AUTOMATIC SPEECH ACTIVITY DETECTION, SOURCE LOCALIZATION, AND SPEECH RECOGNITION ON THE CHIL SEMINAR CORPUS

Dusan Macho, Jaume Padrell, Alberto Abad, Climent Nadeu, Javier Hernando, [1]
John McDonough, Matthias Wölfel, Ulrich Klee, [2]
Maurizio Omologo, Alessio Brutti, Piergiorgio Svaizer, [3]
Gerasimos Potamianos, Stephen M. Chu [4]

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe, Karlsruhe, Germany
[3] ITC-IRST, Povo, Trento, Italy
[4] Human Language Technologies, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

http://chil.server.de *

ABSTRACT

To realize the long-term goal of ubiquitous computing, technological advances in multi-channel acoustic analysis are needed in order to solve several basic problems, including speaker localization and tracking, speech activity detection (SAD) and distant-talking automatic speech recognition (ASR). The European Commission integrated project CHIL, "Computers in the Human Interaction Loop", aims to make significant advances in these three technologies. In this work, we report the results of our initial automatic source localization, speech activity detection, and speech recognition experiments on the CHIL seminar corpus, which is comprised of spontaneous speech collected by both near- and far-field microphones. In addition to the audio sensors, the seminars were also recorded by calibrated video cameras. This simultaneous audio-visual data capture enables the realistic evaluation of component technologies as was never possible with earlier data bases.

1. INTRODUCTION

In the workspace of the future, a so-called "ambient intelligence" will be realized through the widespread use of sensors (e.g., cameras, microphones, directed audio devices) connected to computers that are unobtrusive to their human users. Towards this end of *ubiquitous computing*, technological advances in multi-channel acoustic analysis are needed in order to solve several basic problems, including speaker localization and tracking, speech activity detection (SAD) and distant-talking automatic speech recognition (ASR). The long-term goal is the ability to monitor speakers and noise sources in a real reverberant environment, without any constraint on the number or the distribution of microphones in the space nor on the number of sound sources active at the same time. This problem is surpassingly difficult, given that the speech signals collected by a given set of microphones are severely degraded by both background noise and reverberation. The European Commission integrated project CHIL, *Computers in the Human Interaction Loop*, aims to make significant advances in the three technologies mentioned above, and to integrate them in several technology demonstrators.

* This work was supported by the European Commission under the integrated project CHIL, "Computers in the Human Interaction Loop", contract number 506909. The CHIL Consortium is located at Fraunhoferstr. 1, 76131 Karlsruhe, Germany. The authors wish to thank Kai Nickel and Keni Bernardin for their assistance in collecting and labeling the video data described in this work.

Although significant advances have been achieved during the last decade in distant-talking ASR [1, Ch. 15], high performance systems can be developed only for small vocabularies where training and testing conditions are matched, and for situations where speaker position, head orientation and speaking style are more or less constrained. In such applications, the use of one or more microphone arrays turns out to be effective, thanks to their ability to acquire a higher quality signal than that provided by a single far-field microphone. The effectiveness of the microphone array, however, is dependent on the array geometry, on the related beamforming algorithms used to combine the signals from the several elements in the array, as well as on the other noise reduction and postfiltering techniques that can be integrated into the overall signal processing chain [1].

A second crucial aspect is the capability of the multi-channel signal processing to extract speech sequences from the given far-microphone signals. A few attempts have been made to exploit the redundancy in the far-microphone signals to improve speech activity detection [2, 3]. It is worth noting that speech activity detection plays an important role not only in the front-end of an ASR system, but also as a vital component of a speaker localization system and as a detector of speech or other auditory events (e.g., telephone ring, printer) that require interpretation by acoustic scene analysis algorithms.

The source localization problem has also received significant attention recently for applications that range from distant-talking interaction to videoconferencing to automatic surveillance. Many references can be found in the literature [1, 4]. The most common solutions are based on the adoption of the generalized cross-correlation methods [5] for time difference of arrival (TDOA) estimation, the generalized cross correlation (GCC) or phase transform (PHAT) [6] in particular. In such techniques, the speaker position is derived from a set of delay estimates computed across different microphone pairs. Besides a general weakness of all the proposed methods when dealing with highly reverberant environments, it is worth noting that most of the previous research activities did not address the localization of multiple acoustic sources, which represents one of the most fundamental issues in the scenarios addressed here.

