The IBM Rich Transcription 2007 Speech-to-Text Systems for Lecture Meetings

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. The paper describes the IBM systems submitted to the NIST Rich Transcription 2007 (RT07) evaluation campaign for the speech-to-text (STT) and speaker-attributed speech-to-text (SASTT) tasks on the lecture meeting domain. Three testing conditions are considered, namely the multiple distant microphone (MDM), single distant microphone (SDM), and individual headset microphone (IHM) ones – the latter for the STT task only. The IBM system building process is similar to that employed last year for the STT Rich Transcription Spring 2006 evaluation (RT06s). However, a few technical advances have been introduced for RT07: (a) better speaker segmentation; (b) system combination via the ROVER approach applied over an ensemble of systems, some of which are built by randomized decision tree state-tying; and (c) development of a very large language model consisting of 152M n-grams, incorporating, among other sources, 525M words of web data, and used in conjunction with a dynamic decoder. These advances reduce STT word error rate (WER) in the MDM condition by 16% relative (8% absolute) over the IBM RT06s system, as measured on 17 lecture meeting segments of the RT06s evaluation test set, selected in this work as development data. In the RT07 evaluation campaign, both MDM and SDM systems perform competitively for the STT and SASTT tasks. For example, at the MDM condition, a 44.3% STT WER is achieved on the RT07 evaluation test set, excluding scoring of overlapped speech. When the STT transcripts are combined with speaker labels from speaker diarization, SASTT WER becomes 52.0%. For the STT IHM condition, the newly developed large language model is employed, but in conjunction with the RT06s IHM acoustic models. The latter are reused, due to lack of time to train new models to utilize additional close-talking microphone data available in RT07. Therefore, the resulting system achieves modest WERs of 31.7% and 33.4%, when using manual or automatic segmentation, respectively.

1 Introduction

Meetings and lectures play a central 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. The ultimate goal is to extract higher-level information to facilitate meeting indexing, browsing, summarization, and understanding.

Central to this goal is automatic speech recognition (ASR) or *speech-to-text* (STT) technology, and its complementary technologies of *speech activity detection* (SAD) and *speaker diarization* (SPKR). The three partially address the "what", "when", and "who" of human interaction, and are important drivers of additional technologies, for example speaker localization and recognition, summarization, and question answering. Not surprisingly, significant research effort is being devoted to developing and improving these technologies in the meeting scenario. Resulting systems 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 in developing STT technology and its evaluation in the RT07 campaign. The emphasis of this work is on the "lecture meeting" scenario, central to European-funded project CHIL, "Computers in the Human Interaction Loop" [1]. In this scenario, a subject presents a seminar (lecture) of technical nature in English, with varying interactive participation by a relatively small audience. This represents significant challenges to state-of-the-art STT technology, due to the presence of multiple speakers with often overlapping speech, a variety of interfering acoustic events, strong accents by most speakers, a high level of spontaneity, hesitations and disfluencies, the technical seminar contents, and the relative small amount of in-domain data. Furthermore, the main emphasis is placed on developing STT systems based on far-field microphones, the goal being to achieve interaction with unobtrusive sensors that "fade into the background". This poses additional challenges to the problem, for example low signal-to-noise ratios and reverberation. In particular, the primary evaluation condition in RT07 employs table-top microphones only. These do not have exact positions, therefore their relative geometry is unknown.

In addition to STT, NIST introduced a new variant of it in the RT07 evaluation. This is the so-called speaker-attributed speech-to-text (SASTT) task, which combines both speaker diarization ("Who Spoke When") and STT into a single jointly evaluated task. The purpose of an SASTT system is to correctly transcribe the words spoken, but in addition to also identify the generically labeled speaker of the words. The IBM SASTT system performance will be briefly discussed here, however a detailed description of its speaker diarization component can be found in an accompanying paper [5].

For the IBM team, this effort constitutes the second year of participation in the RT evaluation campaign. A number of technical improvements have been introduced over the IBM STT systems that competed in RT06s [6], most importantly:

- Improved speech activity detection (SAD), which is the very first step for STT, affecting performance of both speaker segmentation and STT.

