The ICSI-SRI Spring 2006 Meeting Recognition System

Adam Janin[1], Andreas Stolcke[1,2], Xavier Anguera[1,3], Kofi Boakye[1], Özgür Çetin[1], Joe Frankel[1], and Jing Zheng[2]

[1] International Computer Science Institute, Berkeley, CA, U.S.A.
[2] SRI International, Menlo Park, CA, U.S.A.
[3] Technical University of Catalonia, Barcelona, Spain
janin@icsi.berkeley.edu

Abstract. We describe the development of the ICSI-SRI speech recognition system for the National Institute of Standards and Technology (NIST) Spring 2006 Meeting Rich Transcription (RT-06S) evaluation, highlighting improvements made since last year, including improvements to the delay-and-sum algorithm, the nearfield segmenter, language models, posterior-based features, HMM adaptation methods, and adapting to a small amount of new lecture data. Results are reported on RT-05S and RT-06S meeting data. Compared to the RT-05S conference system, we achieved an overall improvement of 4% relative in the MDM and SDM conditions, and 11% relative in the IHM condition. On lecture data, we achieved an overall improvement of 8% relative in the SDM condition, 12% on MDM, 14% on ADM, and 15% on IHM.

1 Introduction

Despite ongoing advances in automatic speech recognition technology, natural multi-person meetings continue to be challenging. The acoustic environment, especially with desktop microphones, is quite variable. Noises such as fans, door slams, and paper rustling all contribute to the acoustic background. Reverberation and echo can also be a significant problem. Typically, recordings from different sites (and even within the same site) use many different types of microphones. Another issue is that meetings contain large amounts of overlap — people end each other's sentences, interrupt, encourage ("uh huh"), laugh, and so on. Finally, the relative paucity of in-domain training data makes it vital to leverage methods and data that have been developed for other genres of speech, such as conversational telephone speech (CTS) and broadcast news (BN).

As for all our recent meeting evaluation systems, our development strategy for RT-06S was to base the system on the SRI-ICSI-UW RT-04F conversational telephone speech recognition system,[4] with improvements incorporated from the previous year's NIST evaluation systems [1, 2]. This year, we improved the delay-and-sum algorithm by using a global histogram to discard frames with low correlation and also by using delays selected among the N-best delay scores rather than only the one-best. The nearfield segmenter now uses cross-channel log-energy ratio features (in addition to Mel frequency cepstral coefficients [MFCCs]) integrated directly with a hidden Markov model

[4] As explained later, we also made use of acoustic models developed for BN.

(HMM) segmenter. Language models were updated by the inclusion of new conference and lecture room transcripts, as well as additional Web data. A new procedure was used to train the phone posterior features, including separate adaptation for both the nearfield and farfield sources (whereas the RT-05S system was adapted only from the nearfield data). HMM adaptation was improved using a data-induced regression class trees (rather than hand-crafted classes). For the lecture room condition, we used a small amount of additional in-domain training data, as well as data from the TED Corpus [3]. Finally, farfield models were trained using both the farfield and nearfield data (instead of just the farfield data).

The evaluation task and data are described in Section 2. Section 3 gives the system description, focusing on new developments relative to the 2005 system [1]. Results and discussion appear in Section 4, followed by conclusions and future work in Section 5.

2 Task and Data

2.1 Test data

Evaluation data The RT-06S conference room evaluation data (eval06) consisted of two meetings each from the University of Edinburgh, CMU (Carnegie Mellon University Interactive Systems Laboratory), NIST (National Institute of Standards and Technology), and VT (Virginia Tech), and one meeting from TNO (the Netherlands Organization for Applied Scientific Research). Systems were required to recognize a specific 18-minute segment from each meeting; however, data from the entire meeting was allowed for processing.[5] Separate evaluations were conducted in three conditions:

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

The lecture room data consisted of 120 minutes of seminars recorded by the Computers In the Human Interaction Loop (CHIL) consortium. In addition to the above conditions, lecture data included the following recording conditions:

ADM all distant microphones (optional)
MBF pre-beamformed signal from the Multiple Mark III microphone array (MM3A, optional)

Microphones varied substantially by type and setup, even within each condition. For example, some of the AMI IHM data were recorded with head-mounted lapel microphones, and MDM recording devices ranged from low- and high-quality individual table-top microphones to AMI's circular microphone arrays. Meeting participants included both native and nonnative speakers of English (unlike in CTS evaluations).

