-speaker clusters, as follows [19]: All homogeneous speech

segments are modeled using a single Gaussian density, and are bottom-up clustered into a pre-specified number of clusters employing K -means and a Mahalanobis distance measure. For CHIL data, the number of speaker clusters is set to the ad-hoc value of four over each lecture. This particular scheme proves sufficient for ASR, but is not designed to yield low SPKR error. Indeed, its performance on the RT07 development set is a dismal 70.4%.

For the RT07 evaluation, we made a systematic effort to improve SPKR system performance, by investigating a number of novel ideas. These included the use of word information from a speaker-independent speech recognizer, modifications to speaker cluster merging criteria and the underlying segment model, and the use of speaker models based on GMMs, and their iterative refinement by frame-level re-labeling and smoothing of decision likelihoods. In addition, the new system benefits from employing the improved SAD scheme discussed in Section 4. Details of this work can be found in [17].

Unfortunately however, the developed diarization system heavily depends on appropriately tuning thresholds in the speaker cluster merging process. Possibly as a result of over-tuning such thresholds, performance on the RT07 evaluation test set degrades significantly compared to the one observed on development data. For example, a 7.4% SPKR error is attained on the latter, but the best performance on the RT07 evaluation data is significantly worse (27.7% error).

6. AUTOMATIC SPEECH RECOGNITION

Developing robust far-field ASR technology is of crucial importance in the quest for conversational interaction in smart spaces, as well as for additional technologies, for example summarization and question answering based on spoken input. In this section we proceed with an overview of the IBM ASR system developed for the CHIL scenarios, as part of the RT07 ASR evaluation campaign [13]. This has evolved significantly over the CHIL ASR system submitted to the RT06s evaluation [19]. Three are the main components of interest: Acoustic modeling, language modeling, and the recognition (decoding) process.

Acoustic Modeling: Acoustic modeling commences with training a speaker-independent (SI) model, based on 40-dimensional acoustic features generated by an LDA projection of nine consecutive frames of 13-dimensional PLP features. The features are mean normalized on a per-speaker basis, and are extracted at 100 Hz. The SI model uses continuous density, left-to-right HMMs with Gaussian mixture emission distributions and uniform transition probabilities. Each HMM has three states, except for the single-state silence HMM. The system uses 45 phones, namely 41 speech phones, one silence phone, and three noise phones. The final HMMs have 6k context-dependent tied states and 200k Gaussians. Since the training corpora consists of both CHIL data and additional publicly available sets [12] – see also Section 2, MAP-adaptation of the SI model on the former was deemed necessary to improve performance on CHIL data.

The SI features are further normalized with a voicing model (VTLN) with no variance normalization. A VTLN model is subsequently trained on features in the VTLN warped space. The resulting HMMs have 10k tied states and 320k Gaussians. Following VTLN, a SAT system is trained on features in a linearly transformed feature space resulting from applying speaker-dependent fMLLR transforms to the VTLN normalized features. Following SAT, feature-space minimum phone error (fMPE) transforms are estimated, followed by MPE training and MAP-MPE on the available amount of CHIL-only data [13].

Following the above training procedure, two systems are built, one with the VTLN step present, and one with VTLN removed. Based on the latter, two additional SAT systems are built using the



Fig. 2. Examples of synchronous frontal and profile views of subjects collected in the IBM smart room to enable multi-view AVASR experiments.

randomized decision tree approach, and again having 10k states and 320k Gaussians [13]. Thus, a total of four far-field acoustic models are available for decoding.

Language Modeling: Five separate four-gram LMs are employed for language modeling. The first four – also used in the IBM RT06s system [19], are based on transcripts of CHIL data, additional non-CHIL meetings, text from scientific conference proceedings, and the Fisher data corpus [12], with a total of 43M words. A novel fifth LM employs 525M words of web data available from the EARS program [12]. For decoding, two interpolated LMs are used based on these five models. A reduced-size model, pruned to about 5M n -grams, is employed for static decoding, whereas a larger 152M n -gram model is used in conjunction with on-the-fly dynamic graph expansion decoding. For decoding, a 37k-word vocabulary is employed.

Recognition Process: Following speech segmentation and speaker clustering, and for each trained system, the final output is obtained in three decoding passes for each available far-field microphone: (a) An initial SI pass using MAP-adapted SI models to decode; (b) Using output from (a), warp factors using the voicing model and fMLLR transforms are estimated for each cluster using the SAT model. The VTLN features after applying the fMLLR transforms are subjected to the fMPE transform, and a new transcript is obtained by decoding, using the MAP-adapted MPE model and the fMPE features. (c) The output transcripts from step (b) are used in a cross-system fashion to estimate MLLR transforms on the MPE model. The adapted MPE model together with the large LM are used for final decoding with a dynamic graph expansion decoder.

Since four ASR systems are available, ROVER over their resulting outputs is applied for obtaining a single-microphone decoding [16]. In the multi-sensory setup of CHIL, ROVER can also be applied across various microphone results, amounting to a process akin to decision fusion. In practice, ROVER is first applied over all available table-top microphones, followed by ROVER over the available four systems.

