

The IBM Rich Transcription Spring 2006 Speech-to-Text System for Lecture Meetings

Jing Huang, Martin Westphal, Stanley Chen, Olivier Siohan, Daniel Povey,
Vit Libal, Alvaro Soneiro, Henrik Schulz, Thomas Ross,
and Gerasimos Potamianos

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.

Abstract. We describe the IBM systems submitted to the NIST RT06s Speech-to-Text (STT) evaluation campaign on the CHIL lecture meeting data for three conditions: Multiple distant microphone (MDM), single distant microphone (SDM), and individual headset microphone (IHM). The system building process is similar to the IBM conversational telephone speech recognition system. However, the best models for the far-field conditions (SDM and MDM) proved to be the ones that use neither variance normalization nor vocal tract length normalization. Instead, feature-space minimum-phone error discriminative training yielded the best results. Due to the relatively small amount of CHIL-domain data, the acoustic models of our systems are built on publicly available meeting corpora, with maximum a-posteriori adaptation applied twice on CHIL data during training: First, at the initial speaker-independent model, and subsequently at the minimum phone error model. For language modeling, we utilized meeting transcripts, text from scientific conference proceedings, and spontaneous telephone conversations. On development data, chosen in our work to be the 2005 CHIL-internal STT evaluation test set, the resulting language model provided a 4% absolute gain in word error rate (WER), compared to the model used in last year's CHIL evaluation. Furthermore, the developed STT system significantly outperformed our last year's results, by reducing close-talking microphone data WER from 36.9% to 25.4% on our development set. In the NIST RT06s evaluation campaign, both MDM and SDM systems scored well, however the IHM system did poorly due to unsuccessful cross-talk removal.

1 Introduction

Integrated Project CHIL (“Computers in the Human Interaction Loop”), funded by the EU Information Society Technologies programme, aims to create people-friendly computing by monitoring how people interact, exchange information and collaborate to solve problems, learn, or socialize in meetings. These goals are not possible without a detailed understanding of the human state, human activities, and intentions. An important initial step towards this goal is to be able to automatically generate transcripts of the conversational speech under “always-on” audio capturing from far-field (non-intrusive) microphone sensors.

Conversational speech recorded by distant microphones in noisy reverberant settings poses significant challenges to state-of-the-art automatic speech recognition (ASR) technology. Furthermore, a number of factors such as multiple speakers with often overlapping speech, as well as the non-native accents and technical content of CHIL seminars create additional ASR challenges. With limited in-domain CHIL data available, it is natural to leverage useful out-of-domain data such as the Fisher and broadcast news corpora, available through the Linguistic Data Consortium (LDC) [1]. To face these challenges, during the first CHIL-internal ASR evaluation campaign in January 2005 (“CHIL evaluation run #1”), we opted to employ a number of in-house available acoustic models and combine them using the ROVER technique [2], after adaptation on limited CHIL data [3]. For language modeling, we used a small amount of meeting transcripts and conference proceeding texts, interpolated with Fisher data. However, the approach proved ineffective, resulting for example in the relatively high word error rate of 36.9% on close-talking data [3]. Furthermore, such system would not have been allowed entry into the NIST RT06s evaluation, due to its use of non-publicly available data sources (in addition to LDC corpora) for acoustic model training.

As a result, we have opted to start development “from scratch” for participating in this year’s NIST-sponsored RT06s campaign. We directed most of our effort in organizing most publicly available meeting corpora [1] and in training acoustic models based exclusively on such data. The effort was augmented by training and optimizing language models appropriate for the task. In contrast, no effort has been made to develop and incorporate front-end processing techniques for signal-space noise reduction and combination of distant (far-field) microphones. Overall, the same training procedures were shared between distant and individual headset microphone conditions, of course with a number of small variations. Our progress in this effort was benchmarked over the test set of the first CHIL-internal evaluation (“CHIL evaluation run #1”), which constitutes a subset of the RT05s development data.

The remainder of the paper is structured as follows: The data resources used are described in Section 2. Section 3 describes the system basics including the front end, segmentation, acoustic models, lexicon, and the language model. Results with comments are presented in Section 4, with Section 5 concluding the paper.