- Improved speaker segmentation for diarization (SPKR), based on thresholding schemes instead of a fixed number of speaker clusters. Further refinement is achieved through the use of alignment information.
- System combination (via the ROVER technique [7]) of multiple ASR systems, some of which are now built using a randomized decision-tree growing procedure [8].
- Use of two language models (LMs) in the decoding process: Fast decoding employs static graphs with a small LM for the initial decoding phases, whereas on-line dynamic decoding is used for the final recognition step in conjunction with a very large LM. Both LMs are built by introducing an additional source of data mined from the world wide web.

The remainder of the paper is structured as follows: Data resources used for training, development, and evaluation are overviewed in Section 2. Section 3 is devoted to system descriptions, including speaker segmentation, acoustic and language modeling, and the decoding process. Particular emphasis is placed on improvements over the IBM systems evaluated in RT06s. Experimental results on development data, as well as on the RT07 evaluation test set, are presented in Section 4. Finally, Section 5 concludes the paper.

2 Data Resources

For the lecture meeting domain of the RT evaluation campaign, development and evaluation data are provided by the CHIL consortium [1], that includes five partner sites with state-of-the-art smart rooms of similar sensory setups. However, for STT system training purposes, the available amount of CHIL data is insufficient. 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. For far-field acoustic model training, we select all table-top microphones present in the corpora, with the exception of AMI data, where two microphones from the eight-element circular microphone arrays 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. The TED corpus [10] is also not used, due to its reliance on lapel microphone recordings. Finally, due to time constraints, no new close-talking acoustic model has been developed; instead, the IBM RT06s model is used, trained on a subset of the above-listed resources [6].

2.2 Development, Evaluation Data, and Conditions

For system development (tuning), we utilize as development data (dev) 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. The evaluation data set (eval) is of course the CHIL RT07 test set, which consists of 32 five-minute long segments.

When reporting results, we focus on the following three conditions, two far-field and one close-talking:

(i) *Single distant microphone* (SDM) condition, with only one table-top microphone used, as specified by NIST.
(ii) *Multiple distant microphone* (MDM) condition, where typically all table-top microphones (ranging from three to five) are used / combined to yield a single transcript. This constitutes the primary evaluation condition.
(iii) *Individual headset microphone* (IHM) condition, where all headsets worn by lecture participants are decoded, with the purpose of recognizing the wearer's speech (i.e. decoding cross-talk is penalized). To facilitate cross-talk removal, both manual and automatic segmentations are provided, the latter kindly contributed by ICSI/SRI [11].

3 The IBM Systems

We now proceed to describe the IBM systems developed for the RT07 evaluation on the lecture meeting domain. Since training procedures for both IHM and far-field systems (SDM and MDM conditions) are similar and all share the same language model, the main emphasis of the presentation is placed on far-field STT. Similarly to the RT06s evaluation [6], this is developed around an architecture that combines decoded outputs of multiple systems over multiple table-top microphones (in the MDM case). The section describes front-end processing, speaker segmentation, acoustic and language modeling, and the decoding process, with emphasis on modifications over last year's systems. No description is given for the SASTT system, since this is an obvious union of the far-field STT and SPKR subsystems described in this section.

3.1 Acoustic Front-End

The features extracted from the acoustic signal for STT are 40-dimensional vectors obtained from a linear discriminant analysis (LDA) projection. The source space for the projection is 117-dimensional and is obtained by concatenating nine temporally consecutive 13-dimensional acoustic observation vectors based on perceptual linear prediction (PLP). The PLP features are computed at a rate of 100 frames per second from a Hamming windowed speech segment of 25 ms in duration. The vectors contain 13 cepstral parameters obtained from the LPC analysis of the cubic root of the inverse DCT of the log outputs of a 24-band, triangular filter bank. The filters are positioned at equidistant points on the Mel-frequency scale between 0 and 8 kHz. The cepstral parameters are mean-normalized on a per-speaker basis. No noise filtering is applied to the audio.

