From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System

Nikki Mirghafori[1], *Andreas Stolcke*[1,2], *Chuck Wooters*[1], *Tuomo Pirinen*[1,5], *Ivan Bulyko*[3], *Dave Gelbart*[1,4], *Martin Graciarena*[2], *Scott Otterson*[3], *Barbara Peskin*[1], *Mari Ostendorf*[3]

[1] International Computer Science Institute [2] SRI International [3] University of Washington
[4] University of California at Berkeley [5] Tampere University of Technology
{nikki,stolcke,wooters}@icsi.berkeley.edu

Abstract

We describe the ICSI-SRI-UW team's entry in the Spring 2004 NIST Meeting Recognition Evaluation. The system was derived from SRI's 5xRT Conversational Telephone Speech (CTS) recognizer by adapting CTS acoustic and language models to the Meeting domain, adding noise reduction and delay-sum array processing for far-field recognition, and postprocessing for cross-talk suppression. A modified MAP adaptation procedure was developed to make best use of discriminatively trained (MMIE) prior models. These meeting-specific changes yielded an overall 9% and 22% relative improvement as compared to the original CTS system, and 16% and 29% relative improvement as compared to our 2002 Meeting Evaluation system, for the *individual-headset* and *multiple-distant* microphones conditions, respectively.

1. Introduction

Processing natural multi-party interactions presents a number of new and important challenges to the speech community, from dealing with highly interactive and often overlapping speech to providing robustness to distant microphones recording multiple talkers. Data collected from meeting rooms provide an ideal testbed for such work, supporting research in robust speech recognition, speaker segmentation and tracking, discourse modeling, spoken language understanding, and more.

Recent years have seen increased research activity on meeting data at such sites as CMU/Karlsruhe [13] and ICSI [8], as well as a number of European initiatives. In March 2004 NIST conducted an evaluation of speech recognition systems for meetings (RT-04S), following on its initial Meetings evaluation two years prior (RT-02) [2]. Our team had participated in RT-02 with an only slightly modified CTS recognition system, providing little more than a baseline for future work. For RT-04 our goal was to assemble a system specifically for meeting recognition, although the limited amounts of meeting-specific training data dictated that such a system would still be substantially based on our CTS system. This paper describes and evaluates the design decisions made in the process.

The evaluation task and data are described in Section 2. Section 3 includes the system description, followed by results and discussion in Section 4. Conclusions and future work are presented in Section 5.

2. Task and Data

Test Data. The RT-04S evaluation data consisted of two meetings from each of the recording sites CMU, ICSI, LDC, and NIST, each about one hour or more in length. Systems were required to recognize a specific 11-minute segment from each meeting; however, data from the entire meeting was allowed for purposes of adaptation, etc.[1] Separate evaluations were conducted in three conditions:

MDM Multiple distant microphones (primary)
IHM Individual headset microphones (required contrast)
SDM Single distant microphone (optional)

The CMU meetings came with only one distant mic; for the other meetings between 4 and 10 distant mics were available. The IHM systems were allowed to use all mics (distant or individual). For MDM and SDM conditions, NIST only evaluated regions of speech with a single talker, thus eliminating overlapping speech. Unlike recent CTS evaluations, the Meetings evaluation included non-native speakers of English.

The RT-02 evaluation data (another 8 meetings from the same sources) served as the development test set for RT-04. However, this set was somewhat mismatched to the RT-04 evaluation data in that CMU and LDC used lapel[2] instead of head-mounted microphones. An additional 5 meetings (2 ICSI, 2 CMU, 1 LDC) were available from the RT-02 devtest set.

Training Data. Training data was available from CMU (17 meetings, 11 hours of speech after segmentation), ICSI (73 meetings, 74 hours), and NIST (15 meetings, 14 hours). No data from LDC was available. The CMU data was problematic in that only lapel and no distant microphone recordings were available.

We excluded any data which failed to force-align with the released transcriptions. This eliminated 0.1% of the data from each of ICSI and NIST, and 11% from CMU. For acoustic training of the distant mic systems, we also excluded regions with overlapped speech, based on forced alignments of the individual mic signals.