Experimental Results: The above approach results to a 44.3% word error rate (WER) in the RT07 evaluation set, when all available table-top microphones are used. This corresponds to a significant improvement over the 47.9% WER, achieved using a single table-top microphone. These numbers also demonstrate much progress in the field over the RT06s system that resulted in 50.1% WER in the multi-microphone condition.

It is worth comparing far-field ASR performance to that of a close-talking system. Although not as much effort has been devoted into developing the latter (see [13, 19] for details), performance on the RT07 evaluation set is significantly better, standing at 33.4% WER.

7. AUDIO-VISUAL ASR

The perception technologies reported above have exploited multi-sensory input to improve performance, but of a single modality (audio). All can potentially further benefit (directly or indirectly) by

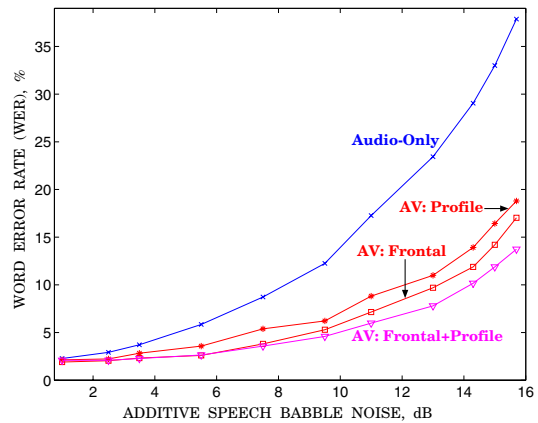


Fig. 3. Audio-only and audio-visual ASR word error rate, reported on the test set of the bimodal multi-view connected-digits database collected at IBM in CHIL [1]. Three AVASR systems are depicted, based on profile view, frontal view, and both views (“multi-view” system). Additive speech babble noise at various dBs has been applied on the far-field audio channel.

extracting information captured from the video sensors in the smart space. An example is ASR, the performance of which is known to improve in noise by inclusion of visual speech information extracted from the speaker mouth region – giving rise to audio-visual automatic speech recognition (AVASR) [20].

However, most of the work in the AVASR literature has concentrated on data with high visual quality containing frontal faces. In contrast, smart space environments, such as in CHIL, provide in general video data with varying head pose obviously not guaranteed to be frontal (see for example Fig. 1). These data can be obtained for example using a 3D tracking algorithm [10] to control one or more PTZ cameras.

In recent work [15], we have started investigating AVASR performance employing non-frontal data. As an initial step, we have opted to compare frontal vs. profile data in a “toy” experiment, where subjects are recorded simultaneously by two cameras in the two poses of interest, uttering connected-digit sequences – in order to keep data requirements for AVASR modeling manageable (see also Fig. 2). Our experiments demonstrate that significant speech information exists in the profile views, albeit much less than in frontal ones. For example, pure visual speech recognition degrades from 25.4% WER in the frontal data to 39.9% in the profile ones. Nevertheless, when combined with the audio signal by means of a simple feature fusion approach, the visual information remains beneficial – especially in the case of acoustic noise. This fact is depicted in Fig. 3, where significant improvements are recorded for both frontal and profile data. Interestingly, when combining the two views into a multi-view system, further gains can be obtained. Additional research is currently in progress to allow AVASR during within-utterance pose-variation and statistical modeling of visual speech across views.

8. CONVERSATIONAL INTERACTION

As already discussed in the Introduction, one of the motivations for developing robust far-field speech recognition technology is use in natural, untethered conversational interfaces. Such are of interest in a variety of smart space applications, for example distant-talking control of interactive television, as is the goal of the DICIT project [7].

To facilitate such applications, IBM Research has been working



Fig. 4. Example of simulated television set-top box and display, driven by conversational interaction, with electronic program guide information shown. The prototype is being developed as part of the DICIT project [7].

on a variety of conversational solutions focusing on entertainment and telematics applications on embedded devices. The systems are based on state-of-the-art dialog management technology and natural language understanding, enabling the user to converse naturally with the device and accomplish objectives easily and efficiently, without the need to navigate through complex menus or remember specific commands [21].

Natural language speech recognition is carried out using a statistical language model with embedded grammars for variable items of interest, such as song names, television programs, or destinations. Some of these embedded grammars get updated dynamically during the dialog session. Interpretation of the user requests is carried out using a multi-level statistical action classifier that operates on the speech recognition output. This approach makes the system robust to non-critical ASR errors.

Such conversational systems are implemented using the IBM *conversational interaction management architecture* (CIMA). CIMA provides a flexible architecture for multimodal dialog management and component integration. CIMA includes a general purpose state machine with a programmable interface for application domain specific dialog management. The base dialog strategy serves as a template for common dialog management functions, and these are customized by writing application-specific dialog logic. This is done using state chart XML (SCXML). CIMA also includes support for accepting asynchronous input from multiple devices and modalities – for example using touch in addition to voice to select from a disambiguation list [21]. An example of a CIMA-based conversational application, being developed for the purpose of the DICIT project [7], is depicted in Fig. 4.