3.2 Speaker Segmentation

The first step of speaker segmentation is speech activity detection (SAD). After SAD, long segments of non-speech (silence or noise) are removed, and the audio is split into shorter segments for fast decoding and speaker segmentation. Last year, IBM developed two schemes for SAD, a complex one used in the RT06s SAD evaluation [12], and a simpler scheme employed for STT [6]. The latter is used again this year, but improved. It basically constitutes an HMM speech/non-speech decoder: Speech and non-speech segments are modeled with five-state, left-to-right HMMs with no skip states. The output HMM distributions are tied across all HMM states, and are modeled with a mixture of diagonal-covariance Gaussians. The non-speech model includes the silence phone and three noise phones, whereas the speech model retains all remaining (speech) phones. Both are obtained by a likelihood-based, bottom-up clustering procedure, applied to the speaker-independent STT acoustic model, MAP-adapted to the available CHIL training data. SAD system details can be found in [5].

The SAD system output is passed as input to the speaker segmenter (SPKR). As compared to our approach last year (part of the IBM RT06s STT system, but not evaluated separately), this year we remove the change-point detection procedure, but keep the simple speaker clustering scheme with few modifications: All homogeneous speech segments are modeled using a single Gaussian and are bottom-up clustered by K-means with a Mahalanobis distance measure. Instead of having a pre-set fixed number of clusters, we first over-segment the data into, let's say eight clusters, merge clusters according to the Mahalanobis distance, and stop merging when a threshold value is reached – optimized on development data. We also switch to extracting 19-dimensional MFCC acoustic features (with no energy term), instead of PLP ones, since the former are widely used in the speaker recognition literature. Details of the system are presented in [5].

3.3 Acoustic Modeling

The speaker-independent (SI) acoustic model is trained on 40-dimensional features, extracted as discussed in Section 3.1. It employs 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 [13, 14] that is updated at each EM training iteration. 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, among which 41 are speech phones, one corresponds to silence, and three model noise, namely background noise, vocal noise, and breathing. The final HMMs have 6k context-dependent tied states and 200k Gaussian mixture components. Since a very small portion of the available training data come from CHIL sites, MAP adaptation of the SI model to CHIL data was deemed necessary to improve its performance on lecture meetings.

The SI features are further normalized with a voicing model (VTLN) with no variance normalization. The frequency warping is piece-wise linear with a breakpoint at 6500 Hz. The most likely frequency warping is estimated among 21 candidate warping factors ranging from 0.8 to 1.2 with a step of 0.02. Warping likelihoods are obtained by means of a voicing model, built on 13-dimensional PLP features. A VTLN model is then trained on features in the VTLN warped space. VTLN warping factors are estimated on a per-speaker basis for all training set data using the voicing model. 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 resulting HMMs have 10k tied states and 320k Gaussians.

Following VTLN, speaker adaptive training (SAT) is performed. The SAT model is trained on features in a linearly transformed feature space, resulting from applying feature-space maximum likelihood linear regression (fMLLR) transforms to the VTLN normalized features [14]. The fMLLR transforms are computed on a per-speaker basis, for all training set speakers. The resulting SAT HMMs have 10k tied states and 320k Gaussians.

Following SAT model training, we estimate feature-space minimum phone error (fMPE) transforms [15], and subsequently perform minimum phone error (MPE) training. The fMPE projection uses 1024 Gaussians obtained from clustering the Gaussian components of the SAT model. Posterior probabilities are then computed for these Gaussians for each frame, and time-spliced vectors of the resulting probabilities become the basis for the features subjected to the fMPE transformation. Such transformation maps the high-dimensional posterior-based observation space to a 40-dimensional fMPE feature space. The MPE model is subsequently trained in this space, with MAP-MPE applied to the available amount of CHIL training data [16]. In this implementation, we have changed the MPE objective function to operate on the frame level [17], instead of the utterance level, as this resulted in slight gains.

The above training procedure provides two systems: System (A) with VTLN present, and system (B) without the VTLN step. Based on experience from the RT06s results [6], we do not use variance normalization in the VTLN and SAT models. To yield better gains by ROVER-based system combination [7], we built two additional SAT systems using the randomized decision tree approach [8] (discussed briefly next). For both, the process starts from SAT system (B), followed by fMPE/MPE training. The two additional resulting systems will be denoted by (C) and (D), and – similarly to systems (A) and (B) – they too consist of 10k tied states and 320k Gaussians.

