The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System

Andreas Stolcke[1,2], Xavier Anguera[2], Kofi Boakye[2], Özgür Çetin[3], Adam Janin[2], Mathew Magimai-Doss[2], Chuck Wooters[2], and Jing Zheng[1]

[1] SRI International, Menlo Park, CA, U.S.A.
[2] International Computer Science Institute, Berkeley, CA, U.S.A.
[3] Yahoo, Inc.
stolcke@speech.sri.com

Abstract. We describe the latest version of the SRI-ICSI meeting and lecture recognition system, as was used in the NIST RT-07 evaluations, highlighting improvements made over the last year. Changes in the acoustic preprocessing include updated beamforming software for processing of multiple distant microphones, and various adjustments to the speech segmenter for close-talking microphones. Acoustic models were improved by the combined use of neural-net-estimated phone posterior features, discriminative feature transforms trained with fMPE-MAP, and discriminative Gaussian estimation using MPE-MAP, as well as model adaptation specifically to nonnative and non-American speakers. The net effect of these enhancements was a 14-16% relative error reduction on distant microphones, and a 16-17% error reduction on close-talking microphones. Also, for the first time, we report results on a new "coffee break" meeting genre, and on a new NIST metric designed to evaluate combined speech diarization and recognition.

1 Introduction

This paper documents the latest in a series of speech recognition systems [1–3] jointly developed by SRI International and the International Computer Science Institute (ICSI) for participation in the annual NIST Rich Transcription evaluations focused on meeting processing (starting with RT-02S in Spring 2002, through RT-07 this year). We give a self-contained overview of the recognition system, while focusing on new aspects of the current version, including several improvements made since the evaluation proper.

Since the beginning of our research on meeting recognition, we have based our systems on existing systems developed for conversational telephone speech (CTS) recognition, by borrowing the decoding architecture and by adapting acoustic models trained originally on telephone corpora. This year, given increasing amounts of in-domain meeting training data, we evaluated if such an adaptation strategy is still worthwhile. We then focused on improvements to the acoustic preprocessing, which aims to minimize the mismatch between meeting speech and our existing acoustic models. New beamforming software for distant microphones and updates to the speech segmenter used for close-talking microphones resulted in improvements in their respective conditions.

Next, we applied several techniques to improve the way acoustic models originally trained on CTS and broadcast news (BN) speech are adapted to the meeting and lecture domain. One successful approach was the combination of three discriminative modeling techniques, at the level of features, feature transforms, and Gaussians [4], modified to work in an adaptive fashion. We also achieved gains by paying special attention to nonnative and non-American speakers in model adaptation, since those dialects are underrepresented in our background training corpora while being more pervasive in the meeting test data.

No significant changes were made to the language models, beyond incorporating additional training data from the Augmented Multi-party Interaction (AMI) project. As we will show, this additional data had limited effect, and improved results solely on AMI meeting test data.

2 Task and Data

2.1 Test data

Evaluation data The RT-07 evaluation data (eval07) was divided into three portions according to meeting genre: conference meetings (confmtg), lecture meetings (lectmtg), and coffee breaks (cbreak), the latter being a more interactive variant of the lecture room setup. The conference data consisted of excerpts from 8 meetings recorded at 4 sites in the U.S. and Europe (Carnegie Mellon University, Edinburgh, NIST, and Virginia Tech), totaling 3 hours in duration. The lecture data was collected at 5 different "Computers in the Human Interaction Loop" (CHIL) consortium sites and comprised 32 lecture excerpts totaling 2.7 hours. Coffee break data originated from the same 5 sites and added up to 0.7 hour.

Separate evaluations were conducted in three acoustic conditions:

MDM multiple distant microphones (primary)
IHM individual headset microphones (required contrast)
SDM single distant microphone (optional)

Lecture and coffee break rooms had more extensive instrumentation and provided the following additional conditions:

MSLA multiple source localization array microphones (optional)
MM3A multiple Mark-III microphone arrays (optional)
ADM all distant microphones (optional)

Although NIST evaluates recognition error on all speech, including portions where speakers overlap, our recognition system presently ignores this fact, and was optimized for non-overlapping speech. Consequently, all results presented here exclude overlapping speech in the distant-microphone conditions, unless noted otherwise.

