Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System*

Andreas Stolcke[1,2], Xavier Anguera[1,3], Kofi Boakye[1], Özgür Çetin[1], Frantisek Grezl[1,4], Adam Janin[1], Arindam Mandal[5], Barbara Peskin[1], Chuck Wooters[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
[4] Brno University of Technology, Czech Republic
[5] University of Washington, Seattle, WA, U.S.A.
stolcke@icsi.berkeley.edu

Abstract. We describe the development of our speech recognition system for the NIST Spring 2005 Meeting Rich Transcription (RT-05S) evaluation, highlighting improvements made since last year [1]. The system is based on the SRI-ICSI-UW RT-04F conversational telephone speech (CTS) recognition system, with meeting-adapted models and various audio preprocessing steps. This year's system features better delay-sum processing of distant microphone channels and energy-based crosstalk suppression for close-talking microphones. Acoustic modeling is improved by virtue of various enhancements to the background (CTS) models, including added training data, decision-tree based state tying, and the inclusion of discriminatively trained phone posterior features estimated by multilayer perceptrons. In particular, we make use of adaptation of both acoustic models and MLP features to the meeting domain. For distant microphone recognition we obtained considerable gains by combining and cross-adapting narrowband (telephone) acoustic models with broadband (broadcast news) models. Language models (LMs) were improved with the inclusion of new meeting and web data. In spite of a lack of training data, we created effective LMs for the CHIL lecture domain. Results are reported on RT-04S and RT-05S meeting data. Measured on RT-04S conference data, we achieved an overall improvement of 17% relative in both MDM and IHM conditions compared to last year's evaluation system. Results on lecture data are comparable to the best reported results for that task.

1 Introduction

Meeting recognition continues to be a challenging task for speech technology for several reasons. Unrestricted speech, recognition from distant microphones, varying noise conditions, and multiple and overlapping speakers pose problems not found in other

* This paper contains corrections and additions to the version distributed at the NIST MLMI Meeting Recognition Workshop.

widely used benchmark tests. Furthermore, meetings pose the interesting problem of designing *portable* recognition systems, in two regards. First, because of the relative novelty of the task, and limited size of in-domain training corpora, it is advantageous to try to leverage methods and data that have been developed for other genres of speech, such as conversational telephone speech (CTS) and broadcast news (BN), for which one can draw on a longer development history and an order of magnitude more data. The second motivation for portability is that the meeting domain itself is varied, with different collection sites, acoustic conditions, and conversational styles and topics.

As for last year's meeting evaluation (RT-04S), our development strategy for RT-05S was to start with an existing CTS systems[6] and adapt it to the meeting domain. This allowed us to leverage research between the corresponding CTS evaluations, from the Fall of 2003 (RT-03F) to the Fall of 2004 (RT-04F), and was crucial to developing a meeting system in the short period available. Acoustic models were adapted to the available conference room data (some of it new for this year), and language models were rebuilt for the conference and lecture room domains (no special acoustic models were created for the lecture domain). A new aspect in our acoustic modeling this year was the use of discriminatively trained Tandem/HATS features, and the fact that features were adapted to the new task, in addition to the more standard model adaptation. The acoustic preprocessing for meetings was also improved significantly, for both distant and individual microphone conditions.

The evaluation task and data are described in Section 2. Section 3 gives the system description, focusing on new developments relative to the 2004 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-05S conference room evaluation data (`eval05`) consisted of two meetings from each of the recording sites AMI (Augmented Multi-party Interaction project), CMU, ICSI, NIST, and VT (Virginia Tech). Systems were required to recognize a specific 12-minute segment from each meeting; however, data from the entire meeting was allowed for processing.[7] 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 provided the following recording conditions:

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

[7] We did find significant gains from adapting on entire meetings, and, except in the acoustic preprocessing, used only the designated meeting excerpts.

MSLA Multiple source-localization arrays (optional)

MM3A Multiple Mark III microphone arrays (optional). The MM3A condition has not yet been delivered for the evaluation set, and could be evaluated only on development data, using a single array.

It should be noted that microphones varied substantially even within each condition. For example, some of the AMI IHM data was 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).