2 Data Resources

We briefly describe the corpora used for training, development, and evaluation of our acoustic models for both far-field and close-talking conditions. In addition to these, a number of publicly available text sources were also used for language modeling, as discussed in Section 3.4.

2.1 Training Data

The following meeting resources, available to all RT06s participant sites, were used for acoustic model training:

- ICSI meeting data, about 70 hours.
- NIST meeting pilot corpus, about 15 hours.
- RT04 development and evaluation data, about 2.5 hours.
- RT05s development data (excluding the CHIL 2005 evaluation data – see Section 2.2), about 4.5 hours.
- AMI meetings, about 16 hours.
- CHIL 2006 development data, about 3 hours.
- Additional CHIL data from the CHIL evaluation dry run (June 2004) and an intermediate collection during the Summer of 2004, for a total of about 4 hours.

With the exception of the last source (no far-field data were used), all other datasets provided both close-talking and multiple far-field microphone data. For training on the latter, we selected all table-top microphones present in the corpora. An exception was made for the AMI data, where four microphones from the eight-element circular microphone arrays were selected based on their location in the room or their SNR estimate. This approach resulted to approximately 473.5 hours of training data for the far-field MDM/SDM systems. For the IHM condition, all available close-talking channels were used, resulting to about 124 hrs of training. Notice that additional available and relevant data to this task, such as the RT05s evaluation data set and the TED corpus [1] were not used due to lack of time.

2.2 Development and Evaluation Data

In order to benchmark improvements and guide our development, we chose as development data the evaluation set of the CHIL-internal evaluation of January 2005 (“CHIL evaluation run #1”). This set provided us with about 1.8 hours of IHM data and four table-top microphones for a total of 8.7 hours. The set was later made available as part of the development data for the NIST RT05s evaluation campaign of last year, and it will be further referred to in this paper as the “CHIL eval05” set.

For evaluation, the lecture meeting data part of the RT06s test set was used. This contained a total of 190 minutes of data recorded in the smart rooms of five CHIL sites: AIT, IBM, ITC, UKA, and UPC. Multiple headset and table-top microphones were present in these data, with the choice of microphones to be used in the evaluation designated by NIST. IBM participated in three conditions, namely the:

- *Multiple distant microphone* (MDM) condition, which constituted the *primary* condition of this evaluation, with typically all table-top microphones allowable for use;
- *Single distant microphone* (SDM), with only one table-top microphone allowed to be used, specified by NIST; and the
- *Individual head microphone condition* (IHM), where all headset microphone channels were to be decoded.

Note that in contrast to the MDM and SDM conditions, where all speech was to be decoded, in the IHM condition the purpose was to recognize the speech of the person wearing the headset (i.e., decoding cross-talk was penalized).

3 The IBM Systems

For this evaluation, two main systems were developed by the IBM team: One for far-field acoustic conditions, with obvious MDM- and SDM-condition specific variants, and one system for the IHM condition. Both used an identical language model and a quite similar acoustic model training procedure. Details are described in the next sub-sections.

3.1 Front-End, Segmentation, and Cross-Talk Removal

The features used to represent the acoustic signal for recognition 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 9 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 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 of this bank 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 was applied to either MDM data or IHM data.

For segmentation, the following procedure is employed: We first segment audio files into speech and non-speech segments, followed by deletion of the non-speech segments. We use an HMM-based segmentation system that models speech and non-speech segments with five-state, left-to-right HMMs with no skip states. The output distributions in each HMM are tied across all states in the HMM, and are modeled with a mixture of diagonal-covariance Gaussian densities. The speech and non-speech models are obtained by applying a likelihood-based, bottom-up clustering procedure to the speaker-independent acoustic model. The speech segments are then segmented into homogeneous segments using Ajmera and Wooters' change-point detection procedure [4]. This is followed by a clustering procedure to cluster the segments into pseudo-speaker clusters that can then be used for speaker adaptation. We use the following clustering mechanism: All homogeneous speech segments are modeled using a single Gaussian density, and are clustered into a pre-specified number of clusters using K -means and a Mahalanobis distance measure. For the lecture meeting data, the number of speaker clusters is set to four.