Development data The NIST RT-06 (eval06), and to a lesser extent, RT-05 (eval05) evaluation data sets were used as development data. Lecture system development used eval06 only, and confmtg results on eval05 were somewhat discounted since eval05 contains one data source (ICSI) that yields very low error rates and does not occur in more recent test sets. Several system parameters (such as rescoring weights) had been optimized on even older NIST evaluation sets, and have not been re-tuned this year. Also, due to the paucity of lecture development data, those parameters were never tuned specifically for the genre, and simply copied from the confmtg system.

2.2 Training data

In-domain training data for the conference room consisted of the same meeting recordings from AMI, CMU, ICSI and NIST as used in previous years, plus additional data released by AMI and NIST since RT-06. The total amount of IHM data was about 213 hours after speech/nonspeech segmentation (AMI: 100 meetings, 100h; CMU: 17 meetings, 11h; ICSI: 73 meetings, 74h; NIST: 27 meetings, 28h).

The training data aimed at the lecture domain was unchanged from last year—due to time constrains we did not make use of some new lecture and coffee break data released prior to RT-07. As a result, the only lecture-type data used was about 7 hours of CHIL training data (close-talking microphones only), the CHIL dev06 distant-microphone development data, and about 9 hours of transcribed lectures available as part of the Translingual English Database (TED) [5].

As in previous years, we used background models trained on old CTS and BN corpora for adaptation to the meeting and lecture domains. These out-of-domain corpora included about 2300 hours of telephone speech from the Switchboard, CallHome English, and Fisher collections, and about 900 hours of BN data from the Hub-4 and Topic Detection and Tracking (TDT) corpora.

3 System Description and Development

3.1 Signal processing and segmentation

Distant microphone processing All distant microphone channels (in both training and test) were Wiener-filtered for noise reduction using a filter developed for the Qualcomm-ICSI-OGI Aurora system [6], identical to previous years [2].

Subsequently, for the MDM, MDM, MSLA, and MM3A conditions, a delay-and-sum beamforming technique was applied to combine all available distant microphone channels into a single "enhanced" channel. The algorithm used was essentially the same as last year [7], but used a new implementation that is freely available under the name BeamformIt (version 2.0) [8].

Once the enhanced signal was generated, speech regions were identified using a speech/nonspeech two-class hidden Markov model (HMM) decoder. Resulting segments were combined and padded with silence to satisfy certain duration constraints that had been empirically optimized for recognition accuracy. The algorithm and models were unchanged from last year [2]. Finally, the segments were clustered into acoustically homogeneous partitions, which served as pseudo-speaker units for normalization and adaptation. This aspect was also identical to last year's system.

Table 1. Comparison of old and new beamforming implementation in terms of word error rates (WER) using RT-06 recognition models.

	eval06 confmtg	eval06 lectmtg	
	MDM	MDM	ADM
RT-06 beamformer	34.2	55.5	51.0
BeamformIt v2.0	33.9	55.8	46.6

Table 2. Comparison of IHM speech/nonspeech segmentation without and with per-channel energy normalization for cross-channel feature computation, and for recognition from reference segments. eval06 results were obtained with the RT-06 recognition system, eval07 results with the current system.

	eval06		eval07	
	confmtg	lectmtg	confmtg	lectmtg
W/o energy norm.	24.0	30.8	25.6	29.5
with energy norm.	22.8	31.7	25.7	30.5
Reference seg.	20.2	29.3	22.8	28.1

To assess the effect of the new beamforming implementation on recognition performance, we reprocessed the eval06 data with BeamformIt, and then ran RT-06 confmtg and lectmtg systems that were otherwise unchanged. Table 1 shows that MDM performance is virtually unchanged, but that ADM is much improved. This seems to indicate that the new implementation is more robust to heterogeneous and/or very large sets of microphones.

Close-talking microphone processing The IHM input channels are segmented (without Wiener filtering) into speech and nonspeech regions using an HMM-based speech/nonspeech segmenter [9]. The segmenter is a two-class HMM decoder with each class represented by a three-state phone model. The states are modeled by 256-component multivariate Gaussian mixtures with diagonal covariance matrices. The segmentation proceeds via decoding of the full IHM channel waveform, potentially in a multi-pass fashion with decreased transition penalty between the speech and nonspeech classes. This is done so as to generate segments that do not exceed 60 seconds in length.

