

Automatic Speech Recognition and Speech Activity Detection in the CHIL Smart Room

Stephen M. Chu, Etienne Marcheret, and Gerasimos Potamianos

Human Language Technologies, IBM T. J. Watson Research Center,
Yorktown Heights, New York 10598, USA
{schu, etiennem, gpotam}@us.ibm.com

Abstract. An important step to bring speech technologies into wide deployment as a functional component in man-machine interfaces is to free the users from close-talk or desktop microphones, and enable far-field operation in various natural communication environments. In this work, we consider far-field automatic speech recognition and speech activity detection in conference rooms. The experiments are conducted on the smart room platform provided by the CHIL project. The first half of the paper addresses the development of speech recognition systems for the seminar transcription task. In particular, we look into the effect of combining parallel recognizers in both single-channel and multi-channel settings. In the second half of the paper, we describe a novel algorithm for speech activity detection based on fusing phonetic likelihood scores and energy features. It is shown that the proposed technique is able to handle non-stationary noise events and achieves good performance on the CHIL seminar corpus.

1 Introduction

Speech is one of the most effective means of communication for humans. It is therefore an essential modality in multimodal man-machine interactions. Speech-enabled interfaces are desirable because they promise hands-free, natural, and ubiquitous access to the interacting device. Much progress in speech technologies has been made in recent years. However, the majority of the successful applications to date, e.g. call-center automation, broadcast news transcription, and desktop dictation, all but confine the speaker to a nearby microphone.

An important step to bring speech technologies into wide deployment as a functional component in man-machine interfaces is to free the users from close-talk or desktop microphones, and enable far-field operation in various natural communication environments. In this work, we consider two related aspects of speech technologies, automatic speech recognition (ASR) and speech activity detection (SAD), in a far-field scenario. The experiments are carried out on the *smart room* platform provided by the European Commission integrated project: Computers in the Human Interaction Loop (CHIL). The CHIL *smart room* is a conference room equipped with

multiple audio and visual sensors to facilitate intelligent multimodal interactions. On the acoustic side, the main far-field input is provided by linear microphone arrays. In addition to the linear arrays, T microphone arrays, desktop microphones, as well as close talk microphones may also be available. On the visual side, video cameras with wide-angle lens provide coverage of the entire space; active pan-tilt-zoom (PTZ) cameras allow close-up shots of the subjects in the room. In this paper, we shall concentrate on experiments using only the audio sensors. Note that the concept of *smart room* is not restricted to the context of CHIL. Research results obtained here can be readily extended to other interactive environments where multimodal *ambient intelligence* is sought.

Far-field ASR in conference rooms is a challenging task. Because of reverberation, it would be extremely difficult for a single distant microphone to give satisfactory recognition results, unless the training and testing conditions in terms of acoustic environment, speaker location, and microphone placement are perfectly matched. A promising way of improving far-field ASR performance is to use an array of microphones [1]. Microphone arrays are able to acquire higher quality signal because of the high directivity achieved by beam-forming algorithms, which typically assume that the geometry of the array is regular and known. Ultimately, we would like to have systems that can take advantage of an arbitrary set of distributed acoustic sensors. In such cases, it becomes necessary to look beyond the conventional beam-forming algorithms and consider alternative ways to fuse the information provided by the multiple microphones. Instead of operating in the signal level, it is also possible to carry out the fusion in the hypothesis domain. Through the subsequent experiments, we aim to gain some preliminary understandings of the potentials of the proposed high-level fusion by comparing the recognition performance of single microphone, beam-forming, and hypothesis combination.

Speech activity detection is a crucial aspect in *smart room* applications. Not only does it play an essential role as a front-end step to the ASR process, but also provide important cues to speaker localization and acoustic scene analysis algorithms. Most existing SAD algorithms build classifiers directly on the features extracted from the acoustic signal [2]-[4]. The features may be straight forward energy coefficients or more complex frequency domain representations. The common choice of classifiers ranges from adaptive thresholds to linear discriminants, regression trees, and Gaussian mixture models (GMM). In general, energy-based speech detection is computationally efficient and simple to implement, but lacks robustness to noise. Although performance can be improved by using adaptive thresholds or appropriate filtering of the energy estimates, it remains difficult to address non-stationary noise effectively. It has been shown that frequency based speech features, such as Mel-frequency cepstral coefficients (MFCC), are necessary for further improvement in noise robustness. In this paper, we propose to employ such features indirectly, through the acoustic model that is assumed to have generated them. The resulting acoustic phonetic features are extracted based on the phonetic class conditional MFCC observation vector likelihoods by the acoustic model, and are used to augment baseline energy based features. The two types of features are fused and subsequently considered for speech/silence detection using a GMM classifier.