3. System Description

Our meeting recognition system was based on a fast (5 times real-time) version of SRI's CTS recognizer, which we augmented and adapted for the meeting task. Key aspects of the system are described in the following sections.

3.1. Signal Processing and Segmentation

Noise Reduction of the Far-Field Microphone Signals. The distant mic signals are filtered using a batch version of the noise

[1] Since preliminary experiments had shown only minor benefits from normalizing and adapting on entire meetings, this option was not pursued.

[2] Throughout the text, *individual mic* subsumes both individual lapel and individual head-set mic conditions.

reduction algorithm developed for the Aurora 2 front-end proposed by ICSI, OGI, and Qualcomm [3]. The algorithm performs Wiener filtering with typical engineering modifications, such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor. We modified the algorithm to use a single noise spectral estimate for each meeting waveform. This was calculated over all the frames judged to be nonspeech by the voice-activity detection component of the Qualcomm-ICSI-OGI front end. We applied it independently for each meeting waveform and used overlap-add resynthesis to create noise-reduced output waveforms, which then served as the basis of all further processing.

Segmentation. To identify regions of speech activity and segment them into suitable chunks for further processing, a recognizer with two phones (speech and nonspeech) was used to decode the signal. The phone models impose minimum duration constraints and the language model (LM) penalizes switches between the two models. The resulting segments were postprocessed to satisfy length constraints, and to pad speech boundaries with a few frames of nonspeech. For distant mics, the algorithm performs acoustic clustering to keep different speakers in separate segments, and to group same or similar speakers into clusters that can subsequently be used for feature normalization and acoustic adaptation.

For the headset mics condition, the segmentation models were trained on ICSI and NIST headset mics training data, using forced alignments against the references. For the distant mic conditions, two sets of models were trained: ICSI and NIST data were used to train models for those two sources; the RT-02 devtest data (which included some CMU and LDC far-field data) were used to train models for segmenting the CMU and LDC meetings.

Multiple Distant Microphone Array Processing. For MDM processing, segmentation was performed on a single, central mic. Array processing was then performed separately on each speech region of the noise-reduced signals according to the common segmentation. The waveform segments from the various distant microphones were aligned to compensate for time skew and sound travel delays. Finally the aligned signals were summed to yield a single new segmented waveform.

The rationale behind this processing is that speech will be summed in-phase and amplified, whereas noise components are summed out of phase and will be dampened. Delays for time alignment were estimated using maximal cross-correlation, in which the central mic channel was used as the reference. Since the microphone and speaker locations were unknown, the same search interval was used for all microphone pairs at a given site; an educated guess as to the possible delay ranges was made based on available documentation of the recording room configurations. Note that the method assumes that each waveform segment contains only one speaker and thus that the alignment delays would not vary within a segment (hence the segmentation step had to precede the array processing).

3.2. Acoustic Modeling and Adaptation

Gender-dependent recognition models were derived from CTS models trained on 420 hours of telephone speech from the Switchboard and CallHome English collections. The MFCC models used 12 cepstral coefficients, energy, 1st, 2nd and 3rd order difference features, as well as 2x5 voicing features over a 5-frame window [5]. The 62-component raw feature vector was reduced to 39 dimensions using heteroscedastic linear discriminant analysis [7]. PLP models used a similar configuration, except that no voicing features were included and a two-stage transform, consisting of standard LDA followed by a diagonalizing transform [11] were used to map the feature space from 52 to 39 dimensions. Also, the PLP models were trained with feature-space speaker adaptive training [6].

The CTS models were adapted to the meeting domain using ICSI and NIST training data (the CMU meetings were deemed to be mismatched to the eval data, as discussed in Section 2). Since the prior models had been trained with the maximum mutual information criterion (MMIE) [10] we developed a version of the standard maximum a-posteriori (MAP) adaptation algorithm that preserves the models' discriminative properties. CTS MMIE models were used to collect numerator and denominator counts on the meeting data (downsampled to 8kHz). These counts were combined with CTS numerator and denominator counts, respectively. Finally, new Gaussian parameters were estimated from the combined counts (mixture weights and HMM parameters were left unchanged in the process).