Last year we had introduced a combination of single- and cross-channel features designed to allow discrimination of foreground speech from cross-talk (which should not be recognized). The single-channel features consist of 12th-order Mel-frequency cepstral coefficients (MFCCs), log-energy, and first and second differences. The cross-channel features are maximum and minimum log-energy differences. The log-energy difference represents the log of the ratio of the short-time energy between a given target channel and a nontarget channel. The maximum and minimum values are selected to obtain a fixed number of feature components, given that the number of channels varies between meetings. All features are computed over a window of 25 ms advanced by 20 ms.

Following RT-06, we modified these features by normalizing the log-energies per channel prior to computing cross-channel features, with the goal of accounting for differences in noise floors and gains. This technique gave excellent results on conference meetings, eliminating cross-talk even from speakers for whom only distant-microphone recordings were available [9]. However, when we evaluated this new feature (per-

Table 3. Effect of adjusting speech/nonspeech prior probabilities. All results obtained with RT-07 recognition systems (hence eval06 results differ from Table 2).

	eval06	eval07		
	confmtg	confmtg	lectmtg	cbreak
Old priors	21.9	25.7	30.5	31.2
New priors	20.2	24.0	29.5	30.6
Reference seg.	19.1	22.8	28.1	29.5

channel energy normalization) on lecture data and current test sets, a mixed picture emerged, as shown in Table 2. It seems that the energy normalization does not improve the result on eval07 confmtg data, and in fact degrades accuracy on lecture data by about 1% absolute. Further investigation is needed to understand the reasons for this inconsistent behavior.

We also observed that there is still a considerable word error rate (WER) gap (1.5-3% absolute) between automatic and reference segmentation, largely because of a high deletion error rate. Running our confmtg recognizer on the AMI system's segmenter output gave a marked improvement, from 25.7% to 24.0% WER. In a post-evaluation experiment we tuned the speech/nonspeech prior probability used by the segmenter on eval06 confmtg data, and were able to obtain the same improvement. Furthermore, as shown in Table 3, the prior adjustment resulted in recognition improvements across all meeting genres.

No speaker clustering was performed on the IHM channels, since it was assumed that each IHM channel corresponds to exactly one speaker.

3.2 Acoustic modeling and adaptation

Decoding architecture To motivate the choice of acoustic models, we first describe the decoding architecture, which is unchanged from last year, depicted in Figure 1. An "upper" (in the figure) tier of decoding steps is based on MFCC features; a parallel "lower" tier of decoding steps uses perceptual linear prediction (PLP) features. The outputs from these two tiers are combined twice using word confusion networks (denoted by crossed ovals in the figure). Except for the initial decodings, the acoustic models are cross-adapted to the output of a previous step from the respective other tier using maximum likelihood linear regression (MLLR). Lattices are generated initially to speed up subsequent decoding steps. The lattices are regenerated once later to improve their accuracy, after adapting to the outputs of the first combination step. The lattice generation steps use noncrossword (nonCW) triphone models, and decoding from lattices uses crossword (CW) models. Each decoding step generates either lattices or N-best lists, both of which are rescored with a 4-gram language model (LM); N-best output is also rescored with duration models for phones and pauses [10].

The final output is the result of a three-way system combination of MFCC-nonCW, MFCC-CW, and PLP-CW decoding branches. The entire system runs in under 20 times real time (20xRT).[4]

[4] Runtimes given assume operation with Gaussian shortlists. Since RT-07 did not impose a runtime limit we ran the system without shortlists, in about 25xRT.

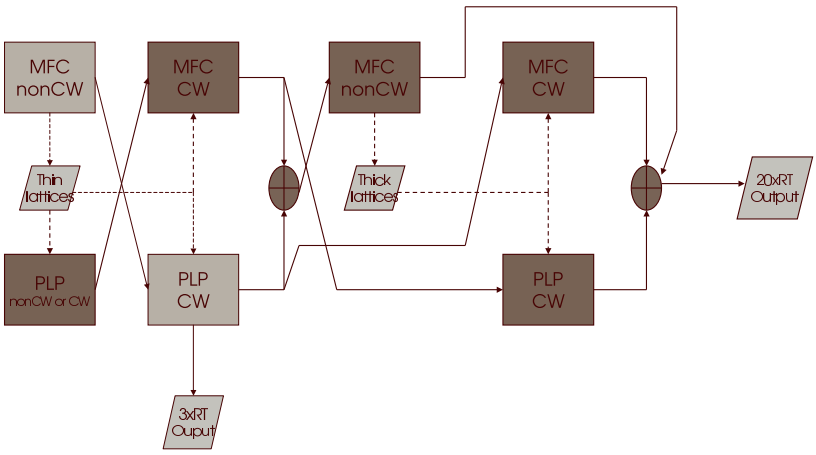

Fig. 1. SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote confusion network system combination steps.