9. SUMMARY

We have made significant advances in far-field speech technologies in smart spaces, exploiting the availability of multiple acoustic sensors. In particular, the state-of-the-art in these technologies has been presented as developed for the CHIL project, focusing on lectures and interactive seminars inside smart rooms. Although the domain remains challenging – as the reported results on acoustic scene analysis, speaker diarization, and automatic speech recognition demonstrate, it is our belief that further advances are possible. One possibility for achieving this goal is the use visual channel information, as the “toy” experiments reported for AVASR show. Our goal is to further improve robustness of these technologies and to expand to additional smart space environments, such as smart homes [6], with one application of interest being natural conversational interaction with domestic devices. One such example includes voice control of interactive television [7], based on state-of-the-art dialog and natural

language understanding technology.

10. REFERENCES

- [1] “CHIL: Computers in the Human Interaction Loop” [Online]. Available: <http://chil.server.de>
- [2] “AMI: Augmented Multi-Party Interaction” [Online]. Available: <http://www.amiproject.org>
- [3] “The NIST SmartSpace Laboratory” [Online]. Available: <http://www.nist.gov/smartspace>
- [4] J.G. Fiscus, J. Ajot, M. Michel, and J.S. Garofolo, “The Rich Transcription 2006 Spring meeting recognition evaluation,” in *Machine Learning for Multimodal Interaction*, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299, pp. 309–322, 2006.
- [5] R. Stiefelhagen and J. Garofolo (Eds.), *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop, CLEAR 2006*, LNCS vol. 4122, 2007.
- [6] “NETCARITY: Ambient technology to support older people at home” [Online]. Available: <http://www.netcarity.org>
- [7] “DICIT: Distant-talking interfaces for control of interactive TV” [Online]. Available: <http://dicit.fbk.eu>
- [8] M. Brandstein and D. Ward (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, 2001.
- [9] K. Nickel, T. Gehrig, H.K. Ekenel, J. McDonough, and R. Stiefelhagen, “An audio-visual particle filter for speaker tracking on the CLEAR’06 evaluation dataset,” in [5], pp. 69–80, 2007.
- [10] Z. Zhang, G. Potamianos, A.W. Senior, and T.S. Huang, “Joint face and head tracking inside multi-camera smart rooms,” *Signal, Image and Video Processing*, vol. 1, pp. 163–178, 2007.
- [11] D. Mostefa, N. Moreau, et al., “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms,” (In Press) *Journal of Language Resources and Evaluation*, 2008.
- [12] “The LDC Corpus Catalog,” Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. [Online]. Available: <http://www.ldc.upenn.edu>
- [13] J. Huang, E. Marcheret, K. Visweswariah, V. Libal, and G. Potamianos, “The IBM Rich Transcription 2007 speech-to-text systems for lecture meetings,” in *CLEAR 2007 and RT 2007 Evaluation Campaigns*, R. Stiefelhagen et al. (Eds.), LNCS vol. 4625, pp. 429–443, 2008.
- [14] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “CLEAR evaluation of acoustic event detection and classification systems,” in [5], pp. 311–322, 2007.
- [15] P. Lucey, G. Potamianos, and S. Sridharan, “A unified approach to multi-pose audio-visual ASR,” in *Proc. Conf. Int. Speech Comm. Assoc. (Interspeech)*, pp. 650–653, Antwerp, Belgium, 2007.
- [16] J.G. Fiscus, “A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER),” in *Proc. ASRU Workshop*, Santa Barbara, CA, pp. 347–352, 1997.
- [17] J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos, “The IBM RT07 evaluation systems for speaker diarization on lecture meetings,” in *CLEAR 2007 and RT 2007 Evaluation Campaigns*, R. Stiefelhagen et al. (Eds.), LNCS vol. 4625, 2008.
- [18] E. Marcheret, G. Potamianos, K. Visweswariah, and J. Huang, “The IBM RT06s evaluation system for speech activity detection in CHIL seminars,” in *Machine Learning for Multimodal Interaction*, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299, pp. 323–335, 2006.
- [19] J. Huang, M. Westphal, S. Chen, et al., “The IBM Rich Transcription Spring 2006 speech-to-text system for lecture meetings,” in *Machine Learning for Multimodal Interaction*, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299, pp. 432–443, 2006.
- [20] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent advances in the automatic recognition of audio-visual speech,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.
- [21] R. Balchandran, M. Epstein, M. Cmejrek, L. Rachevsky, M. Labsky, L. Seredi, and L. Ures, “An embedded conversational solution for telematics,” (submitted to:) *Ann. Meeting Assoc. Computational Linguistics: Human Lang. Techn.*, Columbus, OH, 2008.