For cross-talk removal, a simple algorithm is used to pre-process the close-talking microphone signal, before passing it to a speech recognizer. The objective is to detect the signal segments containing foreground speech and, in effect, reduce the insertion error rate of the recognizer, which typically has difficulty

distinguishing foreground (further also referred as “speech”) from background speech or noise (further also referred as “non-speech”). The algorithm consists of two parts:

- Frame labeling, where each frame is independently labeled as speech or non-speech. Frames are overlapping segments of signal assumed homogeneous with respect to speech/non-speech signal analysis.
- Post-processing, where frames are grouped into larger speech or silence segments.

The frame labeling algorithm is very similar to [5]. There are specific situations, however, where the algorithm fails, for example when the foreground speaker talks very quietly, or the microphone is incorrectly placed relatively far from the wearer’s mouth.

3.2 Acoustic Modeling

The speaker-independent (SI) acoustic model is trained on 40-dimensional features generated by an LDA projection of PLP features (see previous section), 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. The number of mixtures for a tied state s with C_s observations is given by $4 \times C_s^{0.2}$. The final mixture distributions are obtained as the result of a splitting procedure, starting with single Gaussian distributions. Intermediate mixture distribution estimates are obtained using the expectation-maximization (EM) algorithm updating the mixture weights, means and covariances. In addition, the model uses a global semi-tied covariance [6,7] linear transformation, 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 the silence HMM, which is a single-state model. The system uses 45 phones, among which 41 are speech phones, one is the silence phone, and three are noise phones, modeling background noise, vocal noise, and breathing noise. The MDM HMMs use 6,000 context-dependent tied state distributions, obtained by decision tree clustering of quinphone statistics using context questions based on 73 phonetic classes. The total number of Gaussian densities is about 200k. The IHM system is somewhat smaller, with 5000 context-dependent tied states and about 120k Gaussians. Since only 5% of the training data is from CHIL, MAP-adaptation of the SI model was deemed necessary to improve performance on CHIL data.

From this SI model, three different MDM models are derived based on whether variance normalization and/or vocal tract length normalization (VTLN) [8,9] are used. The three speaker adaptive training (SAT) [6,7] models are described in the following:

- **Model A:** This model is trained directly after the SI model on features in a linearly transformed feature space resulting from applying fMLLR transforms to the SI features. fMLLR transforms are computed on a per-speaker basis for all speakers in the training set.

- **Model B:** The SI features are further normalized with a voicing model (VTLN) with no variance normalization. The frequency warping is piecewise linear using a breakpoint at 6500 Hz. The most likely frequency warping is estimated from among 21 candidate warping factors ranging from 0.8 to 1.2 using a step of 0.02. Warping likelihoods are estimated using a voicing model. This model uses the same state tying as the SI model, however its states model emissions using full covariance Gaussian distributions, and the model is based on 13-dimensional PLP features.

A VTLN model is subsequently trained on features in the VTLN warped space. VTLN warping factors are estimated on a per-speaker basis for all data in the training set 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 HMMs have 10k tied states and 320k Gaussians.

Following VTLN, SAT Model B is trained on features in a linearly transformed feature space resulting from applying fMLLR transforms to the VTLN normalized features. fMLLR transforms are computed like the VTLN warping factors on a per-speaker basis for all speakers in the training set. The HMMs have 10k tied-states and 320k Gaussians.

- **Model C:** SAT Model C is trained with the same procedure as Model B, except that variance normalization is now applied.

Following training of SAT models A, B, and C, we estimate feature-space minimum phone error (fMPE) transforms [10] for all three. 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 are 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 [11].

Note that in contrast to the MDM/SDM system, the IHM system is only trained with the procedure in Model C. A total of 5600 context-dependent states and 240k Gaussians are used in this IHM system.

3.3 Recognition Process for MDM Data

After the automatic segmentation and speaker clusters are determined, for each table-top microphone, a final system output is obtained in 3 passes:

- The SI pass uses MAP-adapted SI models to decode.
- Using the transcript from step a., 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 is also trained with MAP on the CHIL data.

