



# Detection, Diarization, and Transcription of Far-Field Lecture Speech

Jing Huang, Etienne Marcheret, Karthik Visweswariah, Vit Libal, Gerasimos Potamianos\*

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.  
 {jghg,etiennem,kv1,libalvit,gpotam}@us.ibm.com

## Abstract

Speech processing of lectures recorded inside smart rooms has recently attracted much interest. In particular, the topic has been central to the Rich Transcription (RT) Meeting Recognition Evaluation campaign series, sponsored by NIST, with emphasis placed on benchmarking speech activity detection (SAD), speaker diarization (SPKR), speech-to-text (STT), and speaker-attributed STT (SASTT) technologies. In this paper, we present the IBM systems developed to address these tasks in preparation for the RT 2007 evaluation, focusing on the far-field condition of lecture data collected as part of European project CHIL. For their development, the systems are benchmarked on a subset of the RT Spring 2006 (RT06s) evaluation test set, where they yield significant improvements for all SAD, SPKR, and STT tasks over RT06s results; for example, a 16% relative reduction in word error rate is reported in STT, attributed to a number of system advances discussed here. Initial results are also presented on SASTT, a task newly introduced in 2007 in place of the discontinued SAD.

**Index Terms:** speech processing, speech recognition, speaker diarization, speech activity detection, lectures, smart rooms.

## 1. Introduction

Lectures and meetings play a significant role in human collaborative activities in the workplace, with speech constituting the primary mode of interaction. Not surprisingly, speech processing in such scenarios has attracted much interest, being the focus of a number of research efforts and international projects, for example CHIL [1], AMI [2], and the U.S. National Institute of Standards and Technology (NIST) Smartspace effort [3]. In these projects, the interaction happens inside smart rooms equipped with multiple audio and visual sensors. Based on the resulting data, the goal is to extract higher-level information, in order to assist, for example, lecture indexing, browsing, summarization, and understanding.

Central to this goal is automatic speech recognition (ASR) or *speech-to-text* (STT) technology, 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. It is therefore not surprising that significant research effort is being devoted to developing SAD, SPKR, and STT algorithms for lectures and meetings inside smart rooms. Noticeably, these efforts have been rigorously evaluated in the past few years within the Rich Transcription (RT) Meeting Recognition Evaluation campaign series, sponsored by NIST [4].

In this paper, we present a summary of the IBM efforts to address SAD, SPKR, and STT using far-field audio inside smart rooms for the lecture scenario central to European-funded project CHIL, “Computers in the Human Interaction Loop”. In

\* We would like to acknowledge support of this work by the European Commission under integrated project CHIL, “Computers in the Human Interaction Loop”, contract number 506909.

this task, a subject presents a seminar of technical nature in English, with varying interactive audience participation. This represents an extremely challenging scenario, due to the presence of multiple speakers with often overlapping speech, a variety of interfering acoustic events, the strong accents of most speakers, a high level of spontaneity, hesitations and disfluencies, the technical seminar contents, and the relatively small amount of in-domain data. An additional major challenge is the use of far-field sensors. In particular, we use the *table-top microphones* only; these do not have exact positions, therefore their relative geometry is unknown.

We follow a natural sequence in our presentation of the three technologies, starting with SAD, followed by SPKR and STT (Sections 3, 4, and 5, respectively). For each, we briefly describe past efforts and discuss details of the latest developed system, emphasizing novel aspects. The systems share common components and are sufficiently dependent on each other to warrant presentation in a single paper. Compared to last year’s systems [5, 6], the following advances are noteworthy:

- Improved speech activity detection (SAD) based on speech/non-speech hidden Markov models built by bottom-up clustering of the Gaussians of an STT system.
- Effective speaker segmentation (SPKR), involving initial segmentation and subsequent refinement using GMMs.
- STT system combination via ROVER [7] of multiple systems built by randomized decision-tree growing [8].
- Two types of STT decoding, one using static decoding graphs with a small language model (LM) for the initial speaker-independent and MPE model decoding; followed by on-line dynamic decoding with a very large LM for the final MLLR-adapted MPE decoding.

We precede the three technology sections with Section 2, where we briefly describe data resources used for system training and tuning. We report experiments in Section 6, and finally, conclude the paper with a summary in Section 7.

## 2. Data Resources

The CHIL project [1] includes five partner sites with state-of-the-art smart rooms that contain similar sensory setups. However, the available amount of data recorded and annotated so far is insufficient for in-domain system training, especially for STT. To remedy this problem, additional publicly available corpora [9] are utilized that exhibit similarities to the CHIL scenario, as discussed next.