Baseline models and test-time adaptation The MFCC recognition models were derived from gender-dependent CTS models in the RT-04F system, which had been trained with the minimum phone error (MPE) criterion [11] on about 1400 hours of data. (All available native Fisher speakers were used, but to save training time, statistics were collected from every other utterance only.) The MFCC models used 12 cepstral coefficients, energy, first-, second-, and third-order differences features, and 2×5 voicing features over a 5-frame window [12]. Cepstral features were computed with vocal tract length normalization (VTLN) and zero-mean and unit variance per speaker/cluster. The 62-component raw feature vector was reduced to 39 dimensions using heteroscedastic linear discriminant analysis (HLDA) [13]. After HLDA, a 25-dimensional Tandem/HATs feature vector estimated by multilayer perceptrons (MLPs) [14, 15] was appended. Both within-word and crossword triphone models were trained, for lattice generation and decoding from lattices, respectively. PLP models were based on full-bandwidth analysis, producing 12 coefficients, energy, first-, second- and third-order differences, and then reduced to 39 dimensions using HLDA. (No voicing or MLP features were used in this case.) These models were originally trained on about 900 hours of broadcast news data from the Hub4, TDT2, and TDT4 collections. PLP models are gender independent. All models were trained using decision-tree-based state tying.

In testing, all models undergo unsupervised adaptation to the test speaker or cluster, using MLLR with multiple, data-induced regression class trees. The first MFCC and PLP adaptation passes used a phone-loop reference model; later passes adapted to prior recognition output. In addition, all but the first decoding used constrained MLLR in feature space, which was also employed in training (speaker adaptive training) [16].

MLP feature adaptation As in past years, we adapted the MLPs for Tandem and HATs feature computation to the meeting domain by running additional MLP training

Table 4. Meeting recognition results using CTS training data, using MFCC maximum likelihood models and a simplified, 1-pass recognition system.

Training data	eval05 IHM confmtg
Fisher 400h	34.0
Confmtg 100h, 8kHz	33.4
Confmtg 100h, 16kHz	31.7
Fisher + confmtg, 8kHz (pooled)	31.9
Fisher + confmtg, 8kHz (MAP)	31.5

iterations on meeting data, starting with the CTS-trained MLPs. We showed previously that this type of adaptation yields about the same improvements as MAP adaptation of Gaussians alone [17]. In fact, as an expedient we used the adapted MLPs from last year, that is, without taking advantage of the new acoustic training data and using conference meeting data only. For distant-microphone recognition, the MLPs were adapted to both distant and close microphone recordings, whereas MLPs for IHM recognition were trained on close-talking microphones only.

Acoustic model adaptation In preparation for this year's evaluation, we conducted several experiments to determine the best training strategy. First and foremost, we wanted to confirm that adapting CTS models to the meeting domain was still a profitable approach. It entails downsampling meeting data to 8 kHz, raising the question of whether or not the attendant loss of information was more than compensated for by the added data. Table 4 summarizes some relevant results.

Models were trained on 400 hours of Fisher CTS data, as well as on the 100 hours of meeting speech available for RT-06, and tested on eval05 confmtg. We found that the downsampling of meeting data indeed incurs a significant, 6% relative error rate increase. However, this was almost made up for by simply pooling the CTS and (downsampled) meeting data. By using MAP adaptation, which gives control over the weighting of the in-domain versus background data, we were able to do slightly better than the meeting-only broadband models (31.5% versus 31.7% WER). Considering that the actual amount of CTS background data available is 5 times the 400 hours used in this experiment, we concluded that it was a safe bet to continue the MAP-adaptation strategy.

The next issue we addressed was the high percentage of nonnative and non-American speakers in the meeting and lecture data. Spot-checking the eval06 lecture data, for example, we found that almost all of it involved speakers with various European accents, most of them nonnative. The mismatch to our CTS background data was exacerbated by the fact that nonnative and non-American speakers had been excluded from our CTS training set (in accordance with past CTS evaluation sets). We therefore collected this previously excluded CTS data in a separate adaptation training set, comprising 220 hours in 1324 conversation sides, and performed tests on eval06 lectmtg data, summarized in Table 5.