In more detail, randomized decision trees are grown by randomly selecting the split at each node, among the top N-best split candidates (N= 5 in our case). This is in contrast to standard decision trees that are grown by only considering the best split. Systems built on different sets of randomized decision trees will model different clusters of context-dependent units. Multiple systems can then be systematically obtained by simply changing the random number generator seed. It has been experimentally shown that such systems are good candidates to be used in the ROVER voting procedure [8]. Results in Section 4 support this observation.

3.4 Language Modeling

To improve language modeling over our RT06s system [6], we complement the four training sources used last year with web data. We thus construct five separate four-gram language models (LMs), all smoothed with the modified Kneser-Ney algorithm [18]: The first LM is based on 0.15M words of CHIL meeting transcript data; the second uses 2.7M words of non-CHIL meeting corpora resources; the third one utilizes 37M scientific conference proceedings (primarily from data processed by CHIL partner LIMSI); the fourth LM uses 3M words of Fisher data [9]; and finally the fifth employs 525M web data available from the EARS program [9]. To construct the LM used for the static decoding graph, we interpolate the five models with weights of 0.31, 0.24, 0.20, 0.06, and 0.19 respectively (optimized on CHIL 2007 development data, as well as 11 segments of the RT06s evaluation test set). We subsequently perform entropy-based pruning [19] to reduce the resulting model to about 5M n-grams. We then employ this LM at the SI and MPE decoding steps (see also Section 3.5). In addition, we consider a much larger LM in conjunction with on-the-fly dynamic graph expansion decoding at the final recognition step (see next subsection). This LM is obtained by pruning only the web data LM, and it consists of 152M n-grams.

A 37k-word vocabulary is obtained by keeping all words occurring in the meeting transcripts and Fisher data and the 20k most frequent words in the other text corpora. Pronunciations use the phone set discussed in Section 3.3, and are based on the Pronlex lexicon, manually augmented whenever necessary.

3.5 Recognition Process

After pseudo-speakers are determined following speaker segmentation, 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 transcript from (i), warp factors are estimated for each cluster 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 MPE model and the fMPE features. The MPE model is also trained with MAP on the CHIL data. The one-best transcript at this step is referred to as ctm-n, where n stands for model (A), (B), (C), or (D).

(iii) The output transcripts from step (ii) are used to estimate maximum likelihood linear regression (MLLR) transforms on the MPE model. The adapted MPE model together with a large 152M n-gram language model are used for final decoding with a dynamic graph expansion decoder. The final one-best transcript at this step will be referred to as CTM-n, where n stands for model (A), (B), (C), or (D).

For each system at step (iii), instead of using the corresponding decoding output from step (ii), we employ cross-system adaptation as follows: ctm-(A) is used as input to system (C), ctm-(C) to system (B), ctm-(B) to (D), and finally ctm-(D) is used as input to system (A). The above process is shown to be beneficial in Section 4.

For the SDM evaluation condition, all four system outputs are combined using ROVER. For the MDM condition, two rounds of ROVER are applied: First, for each system we combine the outputs from all available table-top microphones; subsequently, we combine the four results across systems to obtain the final transcript.

Concerning runtime performance, the MDM system runs at approximately 90 times slower than real time (× RT). This can be broken down to 3.6 × RT for the SI decoding stage, 4.2 × RT for each of the four MPE system decodings, and 17.5 × RT for each of the four final MLLR adapted MPE decodings employing the large LM.

3.6 The Close-Talking STT System

For the close-talking STT system (IHM condition), we used the acoustic model trained in last year's evaluation (RT06s), in conjunction with the LMs developed for RT07, described in Section 3.4. In particular, our RT06s acoustic model used a subset of the resources listed in Section 2.1 with a total duration of 124 hrs. In contrast to the far-field, only one acoustic model was developed, with both VTLN and variance normalization present, consisting of 5k context dependent states and 240k Gaussians. Details can be found in [6]. For decoding, all three steps described in Section 3.5 are used, with the obvious modification that MLLR adaptation in step (iii) is carried out within the single available system. Furthermore, for cross-talk removal, the ICSI/SRI system output is employed [11].