### 2.1. Training Data

The following data resources are used for system *training*:

- ICSI meeting data corpus, about 70 hours in duration.
- NIST meeting pilot corpus, about 15 hours.
- RT04 development and evaluation data, about 2.5 hours.
- RT05s development data, about 6 hours.

- AMI meetings, about 16 hours.
- CHIL 2003 and 2004 data, for a total of 4 hours.
- CHIL 2006 and 2007 development data, about 6 hours.
- Part of the CHIL RT06s evaluation test set, consisting of 11 five-minute segments, about 1 hour in total.

All datasets contain close-talking and multiple far-field microphone data. Since in this work we focus on the latter, for acoustic modeling we select all table-top microphones present in the corpora, with the exception of AMI data, where two microphones from each eight-element circular microphone array are chosen based on their location. This results to approximately 500 hours of far-field data. Notice that additional available resources, such as recently released AMI data and NIST meetings are not used for acoustic modeling; however, their transcripts are employed for language modeling (see Section 5.2).

## 2.2. Development and Evaluation Data

For system development (tuning), we utilize as development data the remaining part of the RT06s evaluation test set, not used in system training. This consists of 17 five-minute segments, for a total of 85 mins, recorded in all five CHIL smart room sites. In addition to the development data, our systems are run on the CHIL RT07 test set (Rich Transcription 2007 evaluation), consisting of 32 five-minute long segments.<sup>1</sup> When reporting far-field results in our experiments, we focus on two conditions: (i) *Single distant microphone* (SDM) condition, with only one table-top microphone used, as specified by NIST; and the (ii) *Multiple distant microphone* (MDM) condition, where typically all table-top microphones are used (ranging from three to five).

## 3. The Speech Activity Detection System

Speech activity detection (SAD) is a pre-requisite to both SPKR and STT. 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. For the RT06s evaluation, the IBM team developed two schemes for SAD: One was officially evaluated in the RT06s SAD task, and was based on a complex scheme of fusing acoustic likelihood and energy features for modeling three classes by full-covariance GMMs. During testing, the classes were collapsed into speech and silence, and appropriately smoothed to yield the final SAD result. Significant performance gains were observed when combining SAD results across multiple far-field channels by simple “voting” (decision fusion) [5].

The second scheme was employed as a first step in the IBM RT06s STT system, but was not independently evaluated [6]. It was basically a hidden Markov model (HMM)-based speech/non-speech decoder; speech and non-speech segments were modeled with five-state, left-to-right HMMs. The HMM output distributions were tied across all states and modeled with a mixture of diagonal-covariance Gaussian densities. The non-speech model included the silence phone and three noise phones. The speech model contained all speech phones. Both were obtained by applying a likelihood-based, bottom-up clustering procedure to the speaker-independent acoustic model developed for STT (see Section 5), but MAP-adapted to the CHIL part of the training data (see Section 2).

By varying the number of Gaussians in the two HMMs, we are able to obtain different operating points for SAD performance, i.e. ratio of missed speech vs. false alarm speech. Because missed speech cannot be recovered in the later stages

<sup>1</sup> Results on the latter set were made available by NIST after the submission of this paper. They are reported in two papers in the Proceedings of the CLEAR 2007 / RT07 evaluation workshop, held in Baltimore, Maryland, in May 2007.

of processing (SPKR, STT), we choose an operating point that only misses minimal amounts of speech, without of course introducing many false alarms. This simple scheme works extremely well, and as a result, very little gain is observed when combining multiple microphone channels, as compared to just using a single far-field microphone with the best SNR. Relevant results are reported in Section 6.

## 4. The Speaker Diarization System

For the RT06s evaluation, the IBM team developed a simple speaker clustering procedure to combine SAD segments into pseudo-speaker clusters that are subsequently used for speaker adaptation. The following procedure was used for this purpose: All homogeneous speech segments were modeled using a single Gaussian density, and were bottom-up clustered into a pre-specified number of clusters using  $K$ -means and a Mahalanobis distance measure. For CHIL data, the number of speaker clusters was set to the ad-hoc value of four over each lecture. This particular scheme proved sufficient for STT, but was never evaluated as a separate SPKR system in RT06s. As a matter of fact, its performance would have been dismal (see Section 6). A few attempts have been made since to improve performance:

(i) Instead of using a fixed number of speaker clusters, we first over-segment the data into, let’s say eight clusters, and subsequently merge them according to the Mahalanobis distance function. We terminate the merging process, when a threshold value is reached, as determined by development data.