The results are quite dramatic, in that adapting the background models to nonnative/non-American CTS data yields better performance than adapting to confmtg data. This clearly indicates that nativeness is one of the major factors of mismatch between the CTS and meeting data. As is to be expected, combining confmtg and

Table 5. Meeting recognition results using adaptation to nonnative and non-American CTS speakers, using MFCC ML-MAP models based on native-English Fisher data and a simplified, 1-pass recognition system.

MAP adaptation data	eval06 IHM lectmtg
confmtg 100h	41.9
Fisher nonnative/non-American 220h	40.5
confmtg + Fisher nonnat./non-Am.	40.0

Table 6. Results with different MAP adaptation criteria using complete recognition systems.

(a)

	eval06 IHM	
Adaptation method	confmtg	lectmtg
ML-MAP	22.8	34.1
MMI-MAP	n/a	29.8
fMPE-MAP	22.3	28.7
fMPE-MAP+MPE-MAP	22.2	26.3

(b)

	eval06 MDM	
Adaptation method	confmtg	lectmtg
ML-MAP	33.7	58.3
fMPE-MAP+MPE-MAP	30.9	48.6
+ML-MAP(lect-dev06)	n/a	47.8

nonnative/non-American CTS data in adaptation yields the best results. As a result of these experiments, we added the previously excluded Fisher speakers to our meeting adaptation data for MFCC model training. Note that this data was not added to the BN-based PLP model training data, both because of the bandwidth mismatch and because BN data is already more heterogeneous in its dialectal makeup.

fMPE-MAP In addition to MLP feature adaptation and MAP adaptation of the Gaussian models, we employed a discriminative feature transform known as fMPE (feature MPE) [18]. A sparse high-dimensional feature vector generated by Gaussian posteriors is mapped to the standard low-dimensional feature space via a transform trained using the minimum phone frame error (MPFE) [11, 19] criterion, and combined additively with the standard features. However, we used a novel variant of fMPE called fMPE-MAP, in which the transform is estimated only on adaptation data, based on a pretrained non-fMPE reference model (our CTS and BN background models). We found that fMPE-MAP gave better results than fMPE on the combined background and in-domain data, while taking much less training time [20]. The Gaussian posteriors input to the fMPE transform were based on PLP features from a 5-frame window, for both the MFCC and PLP fMPE-MAP models.

Table 6(a) compares results with ML-MAP, MMI-MAP (the method used last year), fMPE-MAP, and fMPE-MAP followed by MPE-MAP for IHM recognition, using complete recognition systems in which both MFCC and PLP models had been trained using the respective estimation criteria. The discriminative methods yield small gains on confmtg data, but substantial gains on lectmtg data. Recall that almost all the adaptation data is from the confmtg domain, highlighting the fact that discriminative training greatly enhances the generalization of acoustic models. Also note that MPE-MAP still gives substantial gains on top of fMPE-MAP in the case of lectmtg test data. The combined WER reduction is by 2.6% relative on confmtg and by 23% relative on lectmtg.

Adaptation for distant microphone recognition Models for recognition from distant microphones were obtained by pooling all close-talking and distant-microphone data

Table 7. Effect of language model update on recognition performance, differentiated by test data source

LM	eval06 confmtg			
	IHM		MDM	
	AMI	non-AMI	AMI	non-AMI
2006	20.1	23.2	28.9	32.9
2007	19.6	23.1	26.9	33.4

for adaptation purposes (similar to MLP adaptation). Table 6(b) shows ML-MAP and fMPE-MAP+MPE-MAP results for MDM recognition. The gains from discriminative adaptation are again substantial: 8.3% for confmtg and 17% for lectmtg. However, since the adaptation set contained only a very small amount of in-domain MDM lecture data (the dev06 set), we felt that the models for that domain might be improved further by giving extra weight to the matched data. This was accomplished by a final ML-MAP step using lectmtg-dev06 data only. As shown in the last row of Table 6(b), this indeed yielded a further 1.6% relative error reduction. The resulting models were used in both lecture and coffee break recognition (since both were recorded under the same acoustic conditions).

3.3 Language models

Language models (LMs) for the RT-07 system had the same structure as in previous years, consisting of an interpolation of various genre-specific LMs, including conference transcripts, lectures, CTS, BN, web data, and conference proceedings [21]. LMs specific to confmtg and lectmtg genres were obtained by finding perplexity-minimizing interpolation weights on held-out data of the respective type.