This work presents the results of initial experiments on the CHIL seminar corpus conducted at four partner sites: ITC-IRST in Trento, Italy; Universität Karlsruhe (UKA) in Karlsruhe, Germany; Universitat Politécnica de Catalunya (UPC) in Barcelona,

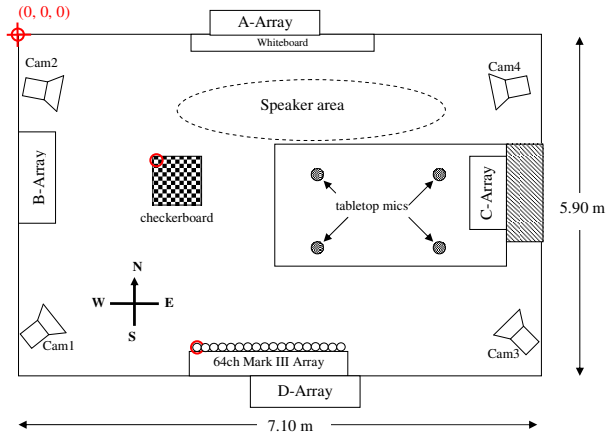

Fig. 1. The CHIL seminar room at the Universität Karlsruhe. The checkerboard is used for camera calibration.

Spain; and IBM Research in Yorktown Heights, USA. The paper is structured as follows: In Section 2, we describe the current and future data collection activities within the CHIL project. Most importantly, we describe why the CHIL seminar corpus provides a unique opportunity to simultaneously evaluate several technologies, including the audio technologies mentioned above. In Section 3, we present our initial results in SAD, as well as the SAD metrics that have been developed for use within the CHIL project. Our activities in source localization are presented in Section 4, and those in ASR in Section 5. The final section presents our conclusions and plans for future work.

2. DATA COLLECTION AND LABELING

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

The seminar speakers were recorded with a Sennheiser close-talking microphone (CTM), as well as two linear eight-channel microphone arrays. The sample rate of the recordings was 16 kHz with 16 bits per sample. In addition to the audio data capture, the seminars were simultaneously recorded with four calibrated video cameras at a rate of 15 frames per second.

The data from the CTM was manually segmented and transcribed. The data from the microphone arrays was labeled with speech and non-speech regions. Prior to the start of the seminar, the video cameras had been calibrated as in [7]. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.67 sec. Using these hand-marked labels, the true position of the speaker's head in three dimensions was calculated using the technique developed by Focken and Stiefelhagen [8]. These positions are accurate to within approximately 10 cm.

Given the naturalness of the setting and the variety of sensors, the CHIL seminar corpus is a very useful test bed for evaluating many audio and video technologies concurrently: The manually transcribed audio data is useful for traditional ASR evaluations. The availability of the true speaker positions derived from the video labels enables the evaluation of source localization techniques under far more realistic conditions than reported in prior work [1, Ch. 8-10]. Additionally, the door slams and other acoustic events that occur during the recordings are challenging for SAD systems, and of interest for acoustic event classification and acoustic scene analysis. For the initial evaluations, two segments of five minutes length were chosen from seven distinct seminars. This provided a total of 70 minutes of speech material.

Since the Fall of 2003, the CHIL seminar room at UKA has been enhanced with a variety of new sensors as shown in Fig. 1. We have obtained several Countryman E6 close-talking microphones, which are preferable to the Sennheiser CTMs for this application because they do not obscure the speaker's lips, thereby allowing for audio-visual ASR [9]. A 64-channel Mark III microphone array developed at the National Institute of Standards and Technology (NIST) has been installed on the wall facing the seminar speaker. In addition, T-shaped microphone arrays with four elements each have been installed on the four walls of the seminar room, to allow accurate three-dimensional source localization. On the work table in the seminar room are four Shure Microflex table-top microphones. For future seminars, all audio sensors will be recorded at 44.1 kHz with 24 bits per sample. The higher sample rate is preferable to permit more accurate TDOA estimation, while the higher bit depth is necessary to accommodate the large dynamic range of the far-field speech data. The CHIL consortium plans to make a portion of the data from subsequent seminars available to NIST for use in upcoming meeting evaluations.