Experiments showed that an adaptation weight near 20 for the numerator and 5 for the denominator was optimal. Furthermore, as reported in Section 4, most of the improvement can be achieved by only adapting the numerator counts; this could be convenient for some applications since denominator training requires lattices to be generated for the adaptation data.

Feature Mapping. We also experimented with the probabilistic optimum filtering (POF) [9] approach to cope with the mismatch between far-field signals and our CTS-based recognition models. In this approach a probabilistic mapping of noisy (distant mic) to clean (headset mic) features is trained based on stereo recordings. However, the method is complicated by time skew between channels, changing speakers, and location-specific background noise. We obtained an error reduction with a feature mapping trained on test data, but were not able to obtain an improvement when using only training data, and therefore did not include this method in our eventual system.

3.3. Language Model and Vocabulary

Our CTS language model is a mixture LM trained on 4M words of Switchboard transcripts, 150M words of Broadcast News, and 191M words of web data chosen for style and content [4]. It was adapted for meeting recognition by adding two meeting-specific mixture components: Meetings transcripts from ICSI, CMU, and NIST (1.7M words), and newly collected web data (150M words) related to the topics discussed in the meetings and also aimed at covering new vocabulary items. Also, 5.3M words from the CTS Fisher collection were added for coverage of current topics. The mixture was adapted by minimizing perplexity on a held-out set consisting of approximately equal amounts of transcripts from the four sources. We also experimented with source-specific LMs, but found that the available tuning data was insufficient to estimate source-specific mixture weights robustly.

The vocabulary was extended (relative to the baseline CTS system) to include all non-singleton words from Fisher and Meetings transcripts. The vocabulary size was close to 50,000, and yielded a 0.9% out-of-vocabulary rate on the development test transcripts.

3.4. Decoding

The recognition search was structured as in the SRI "fast" (5xRT) CTS system. Within-word MFCC models were adapted with phone-loop MLLR and used to generate bigram lattices. The lattices were then rescored with a 4-gram LM and

Table 1: Improvement of the new baseline CTS system as compared to the system used in the RT-02 evaluation, reported on RT-02 eval set.

	All	ICSI	CMU	LDC	NIST
Individual Mics					
RT-02 System	36.0	25.9	47.9	36.8	35.2
RT-04 Baseline	32.8	24.0	44.3	33.2	31.5
Single Distant Mic					
RT-02 System	61.6	53.6	64.5	69.7	61.6
RT-04 Baseline	56.6	48.8	61.9	60.5	60.3

consensus-decoded to obtain preliminary hypotheses. These were then used to estimate speaker-adaptive feature transforms and MLLR model transforms for the cross-word PLP models, which were employed to generate 2000-best lists from trigram-expanded lattices. The N-best lists were then rescored with a 4-gram LM, pronunciation, pause, and duration models [12], and combined into final confusion networks, from which 1-best hypotheses and confidence values were extracted.

3.5. Cross-Talk Suppression

The decoded word hypotheses from the IHM system were post-processed in an attempt to eliminate cross-talk. We assumed that when cross-talk was sufficiently loud, recognized words with low confidence would be produced, and that most speech was not overlapped. Therefore, we time-aligned the words on all channels, and deleted those words which had confidence score below a given threshold, and overlapped, by at least 50%, with a word on another channel.

4. Results and Discussion

Since both the old RT-02 system and this year's baseline system were developed for the CTS domain, we were interested to see how much of the improvements made on the CTS recognition task would carry over to the Meeting task. Using RT-02 system components comparable to the current 5xRT system, the WER on the 2002 CTS task reduced from 29.4% to 23.6%, a 20% relative reduction. As shown in Table 1, the same system achieved relative improvements of 8% and 9% on the RT-02 meeting evaluation data, in the individual and distant mic conditions, respectively.