The only change for this year's system was the addition of new AMI and NIST conference meeting transcripts. While this almost doubled the amount of in-domain LM data, we found only small gains in overall recognition accuracy, as shown in Table 7. Since most of the new data came from the AMI data collection, we broke eval06 recognition results down according to whether or not the test meeting came from an AMI site (Edinburgh or TNO). It becomes evident that the additional training data helps significantly on AMI test data, but not on other data. We attribute this to the special scenario-driven character of the AMI meetings. Still, since the RT-07 test set was expected to contain AMI sources as well, we incorporated the updated LM into our confmtg system. On lectmtg tests, however, the new LM data made no impact whatsoever, so we simply kept last year's lectmtg LM. The lecture LM was also used in coffee break recognition. We again note that, because of time constraints, none of the CHIL lecture data released since RT-06 was used in LM training.

3.4 Speaker clustering revisited

As mentioned, our distant-microphone recognition system groups waveform segments into pseudo-speaker clusters for feature normalization and model adaptation purposes. However, we had found in previous years that this clustering slightly degrades performance on lecture data, presumably because the lecture is dominated by a single speaker and the clustering algorithm is not accurate enough to identify small sets of non-lecturer

Table 8. Effect of acoustic clustering parameters on MDM recognition accuracy. Values chosen in the RT-07 evaluation system appear in boldface.

	eval06 MDM		eval07 MDM		
Clustering	confmtg	lectmtg	confmtg	lectmtg	cbreak
1 cluster		47.8		**44.6**	44.0
4 clusters	30.3		**26.2**		**44.7**
Unlimited	30.2	48.1	26.5	44.7	
Combined	29.4	46.9	25.8	43.7	43.5

speech. Therefore, the RT-07 system again used only a single cluster for lecture recognition.

Post-evaluation we revisited this decision and checked the effect of different clustering parameters for all genres. Three configurations were tried: 1 cluster (the default for lectmtg), 4 clusters (the default for confmtg, close to the average number of meeting participants, and optimized on old evaluation data), and an unlimited number of clusters (constrained only by a minimum amount of data per cluster). The results are summarized in Table 8.

First, we can note that the (blind) choices made for eval07 confmtg and lectmtg turned out to be optimal. The alternative clusterings resulted in minimal degradation only. For coffee break recognition, we had made a poor choice (4 clusters) based on the assumption that they would be more like conference meetings; a single cluster worked best here, too. Most interesting, the error patterns (substitution/insertion/deletion rates) resulting from alternate clusterings were quite different. This suggested combining the different systems by merging the confusion networks produced in their final stages. As shown in the last row of Table 8, this indeed yielded considerable reductions in error over the single best system, of between 0.4% and 1.0% absolute. (Of course, this gain comes at the price of doubled runtime.)

4 Overall Results

4.1 Conference Meetings

Table 9(a) compares results on last year's and this year's evaluation sets for the conference room condition. For last year's test data we also include results from last year's (RT-06) system, thereby allowing us to assess overall progress made. Furthermore, we list results with both the submitted RT-07 system and the improvements made post-evaluation (the retuned priors for IHM recognition and the cluster combination for MDM). On eval06, the progress on MDM data was about 11.4% relative (14.0% post-evaluation), and 8.8% on IHM data (15.8% post-evaluation). We also note that the MDM word error rate on non-overlapped speech is within 8% of IHM performance on eval07, although this looks like an artifact of this particular test set as (eval07 is easier than eval06 on MDM, but harder for IHM recognition).

4.2 Lectures and coffee breaks

Table 9(b) similarly summarizes all the results for the lecture room task, as well as for the new coffee break genre. For eval06 lectures, MDM word error was reduced

Table 9. Results on RT-06 and RT-07 test data summarized.

(a)

System	MDM	SDM	IHM
	eval06 confmtg		
RT-06	34.2	41.2	24.0
RT-07	30.3	40.6	21.9
Post-eval	29.4		20.2
	eval07 confmtg		
RT-07	26.2	33.1	25.7
Post-eval	25.8		24.0

(b)

System	MDM	ADM	MM3A	SDM	IHM
	eval06 lectmtg				
RT-06	55.5	51.0	56.5	57.3	31.0
RT-07	47.8	39.3		49.6	26.3
Post-eval	46.9				25.7
	eval07 lectmtg				
RT-07	44.6	42.1	54.0	50.6	30.5
Post-eval	43.6				29.5
	eval07 cbreak				
RT-07	44.7	41.1	51.0	50.0	31.2
Post-eval	43.5				30.6

13.9% relative (15.5% post-evaluation), and IHM error 15.2% relative (17.1% post-evaluation). The ADM condition saw an even greater improvement of 22.9% relative, largely because of improved beamforming. Comparing across test sets, we find that IHM became harder this year, whereas MDM became easier, similar to what we saw with conference data.