3. SPEECH ACTIVITY DETECTION

In the context of CHIL, it is expected that several speech technologies such as ASR, speaker localization, and speaker identification will benefit from improvements in SAD. Thus, the metrics for the SAD evaluation adopted within the CHIL project have been designed so as to avoid any biasing towards one of these technologies. We use the following three metrics in this evaluation of SAD systems:

- *mismatch rate* (MR) = time of incorrect decisions / time of all utterances.
- *speech detection error rate* (SDER) = time of incorrect decisions at speech segments / time of speech segments. This metric assesses the SAD performance on the speech portions of the signal.
- *non-speech detection error rate* (NDER) = time of incorrect decisions at non-speech segments / time of non-speech segments. This statistic measures how the SAD performs on the non-speech portions of the signal.

For all metrics, an average is evaluated over all utterances.

Table 1 presents the results of our initial SAD experiments. This section summarizes the specific techniques investigated by each site.

ITC-IRST: At this moment, speech activity detection is performed just on one far microphone as follows. The maximum energy of the current frame is compared to the current threshold value to detect speech intervals. The threshold is dynamically updated by calculating it as a nonlinear average value of energy amplitude during speech absence. More precisely, the most recent energy values of non-speech intervals are buffered and resorted in ascending order. The average value of the lower fraction (e.g. the lower half) of the reordered buffer is taken as the new current threshold. Potential speech segments are determined when the threshold is

Site	MR	SDER	NDER
ITC-IRST	17.33	10.06	43.00
UPC	12.56	11.40	14.99

Table 1. Speech activity detection (SAD) results on the CHIL corpus by two sites, expressed %, using the metrics of Section 3.

exceeded. Speech intervals are detected only when the following conditions are satisfied by the energy with respect to the dynamic threshold:

- the detected candidate segment is long enough;
- inside the candidate segment, energy values are below the threshold only for short intervals;
- a sufficiently large percentage of frames inside the candidate segment is over the threshold.

The experiments on the given seminars showed that an energy-based algorithm can provide acceptable performance. However, future work will focus on developing multi-channel SAD; in fact, when dealing with more severe noisy conditions, we expect to achieve satisfactory performance only by exploiting the redundancy in the multichannel signals.

UPC: In the UPC system for SAD, frequency filtered (FF) log filter-bank energy features [10] are calculated on a frame-by-frame basis from the input audio signal. Signal frames are 30 ms long and the frame shift is 10 ms. Delta and delta-delta values are appended to the static FF features, thereby forming a 43-dimensional feature vector: 14 static + 14 delta + 14 delta-delta + 1 delta Energy. The initial size of the FF feature vector is then reduced to a single measure m_1 by applying linear discriminant analysis (LDA) [11].

The FF+LDA measure m_1 is a float number, and it has to be post-processed in order to obtain a binary output. Establishing a threshold on m_1 to reach the speech/non-speech decision gave very poor results. Instead, the well-known C4.5 algorithm was used to train a decision tree classifier on m_1 values. Besides the current m_1 measure, the classifier was also provided with the 15 previous and 15 subsequent m_1 values. From the resulting 31 m_1 values, only seven were automatically selected, which helped simplify the tree classifier. More details can be found in [11].

4. SOURCE LOCALIZATION

The results of our initial source localization experiments are shown in Table 2. A summary of the techniques used by each site is presented next.

ITC-IRST: The location algorithm uses only four of the 16 signals acquired by the linear microphone array. Since the array is composed of two subarrays, each one including eight synchronous channels, two microphone pairs were selected, namely (1,8) and (9,16), and independently used to estimate the direction of arrival with respect to their central points. Time delay estimation is performed by means of the cross-spectrum phase (CSP, PHAT, or GCC). The intersection of the curves describing the potential source locations (i.e. lines approximating the more accurate hyperbolas) is assumed, at each frame, as the candidate source location, provided that it corresponds to an area physically located inside the room. Due to the array geometry, source location is possible only in a two-dimensional plane at fixed height. Furthermore, the distance of the source from the array cannot be estimated with high accuracy. On the other hand, the azimuth with respect to the center of the array is quite accurate, as is apparent from Table 2. The amplitude of the peak of the CSP function can be used as a reliability criterion of the estimate: it is proportional to the coherence of the direct wavefront produced by the acoustic source. Note also that a speaker cannot be accurately located, if not facing the array.