The rest of the paper is organized as follows. In the next section, we first describe the sensor configuration in the CHIL smart room and the CHIL seminar corpus. Section 3 addresses the development of speech recognition systems for the seminar transcription task. In particular, we look into the effect of combining parallel recognizers in both single-channel and multi-channel settings. Section 4 introduces the proposed SAD algorithm based on fusing phonetic likelihood scores and energy features. Instead of putting ASR and SAD results together in a separate section, we shall cover the experimental results in their respective sections immediately after the algorithms are introduced. Finally, conclusions and future work are discussed in section 5.

2 CHIL Seminar Corpus

The speech data used in the experiments are collected in the CHIL smart room at the Universität Karlsruhe in Karlsruhe, Germany. The corpus consists of two parts.

The first part was recorded in the fall of 2003 and made available for the June 2004 evaluation, thus shall be referred to as the June'04 dataset in the remainder of the paper. The content of the speech data contains technical seminars given by students of the university. There are seven seminars and seven distinct speakers with varying degree of fluency in English. During each session, both close talk and far-field recordings are made concurrently, the former through a Sennheiser close talk microphone (CTM), and the latter by two linear eight-channel microphone arrays. The signal is sampled at 16 KHz with 16-bit resolution. The total duration of the recording is 137 minutes. The data is further partitioned into two subsets: a development set with 68 minutes and 3971 utterances, and a test set with 69 minutes and 3077 utterances. All seven speakers appear in both of the subsets.

The second part of the corpus was recorded after a series of hardware updates were made to the smart room. In particular, the Sennheiser CTM is replaced by Countryman E6 microphones; and a new 64-channel Mark III microphone array developed by the National Institute of Standards and Technology (NIST) is now providing the far-field data. The signal is sampled at 44.1 KHz with 24 bits per sample. The schematic of the updated CHIL seminar room is shown in Figure 1. The data was made available for the January 2005 evaluation, and shall be referred to as the Jan'05 dataset here on.

There are five seminars and five speakers in this collection. Among the speakers, one also appeared in the June'04 dataset. The development set contains 46 minutes of speech segmented into 1395 utterances; the test set contains 133 minutes of data and 1764 utterances. Two speakers are found in the development set, including the one shared with the June'04 dataset. In addition to those two speakers, the test set also has three speakers unseen in the development set.

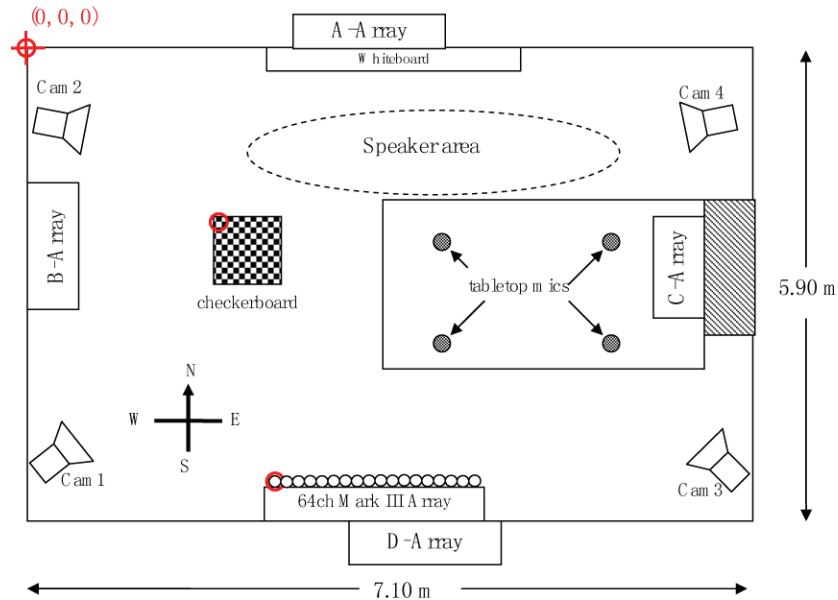


Fig. 1. Schematic of the CHIL smart room at the Universität Karlsruhe. The environment as shown was used to collect the Jan'05 dataset