Development data The RT-04S evaluation data, consisting of eight 11-minute excerpts of meetings from CMU, ICSI, LDC, and NIST was designated as development data for RT-05S, and used by us as an unbiased test set (`eval04`). For most of the development we relied on the RT-04S development set, consisting of another 8 meetings from the same sources, and a newly provided set of 10 AMI meetings. Out of these we formed a balanced 10-meeting set (designated `dev04a`) that served for optimization and system tuning. An additional 5 meetings (2 ICSI, 2 CMU, 1 LDC) were available from the RT-02 devtest set (used by us only for some LM tuning and speech/nonspeech model training). Note that the `eval05` "VT" meetings had no corresponding development data and thus served as a blind test. Excerpts from 5 CHIL lectures were available for development testing in the lecture room domain.

2.2 Training data

Training data was available 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 was of limited use in that only lapel and no distant microphone recordings were available.

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 first individually noise-filtered using a Wiener filter with typical engineering modifications, identically to last year [2, 1].

Subsequently, for the MDM and MSLA conditions, a delay-and-sum beamforming technique was applied to combine all available distant microphone channels into a single enhanced signal, described in more detail in [3]. A time delay of arrival (TDOA) was computed between each input channel and a reference channel every 250 ms, using the GCC-PHAT algorithm [4] on 500 ms segments. The reference channel was chosen

as the most centrally located microphone in the room, as specified by the SDM condition. For each step of 250 ms, a 500 ms segment was extracted for each channel and delayed according to the computed TDOA. Finally, the different channels were summed together, multiplied by a triangular window to avoid discontinuities between steps.

Speech regions were then 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], except that special speech/nonspeech models were trained exclusively from and for AMI meetings.

Finally, the segments were clustered into acoustically homogeneous partitions, which serve as pseudo-speaker units for normalization and adaptation. Last year we fixed the number of clusters at 4; this year the cluster number was chosen automatically, but such that each cluster contained at least 20 segments. We tried using the output of the ICSI-SRI diarization system [3] for segment clustering, so far without improvement in recognition accuracy. This could be because even within one speaker there is important acoustic variation (e.g., due to head movement) that is detected by the current clustering algorithm.

Close-talking microphone processing The IHM input channels were segmented (without Wiener filtering) into speech and nonspeech regions using the same basic algorithm as for the distant microphone signals, using speech/nonspeech models trained on the close-talking training data (again, except for separate AMI processing, models are unchanged from 2004). No speaker clustering was performed, since it was assumed that each IHM channel corresponds to exactly one speaker. However, this year we added a crosstalk detector, with the goal of avoiding insertion of recognized speech from background speakers.[8]

The system generates start and end times for *foreground* speech segments by performing zero-level thresholding of a "crosstalk-compensated" signal derived from channel energies. For each target channel $i = 1, 2, \ldots, N$ in the set of IHM channels, the crosstalk-compensated signal $S_{CC,i}$ is given by

$$S_{CC,i}(n) = E_{\mathrm{offset},i}(n) - \frac{1}{N-1} \sum_{k \neq i} E_{\mathrm{offset},k}(n) \quad . \tag{1}$$

Here $E_{\mathrm{offset},k}$ is computed as $E_k(n) - \min_l E_k(l)$, that is, the signal energy minus the minimum signal energy over the channel. This minimum energy is used as an estimate of the noise floor.

The subtraction of the average of the nontarget energy signals is done with the expectation that regions in which crosstalk appears on the target channel (and most likely on other channels) will have values below the threshold in the resulting signal, as the crosstalk will appear as a region with significant energy in the averaged signal. The energy signals represent an average over a window of 25 ms with a step size of 10 ms. The presumed foreground segments thus detected are then intersected with the output of the speech/nonspeech decoder.

[8] The post-recognition crosstalk detector used last year had proven ineffective and was discarded.

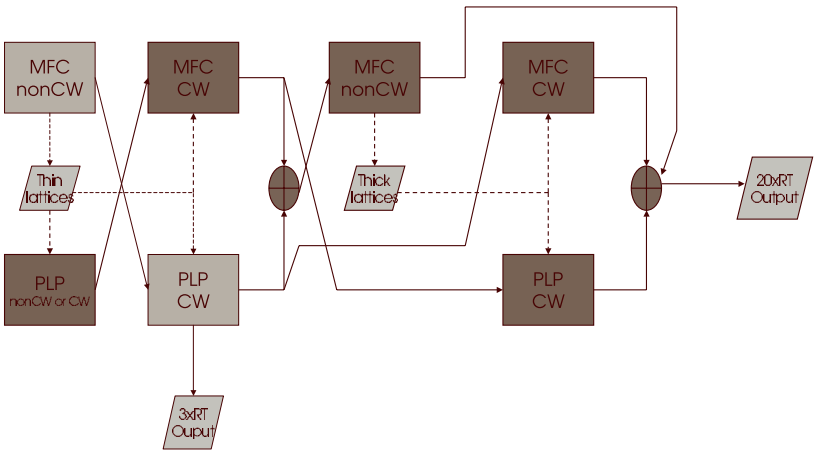

Fig. 1. SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote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

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 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 adapted to the output of a previous step from the respective other tier using MLLR (cross-adaptation). 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 non-crossword (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 LM; N-best output is also rescored with duration models for words and pauses [5].