4 Experimental Results and Discussion

We now proceed to present experimental results on our development (dev) set (a subset of the RT06s evaluation test set, as discussed in Section 2.2), as well as the lecture meeting RT07 evaluation test data (eval), for both STT and SASTT. For the latter task, in addition to the three traditional types of word errors (deletions, insertions, and substitutions), speaker-attributed word errors include

Table 1. STT WERs, %, at various decoding stages, for the RT06s and RT07 systems. For the latter, results with the reference segmentation are also depicted. All WERs are reported on the dev set for the MDM condition, with system (B) considered past the SI decoding level. For the automatic segmentation case, SPKR diarization error is 70.1% for the RT06s system, but only 9.2% for the RT07 one.

STT system	RT07		RT06s
segmentation	reference	automatic	
SI	54.1	54.2	61.2
MPE - sys. (B)	45.8	46.0	50.6
MLLR-MPE - sys. (B)	43.5	43.4	–

speaker label error, i.e. the mapped STT output tokens with matching reference words but non-matching speaker labels. Therefore, SASTT performance really measures both STT word error rate (WER) and diarization error (DER) of the SPKR task. Notice that in this paper, all WER results are reported scored with speech overlap factor set to one [4] – with the exception of Table 6.

We first demonstrate the significant improvement in far-field STT from RT06s to RT07. This is depicted in Table 1 for the MDM condition on our dev set, for one of the four acoustic models (model (B)). For example, SI WER improves from 61.2% (RT06s) to 54.2% (RT07). The improvement is due to a number of factors as discussed in the Introduction, including better acoustic and language modeling, as well as better speaker segmentation. In particular, it is noteworthy that the automatic speaker segmentation scheme used in the RT06s STT system achieves a dismal DER of 70.1%, whereas the RT07 segmentation scheme employed to obtain the tabulated results exhibits a 9.2% DER. The improvement is due to various factors, including the fact that the final number of speaker clusters is no longer fixed to four. Relevant SPKR system experiments can be found in [5]. Notice also, that for the RT07 system, the use of reference (manual) or automatic segmentation results in very similar WERs.

As already mentioned, much of the STT improvement is due to language modeling work. This fact is demonstrated in Table 2, where decoding performance using the RT06s and the two RT07 LMs is depicted, in conjunction with RT07 acoustic model (B). Adding web data to the small LM for SI static-graph decoding achieves a 1% absolute gain in WER (55.2% to 54.2%); employing the very large LM with the dynamic decoder achieves over 3% absolute WER gain, from 46.7% to 43.5%. Clearly, the large LM helps. It is also worth mentioning

Table 2. Dev-set WERs, %, for the MDM STT system developed for RT07, when employing various language models developed for RT06s and RT07. Depicted results are obtained using the reference segmentation.

Language Model	RT06s LM	RT07 – small LM	RT07 – large LM
SI	55.2	54.2	–
MLLR-MPE - sys. (B)	–	46.7	43.5

Table 3. Dev-set WERs, %, of the various developed STT systems for the MDM condition using automatic segmentation.

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			

that decoding with the static graph and the MLLR-adapted MPE model degrades WER from 45.8% (MPE model (B)) to 46.7%, a behavior consistent with the RT06s acoustic model. It therefore seems that our MLLR-adapted MPE model is worth using only with the very large LM.

In Table 3, we depict STT results on the dev set for the various developed systems at the MDM condition. After applying ROVER across all systems, we obtain a WER of 41.9%, which represents a large improvement over the 50.0% WER of our RT06s system – a 16% relative (8% absolute) WER reduction. Clearly, ROVER-based system combination helps, improving performance over the best system by 1% absolute. It is also interesting to remark that cross-system adaptation helps. For example, cross adaptation of system (A) reduces WER from 46.3% at the MPE level to 42.9%. That number would have been 44.1% under a self-adaptation regime.