Both the June'04 dataset and the Jan'05 set were manually segmented and transcribed. For the far-field data, speech and non-speech regions are also labeled to provide ground truth for the SAD experiments.

3 ASR Experiments

To develop an effective speech recognition system for the CHIL seminar transcription, the following characteristics of the task must be considered.

First, the amount of available training data is limited. Given that the total duration of the development set is less than two hours, it is therefore unfeasible to train a large vocabulary continuous speech recognizer (LVCSR) from scratch. A more plausible approach is to start from a set of acoustic models trained on other much larger speech corpora, and then refine these models using the CHIL development data through adaptation.

Second, the smart room environment, the seminar scenario, and the mostly European speaker set make this collection distinct from most of the existing large speech corpora/tasks. Therefore, a system based solely on one of the existing databases is unlikely to give the optimal performance. In our work, we aim to take advantage of different speech datasets by running three systems developed separately on three very different corpora in parallel, and combining the word hypotheses generated by the systems to give the final output.

Lastly, because the domain of the speech content is well defined, further reduction in recognition error can be achieved by developing a domain specific language model. In the June 2003 CHIL evaluation [5], we first experimented using the text from in-domain technical publications for language model development. The merit of the approach was clearly demonstrated in the evaluation results. In this work, we shall continue to use the same method.

3.1 System Description

The three parallel recognition systems considered here are: (1) a wide-band dictation system, (2) a wide-band dialogue system with German accent [6], and (3) a narrow-band conversational system. The front-end specifications and the configurations of the acoustic models are summarized in Fig. 2.

In both the first and the second system, supervised *maximum a priori* (MAP) adaptation was performed using the development set (joint set of the June'04 and Jan' 05 development data). In the third system, supervised MLLR was applied. Note that these adaptations were speaker-independent. In addition to MLLR, the third system also performs speaker-dependent VTLN and feature-level minimum phone error (fMPE) [7] adaptation. These operations are carried out at run-time and are unsupervised.

The language model is developed on three datasets: the CHIL development set, a three million word set from the Switchboard corpus, and a one million word set de-

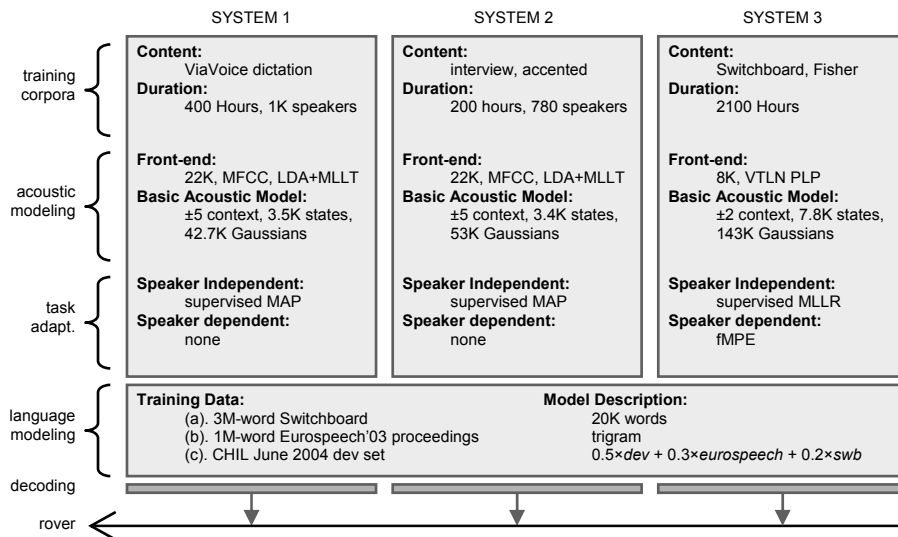


Fig. 2. The IBM CHIL ASR system is composed of three parallel large vocabulary speech recognizers. The acoustic models in each of the individual systems are trained on different speech corpora and adapted to the CHIL seminar transcription task using the development data. The three systems share the same language model. The word hypotheses are combined using ROVER to give the final recognition result