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).[9] For quick turnaround development it is useful to use a "fast" subset of the full system consisting of just two decoding steps (the light-shaded boxes in the figure); this fast system runs in 3xRT and exercises all the key elements of the full system except for confusion network combination.

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

[9] Runtimes given assume operation with Gaussian shortlists. Since RT-05S did not impose a runtime limit we ran the system without shortlists, in about 25xRT.

with the minimum phone error (MPE) criterion [6] 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 [7]. 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) [8]. After HLDA, a 25-dimensional Tandem/HATs feature estimated by multilayer perceptrons (MLPs) [9, 10] was appended. Both within-word and cross-word triphone models were trained, for lattice generation and decoding from lattices, respectively. Baseline PLP CTS models (cross-word triphone only) were trained in analogous fashion, but did not include voicing or MLP features. All models this year were trained using decision-tree-based state tying, rather than SRI's traditional bottom-up genonic model clustering; this change had resulted in improved CTS performance.

In testing, all models underwent unsupervised adaptation to the test speaker or cluster, using maximum likelihood linear regression (MLLR) with multiple, phonetically defined regression classes. After the evaluation we experimented with regression classes that were generated in a data-driven manner by decision 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) [11].

Following work by the CMU-ISL team in RT-04S [12], we also experimented with PLP baseline models trained on BN data. Unlike the CTS versions, these models use the full signal bandwidth and are gender independent. Otherwise, the BN model used similar training, normalization, and adaptation techniques: VTLN, HLDA, feature-space CMLLR, and model-space MLLR.

Acoustic model task adaptation All baseline models were adapted to the IHM and distant microphone conditions using the respective channels in the training data. Based on experiments with last year's system, we chose not to use the CMU data for model adaptation. Also, we found (in 2004) that there was no advantage to delay-summing the training data for MDM recognition, compared to pooling all the individual distant microphone signals into one training set. This meant that, conveniently, a single adapted model set could be used for all distant microphone test conditions. Last year we found only a very minor benefit from adapting acoustic models to individual meeting sources; this year the same pooled adaptation data was used for all meetings. The weight for adaptation data statistics was empirically optimized, and set at 20.

Last year's baseline models had been trained with maximum mutual information (MMI) estimation, and accordingly a version of MAP adaptation that adapted numerator and denominator statistics separately was employed [1]. This year's baseline models were created using MPE training [6], and we found it best to apply the MMI-MAP procedure in adaptation [13]. However, due to lack of time, we applied MMI-MAP only to the IHM models, and used the standard, less-involved ML-MAP procedure on the distant microphone models.

MLP feature adaptation The MLPs estimating Tandem/HATS features had been trained on a large subset of the CTS training data [10] to perform frame-level phone discrimination. In addition to MAP-adapting the models based on these features, we explored adapting the features themselves to better match the meeting domain. This was accomplished by applying three additional backpropagation iterations to the CTS-trained MLPs, using the meeting data as training material. The KLT transform mapping the phone log posteriors was kept unchanged from the CTS system, so as to keep the features compatible with existing models. Because of time constraints and data availability, we were able to carry out this procedure only once, using CMU, ICSI, and NIST close-talking microphone data. However, as described in the next section, the adapted features gave improved results on all meeting sources, and for both IHM and distant microphone conditions; we therefore used the same adapted MLPs in all versions of our system.

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 had 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 collected to match different topics and styles [14], namely RT-04S meeting sources, AMI meetings, and 525M words of Fisher-like conversational web data collected and published by UW for the RT-04F evaluation. We obtained best recognition results with mixture weights that had been tuned to minimize perplexity on heldout CMU, ICSI, LDC, and NIST (but not AMI) transcripts. The LM vocabulary consisted of 54,524 words, comprising all words in our CTS system (including all Hub-5 and all non-singleton Fisher words), all words in the ICSI, CMU, and NIST training transcripts, and all non-singleton 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) 0.1M words of TED oral transcripts and (g) 28M words of speech conference proceedings (suggested by [15]). Also, the Fisher-relevant web data was replaced by web data related to conference proceedings. The lecture LM mixture was then optimized on CHIL development transcripts (LM tuning and testing used a jackknifing scheme to avoid tuning on the data being tested on). No CHIL transcripts were used for N-gram training. 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%.