(ii) Instead of using an acoustic front end identical to the SAD and STT systems (see Section 5), we switch to 19-dimensional MFCC features with no energy. Such features are widely used in speaker identification systems.

(iii) Following SAD, we decode the speech segments with the speaker-independent STT model (see Section 5). We use alignment information from the resulting decoding to isolate the speech frames, and then compare clusters only on these speech frames. This removes short silence, background noise and vocal noise frames that do not help with speaker discrimination. As a result, we obtain better Gaussian models for each speaker.

(iv) We change the distance function from the Mahalanobis to a likelihood gain. Each cluster is modeled by one Gaussian with full covariance. At each step in the bottom-up clustering process, we combine the two nodes that result in the smallest likelihood loss. We terminate the process, when no two nodes can be joined with a loss smaller than a pre-specified threshold.

(v) We add an iterative refinement step after merging stops. We build GMMs for each speaker cluster based on the previous decision. We then relabel each frame by averaging scores of the speaker models for 0.75 seconds on each side of the current frame, and we assign the frame to the speaker with the largest smoothed score. The advantage of this refinement is to split a segment, if there are speaker change points within it. This turns out to be better than our last year’s change-point detection scheme with two Gaussians built locally using a small window.

## 5. The Speech-to-Text System

We now proceed to describe the IBM STT system developed for the RT07 evaluation campaign. The system has significant variations compared to the RT06s one in all acoustic modeling, language modeling, and the recognition process employed.

### 5.1. Acoustic Modeling

For acoustic modeling, first a speaker-independent (SI) model is trained, based on 40-dimensional acoustic features generated by an LDA projection of nine consecutive frames of 13-dimensional perceptual linear prediction (PLP) features, ex-

condition	RT06s system	RT07 system
SDM	15.0 %	5.1 %
MDM	10.6 %	5.0 %

Table 1: SAD diarization error (%) of the RT06s and RT07 systems on the development set for the two far-field microphone conditions.

tracted at 100 Hz. The features are mean normalized on a per-speaker basis. The SI model uses continuous density, left-to-right HMMs with Gaussian mixture emission distributions and uniform transition probabilities. In addition, the model uses a global semi-tied covariance linear transformation [10], which is also updated at every EM training stage. The sizes of the mixtures are increased in steps interspersed with EM updates, until the final model complexity is reached. Each HMM has three states, except for a 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 a small only part of the training data is from the CHIL domain (see Section 2), MAP-adaptation of the SI model was deemed necessary to improve performance on CHIL data.

The SI features are further normalized with a voicing model (VTLN) with no variance normalization. The most likely frequency warping is estimated among 21 candidate warping factors ranging from 0.8 to 1.2. Warping likelihoods are estimated using a voicing model built on 13-dimensional PLP features. A VTLN model is subsequently trained on features in the VTLN warped space. In that feature space, a new LDA transform is estimated and a new VTLN model is obtained by decision tree clustering of quinphone statistics. The 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 fM-LLR transforms to the VTLN normalized features. fM-LLR transforms are computed like the VTLN warping factors on a per-speaker basis for all speakers in the training set. The SAT HMMs again have 10k tied-states and 320k Gaussians.

Following SAT, we estimate feature-space minimum phone error (fMPE) transforms [11], followed by minimum phone error (MPE) training. The fMPE projection uses 1024 Gaussians obtained from clustering the Gaussian components in the SAT model. Posterior probabilities are then computed for these Gaussians for each frame, and time-spliced vectors of these posterior probabilities become the foundation for the features that are subjected to the fMPE transformation. The fMPE transformation maps the high-dimensional posterior-based observation space to a 40-dimensional fMPE feature space. The MPE model is then trained in this feature space with MAP-MPE on the available amount of CHIL-only data [12].

Following the above training procedure, we build two systems: (A) with the VTLN step present, and (B) with VTLN removed. Both types of systems were also used in the IBM RT06s submission [6]. However, a third system based on variance normalization is no longer employed this year. Instead, two additional SAT systems are built using the randomized decision tree approach [8], resulting in systems (C) and (D). Randomized decision trees are grown by randomly selecting the split at each node, among the top  $N$ -best split candidates ( $N=5$  here) – instead of always considering the best one. The two resulting systems were built again to have 10k states and 320k Gaussians. The purpose of building such multiple systems is to combine them at final decoding using the ROVER technique [7]. Experiments in support of this approach are given in Section 6.