rived from Eurospeech '03 proceedings using automated PDF to ASCII conversion. The 20k vocabulary contains the following words: all words in the CHIL development set, words in the Switchboard set with 5+ counts, words in the Eurospeech set with 2+ counts. A trigram model is built on each of the three corpora with modified Kneser-Ney smoothing. The final language model is obtained through the following interpolation,

$$0.5 \times chil + 0.3 \times eurospeech + 0.2 \times switchboard \quad (1)$$

The language model is shared by all three recognition systems. During testing, the input speech is first decoded separately by the three individual systems. Then the outputs are combined using the NIST ROVER [8] system to produce the final hypothesis.

3.2 Experimental Results

To establish appropriate benchmarks for the multi-channel far-field experiments, we first test the recognition system with the CTM recording and data from a single channel in the microphone array. Adaptations are applied to the basic acoustic models using the development data from the corresponding channels. With respect to the given system, the CTM result should give an upper bound for the multi-channel far-field performance; while the single channel result should provide an estimate of the baseline.

We also compare the performance from the current recognition system with an earlier system (06/04), which is essentially *system 2* in Fig. 2, adapted with only the June'04 development set. The benchmark results are summarized in Table 1.

Table 1. Recognition results in word error rate for the close talk and far field microphones. The results from the 06/04 system and the 01/05 system are compared

Test set : system	close talk	far field
06/04 : 06/04	35.1%	64.5%
06/04 : 01/05	31.5%	63.2%
01/05 : 01/05	36.9%	70.8%

The results clearly show the difficulty posed by far-field ASR. In all three cases, the word error rates (WER) for the single far-field microphone are approximately doubled comparing with the corresponding CTM performance. The results also confirm the benefit of parallel decoding using diverse acoustic models. On the same June'04 test set, the parallel system is able to reduce the WER from 35.1% to 31.5%, which translates to a 10.3% relative reduction. The differences are illustrated in Fig. 3

Two multi-channel far-field ASR experiments are carried out. Both use the same 16 channels of microphone array data found in the June'04 dataset. In the first experiment, beam-forming is applied to the multiple outputs of the microphone array to generate a single channel of acoustic signal. The signal is then passed to the ASR system for adaptation and recognition. In the second experiment, each channel of the

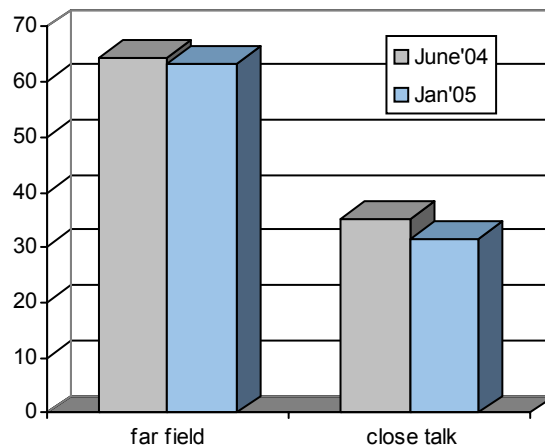


Fig. 3. Benchmark results of the ASR systems using data from CTM and single channel far-field microphone. The Jan'05 system employing three parallel acoustic models gives superior recognition performance to the June'04 system

microphone array is first processed independently. For a given utterance, 16 word-level hypotheses with word confidence scores are generated. These hypotheses are then combined using the ROVER program to give the final word sequence.

The results of the multi-channel ASR experiments are shown in Fig. 4. In addition to the beam-forming and hypothesis-combination results, the single-channel WERs for all 16 channels are also listed. The single-channel WER is used to create a ranked list of the channels; and hypothesis integration experiments are repeated for top n channels for $n = 1 \dots 16$.