Table 1. Comparison of SDM and new and old MDM delay-sum methods, using RT-04S models and fast decoding system. WER on `eval04` with and without CMU meetings (which had only one distant microphone channel) are given.

Method			`eval04`	
Input	Delay-sum	Segmentation on	w/CMU	w/o CMU
SDM			51.3	48.9
MDM	old	SDM	47.4	42.9
MDM	new	delay-summed	45.5	40.1
MDM	new	SDM	45.5	40.3

Table 2. IHM performance without and with crosstalk filtering, and with reference segmentation. Results on `eval04` obtained with RT-04S models, fast system; `eval05` with RT-05S models, full system. WERs are broken down by meeting source, and for `eval05` by error type (substitutions, deletions, insertions).

Crosstalk	`eval04`					`eval05`								
filter	**All**	CMU	ICSI	LDC	NIST	**All**	AMI	CMU	ICSI	NIST	VT	Sub	Del	Ins
No	**35.4**	39.7	27.4	44.7	27.1	**29.3**	22.1	23.3	20.5	45.8	23.3	11.0	10.3	8.0
Yes	**34.3**	39.3	25.2	43.0	27.3	**25.9**	23.3	23.3	24.5	34.5	23.3	11.0	11.5	3.4
Reference	**32.1**	36.9	23.9	40.3	24.3	**19.5**	19.2	19.9	16.8	21.4	20.6	11.2	6.7	1.6

4 Results and Discussion

Here we present results measuring the effects of the system features and improvements described in the previous section. We will first present mostly conference meeting results, since those were the focus of our development. Lecture recognition results are summarized at the end.

4.1 MDM delay-sum processing

For RT-04S, we applied the delay-sum beamforming *after* the segmentation step, with one TDOA estimate per waveform segment. This year, delay-summing was performed *before* segmentation, as described above. Table 1 compares SDM and MDM results with both methods. We observe that the new delay-sum method reduces word error rate (WER) for meetings with multiple distant microphones by 6.5% relative over the old method, and by 18.0% relative over the single-microphone condition. We also tested the new delay-sum algorithm, but retaining the old segmentation computed from SDM input, and found almost the same improvement as with segmentation based on the delay-summed signal (last row in Table 1. This shows that the improvement stems mostly from better recognition on the enhanced signal.

4.2 IHM crosstalk filtering

Table 2 shows IHM recognition results, without and with the new crosstalk suppression algorithm, as well as for an ideal segmentation derived from the NIST STM reference files. On both test sets, our crosstalk processing eliminates about one third of the word error difference between automatic and reference segmentation. However, broken down by meeting source the error patterns on the two test sets differ somewhat. On `eval04`

Table 3. IHM word error rates (in **bold**) and perplexities (in *italics*) with various LM mixtures on development data.

	eval04					2005 devtest	
Language model	All	CMU	ICSI	LDC	NIST	AMI	CHIL
RT-04F (CTS)	**28.7** *95*	**32.1** *111*	**24.5** *78*	**34.1** *85*	**21.5** *109*	**39.8** *113*	**37.6** *320*
RT-04S	**28.7** *97*	**33.1** *117*	**22.0** *62*	**35.7** *97*	**21.1** *99*	**38.3** *107*	**31.5** *212*
RT-05S, no web	**28.9** *103*	**33.4** *121*	**22.0** *66*	**36.0** *101*	**21.5** *110*	**38.4** *100*	**27.6** *155*
RT-05S, w/web	**27.9** *92*	**32.5** *111*	**21.4** *59*	**34.9** *93*	**20.2** *90*	**37.3** *94*	**26.9** *148*

(as well as on our development data) the crosstalk filter never increased WER significantly.[10] On `eval05`, however, WER sometime increases, because of occasional deletions of foreground speech by the filter. This is especially a problem on the `eval05` ICSI meetings, for reasons yet to be investigated. It should be noted that one of the `eval05` NIST meetings is anomalous, in that a talker participates via a speakerphone, but without an associated IHM channel. Also, three channels do not contain any speech. Both these factors lead to very high insertion rates that our algorithm cannot yet effectively suppress.