## 5.2. Language Modeling

To improve language modeling over our RT06s system, we complement the four training sources used last year with web

SPKR System	SPKR DER (%)	SI-STT WER (%)
RT06s, fixed $N=4$	70.1	61.2
RT07: thresholding	9.2	54.2
RT07: + alignment	8.2	n/a
RT07: likel. merging	8.0	n/a
RT07: + refinement	7.4	54.2
RT07: + alignment	7.4	n/a

Table 2: SPKR diarization error (DER, %) and its corresponding speaker-independent MDM STT word error rate (WER, %), for various speaker segmentation systems.

data. We thus construct five separate four-gram models: The first is based on 0.15M words of CHIL lecture transcripts; the second uses 2.7M words of non-CHIL meeting transcripts; the third is built on 37M words of scientific conference proceedings; the fourth on 3M words of Fisher data [9]; and the fifth uses 525M words of web data available from the EARS program [9]. For decoding, two language models (LMs) are used: For static decoding, we interpolate the five models with weights of 0.31, 0.24, 0.20, 0.06, and 0.19 respectively (optimized on the union of CHIL 2007 development data and 11 segments of the CHIL RT06s evaluation test set), and use entropy-based pruning [13] to reduce the resulting model to about 5M n-grams. For on-the-fly dynamic graph expansion decoding, we only prune the LM from the web data and construct a large interpolated LM with 152M n-grams. A 37k-word vocabulary is obtained by keeping all words occurring in meeting and Fisher data and the 20k most frequent words in the other text corpora.

## 5.3. Recognition Process

After speech segmentation and speaker clustering, for each table-top microphone, a final system output is obtained in the following three decoding passes:

- (i) The SI pass uses MAP-adapted SI models to decode.
- (ii) Using the output from (i), warp factors are estimated for each cluster using the voicing model, and fM-LLR transforms are estimated for each cluster using the SAT model. The VTLN features after applying the fM-LLR 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. The resulting transcripts are denoted by ctm-n, where n stands for model (A), (B), (C), or (D).
- (iii) The output transcripts from step (ii) are used to estimate M-LLR transforms on the MPE model. The adapted MPE model together with the large 152M n-gram LM are used for final decoding with a dynamic graph expansion decoder. The final transcripts at this step will be referred to as CTM-n, where n stands for model (A), (B), (C), or (D). Note that instead of using its own decoding outputs from step (ii), we employ cross-system adaptation; i.e., ctm-(A) is input to system (C), ctm-(C) is input to (B), ctm-(B) to (D), and ctm-(D) to system (A).

The final system output for MDM is obtained using ROVER in the following way: First, for each system, we apply ROVER to the outputs from all available table-top microphones; then we combine the four systems to obtain the final transcripts.

## 6. Experimental Results and Discussion

We now proceed to present experimental results for all SAD, SPKR, and STT systems on our selected development set (a subset of the RT06s evaluation test set, as discussed in Section 2.2). In particular, with respect to SAD and SPKR scoring, all reported results are based on forced alignment references, in accordance to the NIST planned scoring of RT07 systems. We also briefly report SASTT system results – this is basically an

	ref. RT07	auto. RT07	auto. RT06s
SI	54.1	54.2	61.2
MPE	45.8	46.0	50.6
MLLR-MPE	43.5	43.4	n/a

Table 3: Comparison of STT WERs (%) at various decoding stages using reference segmentation (RT07 STT) and automatic segmentation with the SAD/SPKR/STT RT07 and RT06s systems. In all cases, STT system (B) in the MDM condition is considered.

SPKR system	SPKR	STT	SASTT
thresholding	9.2 %	41.9 %	44.1 %
likel. merging + refinement	7.4 %	41.9 %	43.5 %

Table 4: Impact of improving SPKR DER on the WER of the RT07 STT and SASTT systems for the MDM condition.

STT system with words assigned to the various SPKR clusters. Wrong assignments are penalized based on a recently developed scoring tool by NIST.

Table 1 compares the SAD results of our new system to last year’s RT06s entry (see Section 3). In particular, the SAD error (so called “diarization error” – see [4]) improved dramatically by 66% relative, from 15.0% to 5.1% in the SDM condition. Interestingly, channel combination no longer provides significant gains, as it was the case with last year’s system.