Decoding the beam-formed signal gives a WER of 58.5%. This is a clear improvement over the single-channel results, which has an average WER of 64% approximately. The lowest WER achieved by hypothesis integration is 59.5%, which is not far behind the beam-forming performance. This is indeed encouraging considering the facts that the beam-forming algorithm explicitly relies on the known geometry of the array, and that the hypothesis integration approach is able to attain more than 80% of the gain without the advantage of this prior knowledge. Therefore, in the situation where the sensor configuration is not known, hypothesis integration can serve as a viable alternative to signal domain algorithms.

One disadvantage of running multiple recognition engines is the increased demand on computational resource. However, from Fig. 4, it can be observed that in this particular experiment, most of the gain of occurs when the first few channels are added. If this observation is true in general, then the computational load of the approach can be significantly reduced. Further experiments are still required.

In essence, array-processing is an information fusion problem. The fusion problem arises when different observations about the same underlying process are available from multiple sources. The goal then is to find the optimal way to integrate the sources so that the generating process can be inferred. In a conventional beam-

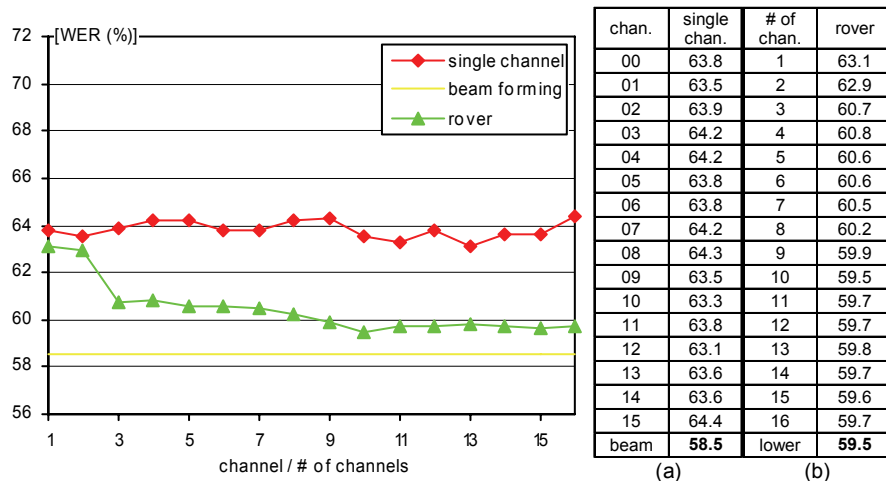


Fig. 4. Comparing the recognition results on the microphone array data. (a). results for each individual channel and the beam-forming signal; (b). Rover top n channels according to (a)

forming algorithm, the information is integrated at a very low-level in the signal domain; whereas in the rather straight-forward alternative approach taken in the second experiment, the fusion takes place at a much higher level in the hypothesis space. In fact, it is worthwhile to look into mechanisms for multi-channel ASR that permit intermediate level fusion.

4 Speech Activity Detection Experiments

The proposed SAD system operates on two types of features, the energy features generated directly from the signal, and the acoustic phonetic features defined from observations generated by the ASR acoustic model. The energy features are five-dimensional vectors computed from band-passed signals. Conceptually, the five components track the energy envelope of the waveform with different sensitivities, thus providing an evolving statistics about the signal. The details of the computations can be found in [9], and are omitted here for brevity. The emphasis will be given to the acoustic phonetic features proposed in this work.

4.1 Acoustic Phonetic Features for SAD

The acoustic phonetic feature space employed for speech activity detection is derived from the acoustic model used for ASR. The acoustic model is generated from partitioning the acoustic space by context-dependent phonemes with the context defined in this work as plus and minus five phonemes, cross-word to the left only. The context-dependent phoneme observation generation process is modeled as a GMM within the hidden Markov model (HMM) framework, and in typical large-vocabulary ASR sys-