[5] We did not find significant gains from adapting on entire meetings, and, except in the acoustic preprocessing, used only the designated meeting excerpts.

Development data The RT-05S evaluation data were designated as development data for RT-06S, and used by us as an unbiased test set (designated eval05). For the conference room task, the data consisted of ten 12-minute excerpts of meetings from AMI, CMU, ICSI, VT, and NIST. For the lecture room task, the data consisted of 120 minutes of seminars recorded by the CHIL consortium. We also used the same development set as was used in RT-05S [1] for additional tuning.

2.2 Training data

Training data for the conference room task were identical to that used in RT-05S, and included data from AMI (35 meetings, 16 hours of speech after segmentation), CMU (17 meetings, 11 hours), ICSI (73 meetings, 74 hours), and NIST (15 meetings, 14 hours). The CMU data were of limited use in that only lapel and no distant microphone recordings were available. For the lecture room task, we included the small amount of available CHIL data that were not in the development sets.[6] These data consisted of only the nearfield signals from excerpts of 38 meetings, totaling about 7 hours of speech. We also included the Translingual English Database (TED) [3], using the boom-microphones only and consisting of 39 lectures for about 9 hours worth of speech.

Background training data for the (pre-adaptation) acoustic models consisted of the publicly available CTS and BN corpora. These 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 TDT corpora.

3 System Description

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 [4]. The process was identical to last year [1].

Subsequently, for the ADM and MDM conditions, a delay-and-sum beamforming technique was applied to combine all available distant microphone channels into a single "enhanced" channel. The system is very similar to the one used in the ICSI RT-06S speaker diarization system [5], and is based on last year's system [6] with two main improvements.

The first improvement affects the noise filtering based on the value returned by the generalized cross-correlation algorithm (GCC-PHAT [7]). Frames with a low correlation value indicate increased uncertainty as to whether the returned delay represents the actual TDOA (time delay of arrival). In last year's submission, we filtered out any value smaller than 0.1, assigning the previous nonfiltered delay to such frames (ensuring delay continuity). This caused fewer frames to be filtered in "cleaner" acoustic conditions than in noisy conditions or with worse microphones. This year's submission computed a global histogram of all delays in all channels and determined the threshold so that 10% of frames are dropped.

[6] These data were provided only after the evaluation had started.

Another improvement this year involves the delays selected among the N-best GCC-PHAT computed. At every position, we consider the tradeoff between selecting the main peaks of the GCC-PHAT function and ensuring a continuity on the selected delays in the region surrounding that point. To do so, we apply a two-step Viterbi decoding at two levels. First, at a channel level, we decide which 2-best delays are most probable in each position. Second, at a global level, all combinations of the local 2-best among all channels are considered, and the best combination is chosen. In each step, each possible state has an emission probability equal to the GCC-PHAT value for each delay/combination, and the transition probability between two nodes is inverse proportional to the distance between its delays/combinations, ensuring that the N-best probabilities in a particular instant sum up to 1. We apply a relative weight of 25 to emphasize the transition probabilities.

This newly introduced technique aims to find the optimum tradeoff between reliability (cross-correlation) and stability (distance between contiguous delays). Stability is vital, as our aim is to obtain an optimally improved signal, while avoiding quick changes of the beamforming between acoustic events.

Once the enhanced signal was generated, speech regions were identified using a speech/nonspeech two-class 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 [1]. Finally, the segments were clustered into acoustically homogeneous partitions, which serve as pseudo-speaker units for normalization and adaptation. The clustering algorithm was also identical to last year's system.

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 [8]. 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.

The segmenter uses both single- and cross-channel features for speech activity detection. The single-channel features consist of 12th-order Mel-frequency cepstral coefficients, 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 non-target 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. These cross-channel features are included specifically to address errors caused by cross-channel phenomena such as crosstalk. All features are computed over a window of 25 ms advanced by 20 ms.