Finally, we observe that the RT-07 coffee break data shows errors across conditions that are remarkably similar to the corresponding lectmtg results. This, together with the earlier observations about speaker clustering and the fact than these results were obtained with lecture-tuned language model, led us to conclude that, for recognition purposes, the coffee break data is presently not significantly different from lecture data.

4.3 Speaker-attributed speech-to-text

This year NIST introduced a new "speaker-attributed speech-to-text" (SASTT) task, combining diarization and speech recognition (speech-to-text, STT). Systems label each recognized word with speaker tags, and the scoring program counts a word as correct only if both the spelling and the speaker label agree with the reference (speaker labels are treated as arbitrary and only significant to the extent that they indicate identity or nonidentity of speakers). The SASTT task is defined only for distant-microphone conditions.

We had not originally planned to develop a system for this task, but after the submission deadline we decided to generate SASTT output by a simple merging of our speech recognition output with ICSI's diarization output [22]. Each recognized word was labeled with the speaker label that has the longest time overlap with the word. Table 10 summarizes the results, which turned out to be highly competitive even without having performed any joint optimization on the diarization and STT systems.

We also tested a simple model that predicts SASTT error from the error rates and types of the underlying STT and diarization systems. If we assume that diarization errors occur independently of STT errors, we would predict that incorrect speaker labels cause about $\mathrm{ME}_{SPKR} + \mathrm{SE}_{SPKR}$ correct STT words to be SASTT-incorrect, where ME_{SPKR} and SE_{SPKR} are the diarization miss and speaker error rates, respectively. Therefore, we predict the SASTT WER error to be

$$\mathrm{WER}_{SATT} = \mathrm{WER}_{STT} + \mathrm{CorR}_{STT} \times (\mathrm{ME}_{SPKR} + \mathrm{SE}_{SPKR})$$

Table 10. Actual and predicted SASTT error rates obtained by a combination of the SRI-ICSI recognizer with the ICSI diarization system. The error rates of the component diarization and recognition systems are also given. Unlike elsewhere in this paper, the scoring here was performed with as many as three overlapping speakers.

Task	eval07 confmtg		eval07 lectmtg
	MDM	SDM	MDM
SASTT (actual)	40.3	51.7	56.9
SASTT (predicted)	41.9	55.2	58.6
STT	37.4	43.6	49.3
diarization	8.5	21.7	23.3

with CorR_{STT} being the STT word-correct rate. As the second row of Table 10 shows, this prediction is only a slight overestimate for the MDM condition. However, for the SDM condition, the formula overestimates SASTT error substantially, probably because under poor acoustic conditions, STT and diarization errors will be more highly correlated.

5 Conclusions and Future Work

We have made further progress in the recognition of conference and lecture meetings, with first results on "coffee break" data that are comparable to those on lectures. The most significant contributions this year came from a combination of discriminative techniques in acoustic modeling, including a new method, fMPE-MAP, that showed the most substantial error reductions on the "hard" tasks, namely, distant microphone recognition in general and lecture recognition in particular. Additional acoustic modeling gains came from adaptation to nonnative and non-American English telephone data. Acoustic preprocessing was improved by using a new beamforming implementation (for distant microphones) and retuning the speech/nonspeech priors (for close-talking microphones). We found a simple way to improve distant microphone recognition in combining multiple recognition systems differing only in their speaker clustering constraints. Finally, we constructed a first, yet competitive SASTT system by a straightforward merging of our STT system with ICSI's diarization output.

6 Acknowledgments

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811), and by the Swiss National Science Foundation through NCCR's IM2 project. Additional support came from the the Defense Advanced Research Projects Agency (DARPA) to SRI under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC). We thank Thomas Hain from Sheffield for making the AMI system's IHM segmenter output available for testing with our system.