tems, this leads to more than 40k Gaussian mixture components. Calculating all HMM state likelihoods from all Gaussians at each frame would preclude real-time operation. Therefore, we define a hierarchical structure for the Gaussians, where it is assumed that only a small subset of them is significant to likelihood computation at any given time. The hierarchical structure takes advantage of the sparseness by surveying the Gaussian pool in multiple resolutions given an acoustic feature vector \mathbf{x} . As part of the training process, the complete set of available Gaussian densities is clustered into a search tree, in which the leaves correspond to the individual Gaussians, and a parent node is the centroid of its children for a defined distance metric. At the bottom of this tree resides a many-to-one mapping, collapsing the individual Gaussians to the appropriate HMM state. Therefore, the HMM state conditional likelihood of a given observation vector \mathbf{x} at time t is computed as

$$p(\mathbf{x} | s) = \sum_{g \in G(s)} p(g | s) p(\mathbf{x} | g) \quad (2)$$

where $G(s)$ is the set of Gaussians that make up the GMM for state s . Traversing the tree will yield a subset of active Gaussians, denoted by Y . Based on Y and the many-to-one mapping, the conditional likelihood of a state is approximated as

$$p(\mathbf{x} | s) = \max_{g \in Y \cap G(s)} p(g | s) p(\mathbf{x} | g) \quad (3)$$

If no Gaussian from a state is present in Y , a default floor likelihood is then assigned to that state.

To define the acoustic phonetic space used for speech activity detection, we apply an additional many-to-one mapping to the pruned result of the hierarchical tree. This function maps phonemes into three broadly defined classes: (i) the pure silence phoneme, trained from non-speech; (ii) the disfluent phonemes, which are noise like phonemes, namely the unvoiced fricatives and plosives, i.e., the ARPAbet subset $\{/b/, /d/, /g/, /k/, /p/, /t/, /f/, /s/, /sh/\}$; and (iii) all the remaining phonemes, such as the vowels and voiced fricatives. The three classes will be denoted by c_1 , c_2 , and c_3 . From the acoustic feature \mathbf{x} , which is used to traverse the acoustic model hierarchy, we can form the speech detection class posteriors for the three speech detection classes as,

$$P(c_i | \mathbf{x}) = \frac{\sum_{g \in Y \cap G(c_i)} p(\mathbf{x} | g) p(g | c_i)}{\sum_{i=1}^3 \left\{ \sum_{g \in Y \cap G(c_i)} p(\mathbf{x} | g) p(g | c_i) \right\}} \quad (4)$$

And $G(c_i)$ is the set of Gaussians defined by the mapping from the phonemes to the speech detection class c_i .

Pruning at each level of the hierarchical acoustic model is accomplished by using a threshold relative to the maximum scoring likelihood for that level. As a result, the sharper the drop-off in Gaussian likelihoods, the more aggressive the pruning becomes. Therefore, both SNR and the phoneme being pronounced impact the pruning.

Features extracted from vowels and other voiced phonemes will result in more aggressive pruning than unvoiced fricatives, plosives and silence phonemes. This pruning will remain relative to SNR, with increasing SNR resulting in an overall more aggressive pruning.

The above observation results in additional speech detection features based on class-normalized Gaussian counts. Let's denote the number of Gaussians after hierarchical pruning that map to speech detection class c_i , n_{c_i} , and consider the normalized counts

$$\bar{n}_{c_i} = n_{c_i} / \sum_{j=1}^3 n_{c_j}, \text{ for } i = 1 \dots 3 \quad (5)$$

as additional features. Combining (4) and (5), we obtain the six-dimensional acoustic phonetic feature vector

$$\mathbf{v}_a = \begin{bmatrix} v_{a1} \\ v_{a2} \\ v_{a3} \end{bmatrix}, \mathbf{v}_{ai} = \begin{bmatrix} \log(P(c_i | x)) \\ \log(n_{c_i}) \end{bmatrix}. \quad (6)$$

Finally, the five-dimensional energy features and the acoustic phonetic features are concatenated to form an 11-dimensional feature vector.

4.2 Training and Classification

The joint energy and acoustic phonetic feature vectors are projected to an eight-dimensional feature space using PCA, on which a three-class GMM classifier is built. For each class eight Gaussian densities with diagonal covariance matrices are used.

The class labels for the training data is inferred through Viterbi alignment using the ASR acoustic model and the transcripts of the corresponding utterances. Once the phone-level alignment is computed, the class identity of a frame can be readily obtained via the phoneme-to-class mapping described earlier.