A later (i.e., post-evaluation) enhancement to the system consisted of an energy normalization technique being applied prior to computing the log-energy difference features. For a given channel, the minimum frame log-energy of the channel is subtracted

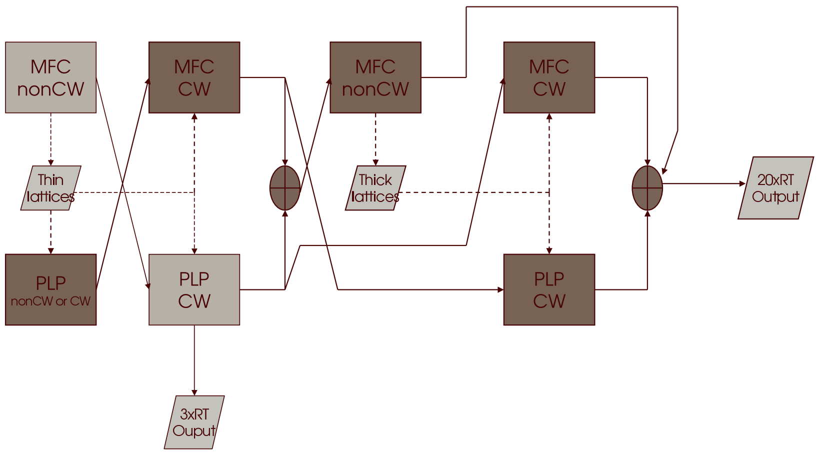

Fig. 1. SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote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

from all log-energy values in that channel. That is, for a channel i at frame n

$$E_{norm}(n) = E_i(n) - E_{min,i} \qquad (1)$$

where E represents log-energy. The minimum frame log-energy is used as an estimate of the noise floor and has the advantage of being largely independent of the amount of speech activity in the channel. This normalization was done to compensate for any significant differences in microphone gains and yielded substantial performance improvements over the unnormalized features.

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 SRI-ICSI-UW RT-04F CTS system, on which the meeting system is based (see 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 words and pauses [9].

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).[7] The "fast" subset consisting of just two decoding steps (the light-shaded boxes in the figure) runs in about 3xRT; it was used for quick-turnaround experiments, but was not used in this year's evaluation.

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 [10] on about 1400 hours of data. (All available native Fisher speakers were used, but to save training time, statistics were collected from only every other utterance). 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 [11]. 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) [12]. After HLDA, a 25-dimensional Tandem/HATs feature vector estimated by multilayer perceptrons (MLPs) [13, 14] 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 underwent unsupervised adaptation to the test speaker or cluster, using MLLR with multiple, data-induced regression class trees. This is in contrast to last year's system, which used hand-defined MLLR regression classes. 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) [15].

Acoustic model task adaptation Nearfield recognition models were adapted to ICSI, NIST, and AMI nearfield microphone data. For the various distant microphone tasks, a single model set was created by adapting to farfield data from ICSI, NIST, and AMI training meetings, plus ICSI, NIST, AMI, and CMU nearfield data. The inclusion of nearfield adaptation data for farfield model training is new this year, and was done with the rationale that farfield recognition involves signals ranging in quality from distant to near-close-talking. Just as in last year's system, we did not delay-sum the training data for the multiple microphone conditions; rather, we pooled all the individual microphone signals into one training set, and used the same pooled adaption data for all meetings. The weight for adaptation data statistics was empirically optimized, and set at 20.

Last year, we applied maximum a posteriori adaptation with a maximum mutual information criterion (MMI-MAP) only to the IHM models, and used the standard, less-involved maximum likelihood (ML) MAP procedure on the distant microphone

[7] Runtimes given assume operation with Gaussian shortlists. Since RT-06S did not impose a runtime limit we ran the system without shortlists, in about 25xRT.

models. This year, MMI-MAP was used for all PLP models, and ML-MAP for all the MFCC+MLP models.

MLP feature adaptation The MLPs used to estimate Tandem and HATS features were originally trained to perform frame-level phone discrimination using a large subset of the CTS training data [14]. To improve the match to the acoustic conditions of the meeting domain, these were adapted by applying four additional epochs of backpropagation using ICSI, AMI and NIST meeting data as training material. The Karhunen-Loeve transform (KLT) used to reduce the feature dimension from 46 (the size of the phone set) to 25 was kept unchanged from the CTS system, in order to keep the features compatible with existing models. Unlike the ICSI/SRI RT-05S system, in which MLPs were only adapted to nearfield sources, separate MLPs were adapted for the nearfield and farfield conditions. In the case of the HATS, only the merger MLP was adapted, and the 15 critical band networks were left unchanged. The initial learning rates were set to be equal to those at the conclusion of training of the CTS MLPs, and halved after each epoch. The input acoustic parameters used an 8-kHz front end to match that used in the original CTS MLP trainings.