Site	Azimuth (^{o})	Depth (cm)
UKA (Benesty)	12.3	91.2
UKA	10.9	95.5
ITC-IRST	9.75	98.1
UPC	9.05	73.5

Table 2. Source localization RMS errors by various sites.

UKA: Six pairs of microphones from each of the two 8-channel arrays were used for source localization. These pairs were chosen as (1,3), (2,4), (3,5), (4,6), (5,7), and (6,8) for the first microphone array, and similarly for the second. Initial experiments at UKA were based on the GCC approach [5, 6], with the usual parabolic interpolation between samples to improve on the accuracy of the TDOA estimates. For each pair of microphones, the azimuth to the speaker was estimated, for a total of six azimuth estimates per 8-channel array. Next, the intersection point between each pair of azimuths was calculated. Those pairs of estimates for which the azimuths did not intersect were discarded, in which case the speaker location was not updated. The final estimate of the speaker position was obtained by averaging the intermediate estimates from each azimuth pair. Thereafter a Kalman filter was used to smooth the series of estimates. This filtering was based on heuristics used to set the Kalman parameters: The Kalman gain was set very low if the azimuth estimate placed the speaker outside of the physical room. If the speaker's position could not be detected for a given frame, the Kalman estimate was not updated.

In addition to the classic GCC method, we also evaluated the TDOA technique recently proposed by Benesty, et al. [4]. As yet we have implemented no interpolation for this new method. Nonetheless this technique provided azimuth and range estimates comparable to those obtained with the GCC plus interpolated TDOA estimates, and hence is of interest for further study.

UPC: The UPC speaker localization system uses the generalized cross-correlation technique with phase weighting (GCC-PHAT) [6] to estimate the direction of sound arrival for a given array. The original 16-microphone linear array was split into four independent sub-arrays assigning four successive microphones to each sub-array. The two extreme microphones of each sub-array are 122.1 mm apart, which means the so called far-field condition is satisfied for such a microphone pair. Thus, the delay between the two signals from this microphone pair can be related with the direction of arrival of sound, assuming a direct sound wave is available. In the present case, four independent directions of arrival are obtained. The delays based on other combinations of microphones have been neglected because nearer pairs present a reduced resolution and usually do not help to enhance the measure.

In these tests, speaker localization was performed only on the portions of signal labeled as speech by the FF+LDA speech activity detector; see Section 3. Additionally, no source tracking model was considered, and each source position was estimated independently from the previous ones. The source position was estimated based on the intersections of the four independent direction estimates obtained from the four sub-arrays. The direction estimates with confidence lower than 0.1 were excluded. Confidence values are functions of the relation between the main and the secondary peak of the cross-correlation estimate. Also, the 15 degree median filter margin was used to eliminate possible outliers. The final azimuth was determined as a connection between the middle of the entire array and the estimated source position. In the case of out-of-room estimates, the azimuth was computed as the median of the

four angle estimates. The speaker position was then estimated as the intersection of the vertical plane defined by this azimuth and a plane parallel to the whiteboard wall (see Fig. 1) and positioned inside the room at a 1 m distance from it.

5. AUTOMATIC SPEECH RECOGNITION

The CHIL seminar data present significant challenges to both modeling components used in ASR, namely the language and acoustic models. With respect to the former, the currently available CHIL data primarily concentrate on technical topics with focus on speech research. This is a very specialized task that contains many acronyms and therefore is quite mismatched to typical language models currently used in the ASR literature. Furthermore, large portions of the data contain spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. On the acoustic modeling side, and in addition to the latter difficulty, the seminar speakers exhibit moderate to heavy German accents in their English speech. The above problems are compounded by the fact that, at this early stage of the CHIL project, not enough data are available for training new language and acoustic models matched to this seminar task, and thus one has to rely on adapting existing models that exhibit gross mismatch to the CHIL data. Clearly, these challenges present themselves in both close-talking microphone data, as well as the far-field data captured using the microphone arrays, where of course they are exacerbated by the much poorer quality of the acoustic signal. Although the results presented here are only for the close-talking microphone (CTM), the CHIL consortium is actively investigating the use of microphone arrays to enhance ASR performance when no CTM is available.