During classification, the likelihood scores of a frame given the three GMMs are first evaluated. The scores are then smoothed over time to give the classification result. As a final step, the three classes are mapped to speech/non-speech according to the following rules.

1. $c_1 \rightarrow$ non-speech
2. $c_3 \rightarrow$ speech
3. $c_2 \rightarrow$ speech, if the one of the two neighboring frames is c_1 , and the other is c_3 ; otherwise $c_2 \rightarrow$ non-speech

This mapping allows the system to correctly handle both consonant-vowel-consonant transitions and non-stationary noise events.

4.3 Experimental Results

The far-field performance of the SAD system is evaluated using the first channel of the linear microphone array in the CHIL June'04 dataset.

The basic acoustic models used to compute the acoustic phonetic features are the same as the ones in *system 1* described in the ASR experiments. Specifically, they consist of 3.5K HMM states and 43K Gaussian mixtures, trained on 400 hours of data from 1000 speakers. The acoustic models use 40-dimensional features derived from 24-dimensional MFCCs through LDA/MLLT. Supervised MAP adaptation is applied on top of the basic acoustic models using the development set.

Three metrics are used to evaluate the SAD performance. They are defined as follows,

- Speech detection error rate (SDER) = time of incorrect decisions at speech segments / time of speech segments
- Non-speech detection error rate (NDER) = time of incorrect decisions at non-speech segments / time of non-speech segments
- Average detection error rate (ADER) = (SDER + NDER) / 2.

In our experiments, the operating points of the SAD are chosen such that the following condition is satisfied,

$$|\text{SDER} - \text{NDER}| / (\text{SDER} + \text{NDER}) \leq 0.1 \quad (7)$$

The SAD results using both the basic and the adapted acoustic models are shown in Table 2. As expected, the performance of the adapted system is significantly better than the system using baseline acoustic models.

Table 2. Speech activity detection results on the June'04 CHIL seminar dataset. The system uses joint energy and acoustic phonetic features and Gaussian mixture models for classification

metric	baseline	MAP adaptation
SDER	16.70%	10.01%
NDER	16.43%	11.92%
ADER	16.57%	10.96%

The reported performance of the adapted system was superior to all five other systems evaluated by the CHIL consortium [10], achieving 4% to 36% relative ADER reduction. All other submitted systems built classifiers directly on the energy or other speech features. The results clearly demonstrate the strength of the acoustic phonetic based approach to speech activity detection.

5 Conclusions

In this work, we consider far-field automatic speech recognition and speech activity detection in the CHIL smart room. We look into the effect of combining parallel recognizers in both single-channel and multi-channel settings for far-field ASR.

Experiments show that for microphone array, word-level hypothesis combination is able to achieve recognition performance comparable to conventional beam-forming algorithms. Ongoing effort aims to make further improvement through intermediate fusion.

A novel algorithm for speech activity detection based on fusing acoustic phonetic features and energy features is proposed and successfully evaluated on the CHIL seminar corpus.

References

1. M. Brandstein and D. Ward, Ed., *Microphone Arrays*, Berlin: Springer Verlag, 2000.
2. Q. Li, J. Zheng, Q. Zhou, and C.-H. Lee, "A robust, real-time end-point detector with energy normalization for ASR in adverse environments," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 233-236.
3. A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 237-240.
4. J. Padrell, D. Macho, and C. Nadeu, "Robust speech activity detection using LDA applied to FF parameters," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.
5. D. Macho et al., "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," to be presented at IEEE International Conference on Multimedia & Expo, Amsterdam, Netherlands, 2005.
6. B. Ramabhadran, J. Huang, and M. Picheny, "Towards automatic transcription of large spoken archives – English ASR for the MALACH project," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
7. D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively Trained Features for Speech Recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 1, pp. 961-964.
8. J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347-354.
9. M. Monkowski, "Automatic gain control in a speech recognition system," U.S. Patent 6,314,396, November 6, 2001.
10. D. Macho, "Speech activity detection: summary of CHIL evaluation run #1," January 2005, <http://chil.server.de/servlet/is/3870/>.