The MLPs for the farfield adaptation were initialized with the nearfield-adapted MLPs after one epoch, and adapted only on regions where there was no overlapping speech. The labels were generated from alignments made on the nearfield data. Initial experiments followed our approach to farfield acoustic model adaptation, in which all available farfield channels were used as training material. Recognition experiments on a development set using these MLPs gave worse performance than using nearfield-adapted MLPs. One possible cause was overtraining, as the MLP was being presented with as many as eight noisy versions of each speech segment during each epoch. We therefore selected a single channel at random to provide the data for each segment (though input normalizations were calculated over all segments for a given speaker/channel combination). This approach led to improved results. Adapting the MLP features to the meeting domain led to reductions in word error rate (WER), in particular for the SDM and MDM conditions, and on the lecture data.

3.3 Language models

Three LMs were used in decoding: a multiword bigram for lattice generation, a multiword trigram for decoding from lattices, and a word 4-gram for lattice and N-best rescoring. The same set of language models is used for all conference meeting sources (we found no advantage in tuning LMs to the meeting source). A second set of LMs is used for the lecture task.

For the conference room domain, the LMs were linearly interpolated mixtures of component LMs trained from the following sources: (a) Switchboard CTS transcripts, (b) Fisher CTS transcripts, (c) Hub-4 and TDT4 BN transcripts, (d) AMI, CMU, ICSI, and NIST meeting transcripts, and (e) World Wide Web data newly collected to match different topics and styles, namely, RT-04S meeting sources and AMI meetings, and 525M words of Fisher-like conversational Web data collected and published by the University of Washington for the RT-04F evaluation. The mixture weights were tuned to minimize perplexity on heldout AMI, CMU, ICSI, LDC, and NIST transcripts. The

LM vocabulary consisted of 54,524 words, comprising all words in our CTS system (including all Hub-5 and all nonsingleton Fisher words), all words in the ICSI, CMU, and NIST training transcripts, and all nonsingleton words in the AMI training transcripts. The out-of-vocabulary rate was 0.40% on eval04 transcripts, and 0.19% on the 2005 AMI development transcripts.

For the lecture room domain, additional LM mixture components were built from (f) 70K words of CHIL development transcripts and (g) 32M words of speech conference proceedings (suggested by [16]). Also, the Fisher-relevant Web data were replaced by about 512M words of the newly collected Web data related to the CHIL transcripts. The lecture LM mixture was then optimized on CHIL development transcripts. The lecture LM vocabulary was an extension of the conference LM vocabulary, with 3791 additional frequent words found in the proceedings data. The out-of-vocabulary rate on the CHIL development data was 0.18%.

The main difference of this year's conference and lecture LMs compared to last year's is in the Web data LM component. The new Web data collected this year employed a different selection criterion for the n-gram queries submitted to the search engine. Instead of using the most frequent 4-grams in the target corpus, we used the 4-grams with the highest likelihood ratio between a target LM trained on the available meeting or lecture data, and a background LM from all the other data [17]. However, overall, we did not see any significant perplexity or WER improvement over last year's LMs in the eval05 conference test set (the perplexity of the final pruned 4-gram meetings LM was 115); we therefore kept the 2005 LM in our conference evaluation system. On the eval05 lecture task, the new LMs reduced perplexity about 5% relative, to 119, but this improvement did not bring any significant improvement in WER. Nevertheless, the updated lecture LMs were used in the evaluation system, because they were thought to provide better coverage for the new test sets due to the inclusion of more recent Web data, and of the CHIL lecture transcripts.

4 Results and Discussion

Note that all results are reported on non-overlapping speech (using an overlap limit of 1 in the NIST scoring software) in order to be comparable to last year's results.