On the SPKR task, we improved significantly over last year’s simple speaker clustering scheme, as it is evident from Table 2. Notice of course that last year’s SPKR system was not submitted for evaluation, rather just used to drive our well-performing STT system. Table 2 depicts the improvement from each technique discussed in Section 4, and the impact on STT based on the speaker-independent (SI) model: Using thresholding and 19-dim MFCC features dramatically improved DER from 70.1% to 9.2%. This is because most CHIL lectures have one dominant speaker – the lecturer. Forcing each segment to four clusters creates many speaker errors. Using alignment definitely helps lower DER by another 1% absolute, because non-speech segments are removed before clustering. As a result, both false alarm speech as well as speaker errors are reduced. The likelihood distance function turns out to be superior to the Mahalanobis one, further lowering DER. Finally, the refinement step also proves to be effective.

The impact of the improved SAD/SPKR on STT decoding is significant. This is depicted in Table 3 for system (B). There, it can be seen that we improved the SI system from the 61.2% WER of last year to 54.2% this year, which turns out to be only 0.1% worse than the WER when decoding using the reference segmentation. In fact, the final decoding with cross-system MLLR-adapted MPE model and the large LM results in a WER of 43.4%, which is a little better than the 43.5% obtained when using the reference segmentation.

We notice from Table 2 that, after a certain point, SPKR DER improvement no longer reduces STT WER. However, on the task of speaker-attributed speech-to-text (SASTT), such improvements help, as shown in Table 4. For example, a 1.8% reduction in SPKR DER results in 0.6% SASTT WER reduction. In all cases, the final RT07 MDM STT system is considered.

The STT improvements reported already in Tables 2, 3, and 4 are due not only to the SAD and SPKR system enhancements, but also due to better acoustic and language modeling. In particular, with respect to the latter, careful vocabulary selection and the use of the large 525M-word web dataset helped significantly. For example, using acoustic model (B), when adding web data to the small LM for SI static decoding achieves a 1% absolute gain (from 55.2% to 54.2% WER); using the large LM with dynamic decoding (at the MLLR-MPE stage) achieves over 3% absolute gain, from 46.7% to 43.4%.

Finally, in Table 5, we depict STT results for the various

System	(A)	(B)	(C)	(D)
MAP-SI			54.2	
MPE	46.3	46.0	47.3	46.3
cross-MLLR+MPE	42.9	43.4	43.3	43.0
final ROVER			41.9	

Table 5: WERs, %, of MDM STT systems at the various decoding stages described in Section 5.3, using automatic speaker segmentation with DER of 9.2%.

systems in the MDM condition. After applying ROVER across all systems, we obtain the final WER result of 41.9%, which represents a dramatic improvement over the 50.0% result of our RT06s system, namely about an 8% absolute WER reduction.

## 7. Conclusions

We have made significant progress in the automatic detection, diarization, and transcription of CHIL lectures using far-field audio. Major highlights of the developed systems are improvements in speech activity detection, speaker segmentation, acoustic model training, system combination, and development of a large language model using web data. The effort has led to a 16% relative reduction in word error rate, and dramatic reductions in speech activity and speaker diarization errors, all benchmarked on a subset of the NIST RT06s evaluation test set.

## 8. 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] 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.
- [6] 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.
- [7] 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.
- [8] O. Siohan, B. Ramabhadran, and B. Kingsbury, “Constructing ensembles of ASR systems using randomized decision trees,” in *Proc. Int. Conf. Acoustics Speech Signal Process.*, Philadelphia, PA, vol. 1, pp. 197–200, 2005.
- [9] “The LDC Corpus Catalog,” Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, [Online]. Available: <http://www ldc.upenn.edu>
- [10] G. Saon, G. Zweig, and M. Padmanabhan, “Linear feature space projections for speaker adaptation,” in *Proc. Int. Conf. Acoustics Speech Signal Process.*, Salt Lake City, UT, pp. 325–328, 2001.
- [11] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltan, and G. Zweig, “fMPE: Discriminatively trained features for speech recognition,” in *Proc. Int. Conf. Acoustics Speech Signal Process.*, Philadelphia, PA, vol. 1, pp. 961–964, 2005.
- [12] D. Povey and P.C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in *Proc. Int. Conf. Acoustics Speech Signal Process.*, Orlando, FL, pp. 105–108, 2002.
- [13] A. Stolcke, “Entropy-based pruning of backoff language models,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 270–274, 1998.