4.3 Language modeling

To evaluate the effectiveness of the various LM mixture components, we ran IHM recognition tests on `eval04`, AMI devtest data, and CHIL devtest data. Results are summarized in Table 3. We note that the addition of AMI meeting transcripts, additional Fisher data, and new web data reduced WER by about 1.2% absolute on conference meetings relative to last year's meeting LM. Naturally, given the difference in topic and speaking styles, the adaptation to the lecture domain has a more dramatic effect, as the WER is reduced by 4.6%. Web data is quite important for conference meetings, lowering WER by 1.3%-1.5%, but less so for lectures, where its effect is only a 0.7% absolute reduction. A possible explanation for this difference is that lecture-relevant material on the web is already available in the conference proceedings used in lecture LM training. Furthermore, we observe that the CTS LM mixture performs the best on the CMU and LDC meetings in terms of WER, and on the LDC meetings in terms of perplexity. This could be due to the sparsity of training data for these two sources, or to intra- and inter-source variability.

4.4 Acoustic modeling

We tested a range of acoustic models to determine the contribution of baseline model improvements, Gaussian adaptation, MLP features (original and adapted), and CTS/BN model combination (for distant microphone recognition). Results are summarized in Table 4. Below we point out the most important contrasts.

[10] Here and elsewhere, we score on all `eval04` personal microphone channels, including one ICSI lapel microphone channel that was removed from official scoring.

Table 4. Word error rates using various sets of acoustic models and evaluation data for the conference test conditions. All results were obtained using the full recognition system and conference meeting LMs. Columns 2 and 3 indicate whether the Gaussian models and/or the MLP features were adapted to the meeting domain. "None" in column 3 indicates that MLP features were used at all, whereas "no" means that CTS-trained MLPs were used. **Highlighted** results correspond to the final evaluation system.[12]

	Baseline	Gaussians	MLP	MDM		IHM	
Line	models	adapted	adapted	eval04	eval05	eval04	eval05
a	RT-03F CTS	no	no	48.3	40.2	33.2	30.8
b	RT-04F CTS	no	no	41.4	34.5	28.9	28.6
c	RT-04F CTS	no	yes	41.1	34.2	28.4	27.0
d	RT-04F CTS	ML-MAP	none	n/a	n/a	29.4	28.6
e	RT-04F CTS	ML-MAP	no	n/a	n/a	28.6	26.9
f	RT-04F CTS	ML-MAP	yes	40.3	32.2	28.3	26.2
g	RT-04F CTS+BN	ML-MAP	yes	**37.1**	**30.2**	28.0	26.3
h	RT-04F CTS	MMI-MAP	yes	n/a	n/a	**27.9**	**25.9**
i	RT-04F CTS	MMI-MAP+DT	yes	n/a	n/a	n/a	25.6

Lines (a) and (b) give results with CTS models underlying the RT-04S and RT-05S meeting models, respectively. We observe between 7% and 16% relative WER reduction as a result of added CTS data and improved modeling techniques. Gaussian adaptation (e) and feature adaptation (c) each give about the same amount of improvement for IHM. Feature adaptation gives only a small gain for MDM because the MLPs were adapted on close-talking data only. Line (f) shows that Gaussian and feature adaptation are partly additive. The combined WER reduction is about 7-8% relative on eval05 and 2-3% on eval04. MMI-MAP (h) gives an extra 1% relative error reduction (for IHM), as does the use of decision-tree-generated regression classes (post-evaluation, line i).

The combination of CTS-based MFCC models with BN-based PLP models (g) results in a large, 6-7% relative error reduction for MDM. Preliminary experiments had shown no gain for IHM, and the post-evaluation results given here show that there is no consistent gain over CTS-based PLP models. The reason might be that while CTS models are a better match to meeting speech in terms of speaking style, BN data contains more samples of distant microphones and noisy speech.

A comparison of adapted models without MLP features (d) and with adapted MLP features (f) shows an improvement of 9.4% relative on eval05. That is very comparable to the 10% relative gain found in the CTS domain [10], and indicates good portability of the Tandem/HATS method.

Comparing the two evaluation sets, we notice that eval05 is slightly easier (2% absolute) for IHM, and considerably easier (almost 10% absolute) for MDM. The absolute WER differences between line (a) and lines (g)/(h) are almost the same for the two test sets (about 10% for MDM and 5% for IHM). However, almost all the win on eval04 seems to come from improvements in the baseline system, whereas for

[12] The original MDM submission had 30.0% WER, due to a VTLN bug that actually helped on eval05. Post-evaluation we found the BN model had not been adapted to female speakers; however, fixing this only reduced the WER to 30.1%.