- c. The lattices resulting from step b. are rescored using an interpolated language model. The one-best at this step will be referred to as CTM-n, where n stands for model A, B, or C.

Clearly, we have three different MPE models at step b, therefore we have three outputs CTM-A, CTM-B, CTM-C for *each* available table-top microphone.

The final output for MDM is obtained using ROVER in the following way: First, for each table-top microphone, we combine the three MPE systems with no empty hypothesis present, the sequence being CTM-A, CTM-B, and CTM-C; following this, we combine the resulting system outputs over the multiple table-top microphones, with the empty hypothesis present. This ROVER arrangement gives us the best results on the development data.

3.4 Language Model

For language modeling, we constructed three separate four-gram models, which were smoothed with modified Kneser-Ney smoothing [12]: One based on 1.5M words of meeting transcript data; a second one using 37M words of scientific conference proceedings (primarily from data processed by CHIL partner LIMSI); and finally, one based on 3M words of Fisher data. To construct the language model (LM) used for the static decoding graph, we interpolated these models with weights of 0.56, 0.37, and 0.07, respectively, and used entropy-based pruning [13] to reduce the resulting model to about 5M n-grams. For the LM employed in lattice rescoring, we used the interpolated models with no pruning.

A 37k-word lexicon was 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 were based on a 45-phone set (41 speech, one silence phone and three noise phones), and were obtained from the Pronlex lexicon, augmented with manual pronunciations.

4 Results and Discussion

As mentioned earlier, we used the CHIL 2005 evaluation data set (“CHIL eval05”) as our development set to allow benchmarking progress from last year’s CHIL-internal evaluation. We first report development set results, followed by results on the RT06s evaluation set (“CHIL eval06”).

4.1 Development Set Results

We first summarize results concerning the acoustic modeling steps discussed in Section 3.2, using the reference segmentation for the MDM condition. Table 4.1 illustrates the gains at each stage, with each word error rate (WER) number obtained by using ROVER on the recognizer output from all four table-top microphones in the CHIL eval05 data (one table-top microphone was omitted due to its very low SNR). The final combined output is obtained by the two-level ROVER process discussed in Section 3.3.

Table 1. Word error rates, %, of MDM systems on the CHIL eval05 dataset, using the reference segmentation

System	Model A	Model B	Model C
SI	55.4	55.4	55.4
MAP-SI	53.1	53.1	53.1
VTLN	—	53.8	55.0
SAT	52.1	51.4	53.4
fMPE+MAP-MPE	46.6	47.5	49.0
final ROVER		45.6	

While Model B provides the best SAT result, Model A gives the best MPE result, with no variance normalization and no VTLN. Discriminative training provides up to 5.5% absolute gain over the SAT system. Applying ROVER on the three MPE systems adds another 1% absolute gain. We see MAP adaptation gaining 2.3% absolute at the initial SI decoding, the output transcripts of which are used for computing the VTLN warping factor and fMLLR transforms. MAP-MPE is necessary after fMPE, for otherwise, MPE after fMPE erases all the gain obtained by fMPE. MAP-MPE gives 0.6% absolute gain on top of fMPE. The reason for MAP-MPE may be that the acoustic conditions of CHIL data are quite different from other meeting data resources.

Table 2. Perplexity and OOV rates of the old LM (used in the CHIL 2005 evaluation) and the newly designed LM

Set	CHIL eval05		CHIL eval06
	old LM	new LM	new LM
perplexity	136.7	110.4	119.0
OOV rate (%)	3.9	0.4	1.0

Compared to our last year's LM [3], the new LM incorporates all meeting transcripts available, and a lot more conference proceedings text. Table 2 lists perplexity values and the out-of-vocabulary (OOV) rate of the old vs. new LM on the CHIL eval05 data. Clearly, the new LM results in significantly reduced values of both perplexity and OOV rate. Notice however that the two LMs have drastically different vocabulary sizes of 20k (old) versus 37k (new) words. It is also interesting to note that the new LM generalizes well to the CHIL eval06 data, achieving reasonable perplexity and OOV rate, as depicted in Table 2.