References

1. Stolcke, A., Wooters, C., Mirghafori, N., Pirinen, T., Bulyko, I., Gelbart, D., Graciarena, M., Otterson, S., Peskin, B., Ostendorf, M.: Progress in meeting recognition: The ICSI-SRI-UW Spring 2004 evaluation system. In: Proceedings NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, National Institute of Standards and Technology (2004)
2. Stolcke, A., Anguera, X., Boakye, K., Çetin, Ö., Grézl, F., Janin, A., Mandal, A., Peskin, B., Wooters, C., Zheng, J.: Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system. In: Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation, Edinburgh, National Institute of Standards and Technology (2005) 39–50
3. Janin, A., Stolcke, A., Anguera, X., Boakye, K., Çetin, Ö., Frankel, J., Zheng, J.: The ICSI-SRI Spring 2006 meeting recognition system. In Renals, S., Bengio, S., Fiscus, J., eds.: Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006. Volume 4299 of Lecture Notes in Computer Science., Springer (2006) 444–456
4. Zheng, J., Cetin, O., Hwang, M.Y., Lei, X., Stolcke, A., Morgan, N.: Combining discriminative feature, transform, and model training for large vocabulary speech recognition. In: Proc. ICASSP. Volume 4., Honolulu (2007) 633–636
5. Lamel, L., Schiel, F., Fourcin, A., Mariani, J., Tillman, H.: The translingual English database (TED). In: Proc. ICSLP, Yokohama (1994) 1795–1798
6. Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S.: Qualcomm-ICSI-OGI features for ASR. In Hansen, J.H.L., Pellom, B., eds.: Proc. ICSLP. Volume 1., Denver (2002) 4–7
7. Anguera, X., Wooters, C., Pardo, J.M.: Robust speaker diarization for meetings: ICSI-SRI RT-06S meetings evaluation system. In Renals, S., Bengio, S., Fiscus, J., eds.: Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006. Lecture Notes in Computer Science. Springer (2007)
8. Anguera, X.: Beamformit (the fast and robust acoustic beamformer). http://www.icsi.berkeley.edu/~xanguera/beamformit/ (2006)
9. Boakye, K., Stolcke, A.: Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In: Proc. ICSLP, Pittsburgh, PA (2006) 1962–1965
10. Vergyri, D., Stolcke, A., Gadde, V.R.R., Ferrer, L., Shriberg, E.: Prosodic knowledge sources for automatic speech recognition. In: Proc. ICASSP. Volume 1., Hong Kong (2003) 208–211
11. Povey, D., Woodland, P.C.: Minimum phone error and I-smoothing for improved discriminative training. In: Proc. ICASSP. Volume 1., Orlando, FL (2002) 105–108
12. Graciarena, M., Franco, H., Zheng, J., Vergyri, D., Stolcke, A.: Voicing feature integration in SRI's Decipher LVCSR system. In: Proc. ICASSP. Volume 1., Montreal (2004) 921–924
13. Kumar, N.: Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. PhD thesis, Johns Hopkins University, Baltimore (1997)
14. Morgan, N., Chen, B.Y., Zhu, Q., Stolcke, A.: TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition. In: Proc. ICASSP. Volume 1., Montreal (2004) 536–539
15. Zhu, Q., Stolcke, A., Chen, B.Y., Morgan, N.: Using MLP features in SRI's conversational speech recognition system. In: Proc. Interspeech, Lisbon (2005) 2141–2144
16. Jin, H., Matsoukas, S., Schwartz, R., Kubala, F.: Fast robust inverse transform SAT and multi-stage adaptation. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, Morgan Kaufmann (1998) 105–109
17. Stolcke, A., Anguera, X., Boakye, K., Çetin, Ö., Grézl, F., Janin, A., Mandal, A., Peskin, B., Wooters, C., Zheng, J.: Further progress in meeting recognition: The ICSI-SRI Spring

2005 speech-to-text evaluation system. In Renals, S., Bengio, S., eds.: Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005. Volume 3869 of Lecture Notes in Computer Science., Springer (2006) 463–475
18. Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., Zweig, G.: fMPE: Discriminatively trained features for speech recognition. In: Proc. ICASSP. Volume 1., Philadelphia (2005) 961–964
19. Zheng, J., Stolcke, A.: Improved discriminative training using phone lattices. In: Proc. Interspeech, Lisbon (2005) 2125–2128
20. Zheng, J., Stolcke, A.: fMPE-MAP: Improved discriminative adaptation for modeling new domains. In: Proc. Interspeech, Antwerp (2007) 1573–1576
21. Çetin, Ö., Stolcke, A.: Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system. Technical Report TR-05-06, International Computer Science Institute, Berkeley, CA (2005)
22. Wooters, C., Huijbregts, M.: The ICSI RT07s speaker diarization system. Lecture Notes in Computer Science, Springer (2007) To appear.