4.1 IHM crosstalk filtering

Table 1 shows IHM recognition results using the eval05 data for the conference room condition. For each row, we show for each meeting recording site the score using unnormalized features and normalized features as described in Section 3 (missing entries were not run for lack of time). We also show the effect of using the SDM channel as a "stand-in" for participants without a microphone. Using the SDM channel does a good job of detecting speech when there is no IHM signal containing the foreground speaker. Notice the dramatic improvement in using normalization and the SDM signal for the NIST meeting, in which there was known to be a speaker without microphone (on a speakerphone). The "Reference" row shows the results of a cheating experiment in which the reference segmentation was used. It shows the best we can expect to achieve

Table 1. IHM word error using the RT-06S system for the eval05 set on the conference room data with and without energy normalization and with and without the SDM signal.

Segmenter Method	Word Error					
	ALL	AMI	CMU	ICSI	NIST	VT
Raw	**25.6**	22.0	23.5	20.9	**37.3**	23.8
Raw + SDM	**24.7**				33.0	
Norm + SDM	**22.7**	21.9	23.1	20.6	**25.2**	22.9
Reference	**19.5**	19.2	19.9	16.8	**21.4**	20.6

Table 2. IHM word error using various segmentation systems on the conference data from the RT-05S and RT-06S evaluations.

Segmenter Method	eval05 data	eval06 data
Baseline	29.3	
RT-05S system	25.9	
RT-06S system (raw energies)	24.7	24.0
RT-06S system (normalized energies)	22.7	22.8
Reference	19.5	20.2

using an automated method. Though we are approaching this threshold, there is clearly more work that can be done.

The SDM channel was not used in the RT-06S IHM system segmenter, since NIST specified that no un-miced speakers would be present.

Table 2 summarizes the results of the IHM segmenter using various systems on both the eval05 and eval06 evaluation sets.

4.2 Acoustic modeling

To highlight the improvement in acoustic modeling, Table 3 shows the word error on the SDM and IHM conditions using acoustic models from RT-05S and RT-06S for the conference room and lecture room results. Since IHM and MDM do not use delay-summed signals, these results exclude changes in the delay-sum algorithm. The IHM results were computed using identical segmentations. Also, the language model was kept constant in these experiments. Notice incremental improvements in all conditions, due to the aforementioned improvements in MLP feature training, Gaussian training, and MLLR.

4.3 Result summary

Table 4 summarizes results on last year's and this year's evaluation sets on the conference room condition. Numbers in parentheses indicate results that were obtained using the new energy normalization technique after the evaluation ended. Relative gains of

Table 3. Word error for SDM and IHM conditions on the conference room and lecture room data using 2005 and 2006 acoustic models.

Models	RT-05S Conference		RT-05S Lecture	
	SDM	IHM	SDM	IHM
RT-05S	40.9	24.7	51.9	30.8
RT-06S	39.3	24.1	47.4	28.6

Table 4. Word error for conference room data using the 2005 and 2006 systems on the 2005 and 2006 evaluation sets. Numbers in parentheses indicate results obtained after the official evaluation had ended.

System	MDM	SDM	IHM
	eval05 data		
RT-05S system	30.2	40.9	25.9
RT-06S system	29.0	39.3	24.1 (23.0)
	eval06 data		
RT-06S system	34.2	41.2	24.1 (22.8)

Table 5. Word error for the lecture room task using the RT-05S and RT-06S systems on the eval05 and eval06 data sets.

System	IHM	SDM	MDM	ADM	UKA/MBF	ICSI/MBF
	eval05 data					
RT-05S system	28.0	51.9	52.0	44.8	-	-
RT-06S system	23.8	47.7	45.8	38.6	-	-
	eval06 data					
RT-06S system	31.0	57.3	55.5	51.0	56.5	56.0

3.9% for SDM, 4.0% for MDM, and 6.9% (11.2% post-eval) for IHM were achieved on the eval05 data. The difficulty of the eval06 set is comparable to the eval05 set, with the possible exception of the MDM condition, which is slightly worse. We have not yet analyzed this discrepancy.

Table 5 summarizes all results for the lecture room task using the RT-05S and RT-06S systems on the eval05 and eval06 data sets. Notice that eval06 was overall much more difficult than eval05, possibly because of more nonnative speakers, more variation in recording sites, and more channels in the IHM condition (causing more insertion errors from crosstalk).

For all conditions, the RT-06S lecture system shows substantial improvements compared to the RT-05S system, as measured on eval05 data. The gains were 8.1% relative for the SDM condition, 11.9% for MDM, 13.8% for ADM, and 15.0% for IHM.