In LM rescoring, silence and noise words are treated as transparent words; while in the static decoding graph, the probabilities of silence and noise are estimated from training data. We optimize the probability value of silence with one table-top microphone in the development data and use it for evaluation data. Table 3 presents the comparison of each table-top microphone before and after the LM rescoring. LM rescoring seems to normalize the text, and the ROVER effect decreases a bit for Model A. The final ROVER result is 0.6% better than

Table 3. Far-field WERs on the CHIL eval05 set using LM rescoring and two-level ROVER over channels and models

Table Mic	Model A	Model B	Model C
	MPE/LM rescoring	MPE/LM rescoring	MPE/LM rescoring
1	51.3 / 48.7	52.0 / 49.6	53.5 / 51.3
2	53.4 / 51.1	54.1 / 52.0	55.0 / 52.7
3	53.5 / 51.7	53.8 / 51.7	55.1 / 52.7
4	53.9 / 51.7	54.4 / 51.9	55.3 / 53.5
ROVER	46.6 / 47.8	47.5 / 46.8	49.0 / 49.0
final ROVER		45.6 / 45.0	

the MPE final ROVER result. Table 3 also demonstrates significant gains from applying ROVER on four table-top microphones (about 4.5% absolute for all three MPE systems).

The models for IHM data are trained as those of Model C. Table 4 presents the gains obtained at each stage of the decoding process using reference segmentation for IHM systems, and the comparison of the old/new LMs. Clearly the new LM gives about 4% absolute gain. MAP-MPE adds 1.4% absolute on top of fMPE, which is more than the gain on MDM data. This is because there exist relatively more CHIL IHM data for MAP adaptation than CHIL far-field data.

The above results are obtained using the reference segmentation. Table 5 shows WERs of Model A on our automatic segmentation of the CHIL eval05 MDM data. Surprisingly, the results are much better (2% – 3% absolute) than those from human segmentation. This may be due to the fact that human transcribers tend to chop a whole sentence for easier processing.

4.2 Evaluation Set Results

Table 6 depicts the WER of the MDM system at various stages of its pipeline, as well as the final system WER, reported on the NIST RT06s lecture meeting data (“CHIL eval06” set), with the overlapping speaker number set to one. Each WER number is obtained by applying ROVER over the table-top microphones of each CHIL site. The final ROVER result is generated by the two ROVER processes discussed in Section 3.3, following the CTM-A, CTM-B, CTM-C sequence.

Table 4. WERs, %, of IHM systems on CHIL eval05 data using reference segmentation

System	old LM	new LM
SI	39.8	35.6
MAP-SI	38.2	34.3
VTLN	38.1	—
SAT	36.9	33.4
fMPE	33.4	29.9
fMPE+MAP-MPE	—	28.5
MLLR	—	26.8
LM rescoring	—	25.4

Table 5. Comparison of WER results, %, using human (reference) versus automatic segmentation on the CHIL eval05 MDM data

system	reference segmentation	automatic segmentation
SI	55.4	53.5
fMPE+MAP-MPE	46.6	43.7

Table 6. MDM system WER on the CHIL eval06 set using automatic segmentation, reported with overlapping speaker number set to one

System	Model A	Model B	Model C
MAP-SI	60.9	60.9	60.9
fMPE+MAP-MPE	51.5	50.6	51.2
ROVER	49.9		
LM rescoring	52.3	50.6	51.4
final ROVER	50.1		
IBM official submission	51.1		

It turns out that Model B produces the best MPE results on the CHIL eval06 data, while Model A is the worst of all three. Furthermore, LM rescoring hurts performance, with the ROVER result on the MPE outputs being 0.2% better (absolute) than the final ROVER system output. Notice also that the IBM official system submission was scored at 51.1% WER, as compared to the 50.1% reported in Table 6. This is due to the fact that, inadvertently, ROVER over the microphone channels was not consistently applied for all CHIL site data. Of course, this inconsistency did not affect the SDM submission, which achieved a 51.4% WER, since only one microphone channel is used in that condition. Nevertheless, the IBM submission for both MDM and SDM conditions scored well in the RT06s STT evaluation on the CHIL seminar (lecture meeting) data. Interestingly, compared to results for CHIL eval05 in Table 3, results for CHIL eval06 are much worse. Inferred from Table 2, the large WER differences between CHIL eval05 and CHIL eval06 sets may be due to acoustic mismatch.