IBM: The ASR system developed by the IBM team uses an interpolated language model and a wideband acoustic model on MFCC speech features, with per-speaker supervised adaptation. In more detail, three language model components are used: The first is built on 3M tokens from the well known "switchboard" task, the second is built on 1M tokens from Eurospeech conference papers, and the third is a tri-gram model of the CHIL development set. The final language model is an interpolation of these three, has a 16k word vocabulary, and achieves a perplexity of 147 on the CHIL evaluation set, with a 0.8% out-of-vocabulary rate. For acoustic modeling, a wideband hidden Markov model is used, with approximately 3.4k context-dependent states and 53k Gaussian components. The model has been originally trained on 200 hrs of mostly accented speech by 780 speakers from the MALACH task [12], and subsequently adapted by supervised MAP adaptation on the entire development set, followed by per-subject MLLR adaptation [13]. The resulting word error rate is 37.3%.

UKA: As no CHIL data is presently available for training ASR systems, the acoustic model was trained on *Broadcast News* and merged with the close-talking channel of the meeting corpus described in [14]. This provided a total of 300 hours of training material.

The speech data was sampled at 16kHz. Speech frames were calculated using a 10 ms Hamming window. For each frame, 13 mel-cepstral coefficients were calculated. Thereafter, linear discriminant analysis was used on seven adjacent frames of speech to reduce the final feature size to 42. The baseline model consisted of 300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks.

Several steps of corpus and speaker adaptation were applied to the speaker independent acoustic model: MLLR adaptation was first used to adapt the model to the seminar corpus. Thereafter a second phase of MLLR was used to adapt the model to the individual seminar speakers. On the seminar data, the final system achieved a word error rate of 41.6% on the CTM channel.

6. CONCLUSIONS

The CHIL consortium is dedicated to making significant advances in the state-of-the-art for the automatic speech activity detection, acoustic source localization and speech recognition technologies discussed in this work, thereby moving closer to the long-term goal of ubiquitous computing. In addition to the data collection and research activities described here, the CHIL project is also developing data labeling and transcription practices, as well as evaluation metrics to enable improvements in the state-of-the-art to be reliably measured. Towards this end, the CHIL consortium is working actively with the National Institute of Standards and Technology in the USA. Moreover, the consortium is actively seeking the participation of external research sites, willing to evaluate their technology on the seminar and meeting data collected by the project.

7. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer Verlag, 2000.

[2] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. ICASSP*, 1990.

[3] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. Eurospeech*, 2003.

[4] J. Chen, J. Benesty, and Y. A. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Proc.*, 11:540–557, 2003.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics Speech Signal Proc.*, 24:320–327, 1979.

[6] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP*, 1994.

[7] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis Machine Intell.*, 22:1330–1334, 2000.

[8] D. Focken and R. Stiefelhagen, "Towards vision-based 3-d people tracking in a smart room," in *Proc. Int. Conf. Multimodal Interfaces*, 2002.

[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the recognition of audio-visual speech," *Proc. IEEE*, 91:1306–1326, 2003.

[10] C. Nadeu, D. Macho, and J. Hernando, "Frequency and time filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, 34:93–114, 2001.

[11] J. Padrell, D. Macho, and C. Nadeu, "Robust speech activity detection using LDA applied to FF parameters," in *Proc. ICASSP*, 2005.

[12] B. Ramabhadran, J. Huang, and M. Picheny, "Towards automatic transcription of large spoken archives – English ASR for the MALACH project," in *Proc. ICASSP*, 2003.

[13] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9:171–185, 1995.

[14] F. Metze, C. Fügen, Y. Pan, T. Schultz, and H. Yu, "The ISL RT-04s meeting transcription system," in *Proc. Rich Transcr., Spring Meeting Recog. Wksp.*, 2004.