Table 5. Evaluation system result summary.

	Conference Meetings						Lecture Meetings							
	eval04			eval05			CHIL devtest				eval05			
System	MDM	SDM	IHM	MDM	SDM	IHM	MDM	MSLA	MM3A	IHM	MDM	MSLA	SDM	IHM
RT-04S	44.9	51.3	33.6											
RT-05S	37.1	43.0	27.9	30.2	40.9	25.9	51.6	51.0	49.7	26.9	52.0	44.8	51.9	28.0

eval05 the adaptation techniques contribute a larger gain. The reasons for this discrepancy still remain to be investigated.

4.5 Result summary

Table 5 summarizes results on last year's and this year's evaluation sets, including on lecture room data. The lecture recognition system differed from the conference meeting system only in the LM (as described earlier), and in the configuration of the signal preprocessing. For MDM processing, we found that one of the CHIL tabletop microphones had much better signal-to-noise ratio than the others, and was best used alone, instead of in beamforming. Also, the speaker clustering for distant microphones proved detrimental and was omitted, no doubt because the lectures are dominated by a single speaker. For comparison with other lecture recognition work we include results on the development data, which corresponds to the CHIL project's January 2005 evaluation set.

Based on eval04 results, the overall reduction in word error compared to last year's system is 17.4% relative for MDM, and 16.9% for the IHM condition. Error rates are broadly comparable on eval04 and eval05, in spite of the latter containing different meeting sources, including one "blind" data source (VT).

Word error rates on CHIL seminar lectures are comparable to conference meetings for the IHM condition. Distant microphone recognition shows 10% or more absolute higher WERs, which is not unexpected given the challenging acoustic conditions and the lack of in-domain training data. Results are in line with error rates reported by CHIL research sites [15].

5 Conclusions and Future Work

We have made considerable progress in the automatic transcription of conference meetings, as measured on NIST evaluation data. Substantial improvements came from meeting-specific preprocessing methods, as well as successful porting of CTS and BN models, MLP features, and decoding techniques, for an overall word error reduction of about 17% relative. We were also pleased that the system generalized well to previously unseen meeting sources and to the lecture domain, the latter with only minimal language model porting effort.

Major challenges remain, for example, in the recognition of distant speakers and overlapping speech. The single most important problem in IHM recognition remains the separation of foreground from background speech, especially when not all meeting participants are recorded individually. We also hope to benefit from tighter integration with our diarization system in the future.

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. We thank Ivan Bulyko for assistance with LM preparation and web data collection; Lori Lamel for information concerning CHIL data and scripts for proceedings processing; Thomas Hain for information about AMI data; Barry Chen, Qifeng Zhu and Nelson Morgan for assistance and advice with MLP feature computation; and members of SRI's Speech Technology and Research Laboratory for assistance with the recognition 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. 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
3. Anguera, X., Wooters, C., Peskin, B., Aguiló, M.: Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 diarization system. In: Proceedings NIST MLMI Meeting Recognition Workshop, Edinburgh (2005)
4. 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
5. 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
6. 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
7. 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
8. Kumar, N.: Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. PhD thesis, John Hopkins University, Baltimore (1997)
9. 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
10. Zhu, Q., Stolcke, A., Chen, B.Y., Morgan, N.: Using MLP features in SRI's conversational speech recognition system. In: Proc. Interspeech, Lisbon (2005)
11. 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
12. Metze, F., Fügen, C., Pan, Y., Waibel, A.: Automatically transcribing meetings using distant microphones. In: Proc. ICASSP. Volume 1., Philadelphia (2005) 989–902
13. Povey, D., Gales, M.J.F., Kim, D.Y., Woodland, P.C.: MMI-MAP and MPE-MAP for acoustic model adaptation. In: Proc. EUROSPEECH, Geneva (2003) 1981–1984
14. Bulyko, I., Ostendorf, M., Stolcke, A.: Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In Hearst, M., Ostendorf, M., eds.: Proc. HLT-NAACL. Volume 2., Edmonton, Alberta, Canada, Association for Computational Linguistics (2003) 7–9
15. Lamel, L., Adda, G., Bilinski, E., Gauvain, J.L.: Transcribing lectures and seminars. In: Proc. Interspeech, Lisbon (2005)