Looking at lecture recognition results across distant microphone conditions, we see that the delay-sum combination method is effective. Compared to last year's system, this year's system is more robust: last year, the MDM results were worse than the SDM results, whereas the improved delay-sum algorithm now ensures that additional distant microphones always improve results (ADM is better than MDM, which is better than SDM).

For the MBF (multiple microphone beam-formed) condition, we ran the same system as for the SDM condition, using the beamformed signal as input. For the evaluation, the University of Karlsruhe provided a single beamformed signal based on all the signals from the MM3A microphones [18] (denoted "UKA/MBF" in the table). We were curious to compare the ICSI blind beamformer with the source-localization-based beamformer employed by Karlsruhe, for recognition purposes. Using the same delay-sum procedure as described in Section 3, we generated a new beamformed signal and ran an otherwise identical recognition system ("ICSI/MBF" in the table). Results show that, if anything, the delay-sum method gives somewhat better recognition results. We attribute this to the fact that our algorithm was tuned specifically for recognition accuracy, whereas Karlsruhe's was presumably optimized for source localization accuracy.

Finally, it is interesting to note that the MBF condition performed worse than MDM despite the larger number of microphones (64) in the array. One possible explanation is that the MM3A arrays were located farther from the main speaker than the MDM microphones.

5 Conclusions and Future Work

We continue to make progress in the automatic transcription of conference and lecture room meetings, as measured on NIST evaluation data. Modest gains were achieved in the conference room domain, with the largest improvement coming from the use of integrated cross-channel features in the IHM segmenter. Substantial gains were achieved in the lecture room task through the use of conference-trained distant microphone MLP features, more robust delay-sum, the use of CHIL and TED data to adapt the models, and a small LM improvement. It should be pointed out that all lecture system development occurred within a couple of weeks before and during the evaluation, and further improvements can no doubt be achieved with more careful experimentation.

Plenty of work remains. Several of the system parameters (such as LM weights and insertion penalties) were not properly optimized due to time constraints. Feature mapping techniques could reduce mismatch of the CTS and BN background training data. Given the large number of nonnative speakers of English especially in the lecture data, models adapted to particular accents may improve performance. Finally, although overlapped speech was considered part of the primary condition in this year's evaluation, we made no special effort to handle this type of speech; we consider the detection and modeling of overlapped speech one of the main challenges for future work.

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 José Pardo for contributions to the segmentation and delay-sum algorithms, the members of SRI's Speech Technology and Research Laboratory, as well as Arindam Mandal from the University of Washington for assistance with the recognition system, the U. Washington SSLI laboratory for the computer resources used for web data collection, and all the researchers at ICSI for their help and patience during the evaluation.

References

1. 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

2. 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)
3. Lamel, L., Schiel, F., Fourcin, A., Mariani, J., Tillman, H.: The translingual English database (TED). In: Proc. ICSLP, Yokohama (1994) 1795–1798
4. 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
5. Anguera, X., Wooters, C., Pardo, J.M.: Robust speaker diarization for meetings: ICSI-SRI RT-06S meetings evaluation system. In: Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006. Springer (2007)
6. Anguera, X., Wooters, C., Peskin, B., Aguiló, M.: Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 diarization 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) 402–414
7. Flanagan, J.L., Johnston, J.D., Zahn, R., Elko, G.W.: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. **78** (1985) 1508–1518
8. Boakye, K., Stolcke, A.: Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In: Proc. ICSLP, Pittsburgh, PA (2006)
9. 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
10. 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
11. 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
12. Kumar, N.: Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. PhD thesis, Johns Hopkins University, Baltimore (1997)
13. 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
14. 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
15. Jin, H., Matsoukas, S., Schwartz, R., Kubala, F.: Fast robust inverse transform SAT and multistage adaptation. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, Morgan Kaufmann (1998) 105–109
16. Lamel, L., Adda, G., Bilinski, E., Gauvain, J.L.: Transcribing lectures and seminars. In: Proc. Interspeech, Lisbon (2005)
17. Wan, V., Hain, T.: Strategies for language model web-data collection. In: Proc. ICASSP. Volume I., Toulouse (2006) 1069–1072
18. Gehrig, T., McDonough, J.: Tracking multiple simultaneous speakers with probabilistic data association filteres. In: Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006. Springer (2007)