The observed improvements in the dev set carry over to the eval set. Table 4 presents STT results on the RT07 evaluation test set (eval) for the MDM and SDM conditions. By selecting the highest-SNR microphones to drive the SAD and SPKR subsystems, and by applying ROVER across all available table-top microphones, the final WER at the MDM condition is 3.6% absolute better than the final WER at the SDM condition. Between SI decoding and the final result, the WER improves by a relative 20% (18%) for the MDM (SDM) condition. These gains are less than the 23% relative observed on the dev set. The reason may be that the DER of the SPKR sybsystem on eval data is much

Table 4. Far-field STT WERs, %, on the eval set at various decoding stages for the MDM and SDM conditions. The final ROVER results are obtained by combining all four systems (A)-(D), and were submitted at the RT07 evaluation with a one week delay. The primary IBM RT07 STT submissions on-time have combined three only systems (due to lack of sufficient time to train the fourth), resulting in WERs of 44.8% and 48.6% for the MDM and SDM conditions, respectively.

condition	MDM				SDM			
system	(A)	(B)	(C)	(D)	(A)	(B)	(C)	(D)
MAP-SI	55.5				58.6			
MPE	48.9	48.6	48.6	48.9	53.7	53.1	53.2	53.4
cross-MLLR+MPE	46.1	46.3	46.0	46.0	50.7	50.9	51.1	51.0
final ROVER	44.3				47.9			

Table 5. Eval-set WERs, %, of the IHM system depicted at various decoding stages using both automatic and reference segmentation.

segmentation	automatic	reference
MAP-SI	44.1	43.2
MPE	34.6	33.4
cross-MLLR+MPE	33.4	31.7

Table 6. SASTT results on dev and eval data. SPKR and STT system performance is also depicted. STT and SASTT performance is also shown when scoring overlapped speech ("o3" condition).

		SPKR DER (%)	STT WER (%)		SASTT WER (%)	
data	condition	o1	o1	o3	o1	o3
dev	MDM	9.2	41.9	–	44.1	–
eval	MDM	27.6	44.3	50.0	52.0	58.4
eval	SDM	27.4	47.9	–	55.4	60.8

higher (27.6%) than that on dev data (9.2% – see also [5] and Table 6). Further improvements are clearly needed in the SPKR system.

For the IHM condition, we used both the automatic segmentation provided by ICSI/SRI [11], as well as the reference segmentation. Table 5 depicts the eval set results. The final WER using the reference segmentation is about 1.7% better than the WER based on automatic segmentation, with 1.4% more substitution errors but 3.2% less deletion errors. It seems that the automatic segmentation misses some speaker segments.

Finally, Table 6 presents our best WERs for SASTT, as well as DERs for SPKR on both dev and eval sets, with overlapping factor set to one (or three) during scoring [4]. Notice that the WER degradation from the STT to the SASTT task is significantly higher in the eval set than in dev. This is due to the poor performance of the SPKR system on the eval set, as already mentioned.

5 Conclusions

We have made significant progress in the automatic transcription of lecture meeting data. Main system advances compared to the RT06s evaluation are improvements in speech activity detection, speaker segmentation, acoustic modeling, system combination, and development of a very large language model that incorporates web data. The effort has led to a 16% relative reduction in word error rate on development data and has resulted in competitive performance in the RT07 evaluation.

6 Acknowledgements

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.

References

1. Computers in the Human Interaction Loop. [Online]. `http://chil.server.de`
2. Augmented Multi-party Interaction. [Online]. `http://www.amiproject.org`
3. The NIST SmartSpace Laboratory. [Online]. `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. J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos, "The IBM RT07 evaluation system for speaker diarization in CHIL seminars," (Same Volume), 2007.
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. Automatic Speech Recognition Underst. Works.*, 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/Catalog`
10. L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann, "The translanguage English database (TED)," in *Proc. Int. Conf. Spoken Language Process.*, Yokohama, Japan, 1994.
11. K. Boakye and A. Stolcke, "Improved speech activity detection using cross-channel features for recognition of multiparty meetings," in *Proc. Int. Conf. Spoken Language Process.*, Pittsburgh, PA, pp. 1962–1965, 2006.
12. 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.
13. M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
14. 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.
15. D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, 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.
16. 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.
17. J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Eurospeech*, Lisbon, Portugal, pp. 2125–2128, 2005.
18. S.F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, pp. 359–393, 1999.
19. A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcr. Underst. Works.*, Lansdowne, VA, pp. 270–274, 1998.