Unfortunately, the IBM submission for the IHM condition has been very unsatisfactory, yielding 52.8% WER. This is mostly attributable to the poor performance of cross-talk removal. Indeed, without applying the cross-talk removal algorithm (i.e., depending on the automatic segmentation alone), the WER gets significantly reduced to 39.5%. Further analysis of the results shows that if all cross-talk was successfully removed, the WER would have been 33.3%; further, if none of the segments were deleted, and assuming that they would have been correctly decoded, the WER would have dropped to 28.7%. It is also interesting to note that decoding the CHIL eval06 IHM data using the *manual* (reference) segmentation (instead of the automatic segmenter) almost halves the WER of the IBM submission to 27.1%. This is of course due to cross-talk elimination and correct speaker segmentation in the manual transcripts, which also positively affect speaker adaptation in the system.

Table 7. Comparison of WER results, %, using manual (reference) versus automatic segmentation (with no additional cross-talk removal) on the CHIL eval06 IHM set. These results are significantly better than the IBM official submission of 52.8% that employed an unsuccessful cross-talk removal algorithm.

segmentation system	reference (manual)	automatic, with no cross-talk removal
SI	39.1	51.1
fMPE+MAP-MPE	29.5	41.0
MLLR	28.3	40.3
LM rescoring	27.1	39.5

5 Conclusions

We have made significant progress in the automatic transcription of CHIL meeting data. The main difference of our systems from last year is that we built systems from meeting data, instead of adapting existing out-of-domain models to CHIL data. This resulted in 11.5% absolute improvement (from 36.9% to 25.4%) on the close-talking evaluation data of the first CHIL-internal evaluation campaign, used as our development data. However, this progress did not translate into a satisfactory IHM system submission to the RT06s evaluation, due to the failure to successfully remove cross-talk present in the IHM channels. On the other hand, our development efforts paid off in the far-field conditions: Both our MDM and SDM system scored well in the RT06s evaluation campaign on lecture meeting data. Improvements in both acoustic and language modeling led to this success, with ROVER applied across three different acoustic models playing an important role in WER reduction. In particular for the MDM system, this was accompanied by applying ROVER across multiple table-top microphones, as a means of combining the available channels.

Acknowledgements

We wish to thank Karthik Visweswariah and John Hershey for organizing the RT04 and RT05s data corpora for training purposes. We would also 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. *The LDC Corpus Catalog*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. [Online]. Available: <http://www.ldc.upenn.edu/Catalog>
2. J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” in *Proc. Wksp. on Automatic Speech Recog. and Understanding (ASRU)*, Santa Barbara, CA, 1997, pp. 347–354.

3. S. Chu, E. Marcheret, and G. Potamianos, "Automatic speech recognition and speech activity detection in the CHIL smart room," in *Proc. Wksp. Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, UK, 2005, pp. 332–343.
4. J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. Wksp. on Automatic Speech Recog. and Understanding (ASRU)*, St. Thomas, US Virgin Islands, 2003, pp. 411–416.
5. A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: the ICSI-SRI Spring 2005 speech-to-text evaluation system," in *Proc. Rich Transcription 2005 Spring Meeting Recog. Eval.*, Edinburgh, UK, 2005, pp. 39–50.
6. M. F. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
7. G. Saon, G. Zweig, and M. Padmanabhan, "Linear feature space projections for speaker adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Salt Lake City, UT, 2001, pp. 325–328.
8. S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Atlanta, GA, 1996, pp. 339–341.
9. G. Saon, M. Padmanabhan, and R. Gopinath, "Eliminating inter-speaker variability prior to discriminant transforms," in *Proc. Wksp. on Automatic Speech Recog. and Understanding (ASRU)*, Trento, Italy, 2001, pp. 73–76.
10. D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 1, Philadelphia, PA, 2005, pp. 961–964.
11. D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Orlando, FL, 2002, pp. 105–108.
12. 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.
13. A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Wksp.*, Lansdowne, VA, 1998, pp. 270–274.