In the rest of this section, we report results on the official RT-04S development test, whose references differed somewhat from the RT-02 evaluation set. We present experiments in cumulative fashion, so that each improvement is the baseline for the following experiment. To be consistent with RT-02, unless otherwise noted, individual mic recognition uses reference segmentations, while distant mic experiments use automatic segmentation, plus noise filtering.

First we examine the effect of LM adaptation (see Section 3.3), shown in Table 2. The improvement is roughly 5% overall and appears to be more substantial for ICSI and NIST, and less so for CMU and LDC data. Besides the lack of training data for LDC meetings, the observed difference could be due to the consistency of meeting topics in the ICSI and NIST data, and their relative variability in the CMU meetings.

Next we tested the MMIE-MAP acoustic adaptation approach described in Section 3.2. Table 3 shows small, yet consistent, improvements over the standard MLE-MAP approach. MMIE adaptation was effective even if only the numerator counts were updated ("NUM-MAP").

For the IHM condition, models were adapted on training data recorded with head-mounted microphones; for the MDM and SDM conditions, training data recorded with distant microphones were used. For the latter conditions, experiments showed that adapting models to duplicate versions of the data from different microphones decreased the WER by 35-63% more than when models were adapted to data from the central microphone only.

Table 2: Effect of language model adaptation on RT-04 devtest data.

	All	ICSI	CMU	LDC	NIST
Individual Mics					
Baseline	33.3	23.5	44.6	34.2	32.0
Adapted LM	31.5	20.9	43.6	33.7	28.5
Single Distant Mic					
Baseline	56.2	45.9	61.0	63.7	59.9
Adapted LM	53.6	43.0	60.8	62.9	52.3

Table 3: Effect of different acoustic adaptation algorithms on the IHM condition (RT-04 dev). The source of the adaptation data is matched to the test data (except for LDC, where ICSI data was used in adaptation).

	All	ICSI	CMU	LDC	NIST
Unadapted	31.5	20.9	43.6	33.7	28.5
MLE-MAP	30.4	18.4	42.8	33.2	28.0
NUM-MAP	30.0	18.3	42.0	33.0	**27.3**
MMIE-MAP	**29.8**	**17.9**	**41.4**	**32.9**	27.6

Table 4 shows the improvement of adapted versus unadapted models. Acoustic adaptation provided an impressive improvement of 12.5% for the SDM condition (12.6% for delay-summed MDM) and 5.3% for the individual mic condition. For the distant mic conditions, combining the ICSI and NIST data for adaptation proved to be more effective than source-matched adaptation. Also for the distant mic condition, the best results for CMU were produced by using ICSI-only adapted models. Acoustic adaptation was most effective for ICSI data. One reason is surely that ICSI was the source with by far the most adaptation data. Another likely reason is that ICSI meetings are dominated by speakers that recur throughout the entire corpus, including in the test sets.

The acoustic front-end processing of delay-summing the test signal (as discussed in Section 3.1) produced a further improvement of 6.6%. The delay-summing technique was also most effective for ICSI data, possibly because we had more information about ICSI's meeting room configuration than for the other sources. Delay-summing the adaptation data proved to be not as effective as using acoustic models that were adapted to multiple versions of the signal from all microphones (by 5% relative). This may be because in the latter case channel variability is better represented in the adaptation data.

Table 5 shows WERs with different segmentations. For individual mics, the automatic segmentation increases the WER significantly compared to using reference segmentations. Research on speaker diarization techniques could be a solution in recognizing cross-talk and producing a better segmentation. The cross-talk suppression technique described in Section 3.5 led to a 2% WER reduction. The improvement was largest for the lapel recordings (CMU and LDC); postprocessing was not done for NIST meetings, which seemed to have very little cross-talk.

Finally, Table 6 shows the results on the RT-04 evaluation set, which turned out remarkably similar to the devtest overall. The CMU individual mic recognition is much improved, presumably as a result of the switch to headset mics, though this doesn't seem to be true for LDC. Note that, for the MDM condi-

Table 4: Effect of acoustic adaptation on RT-04 devset. "SM Adapted" means *source-matched*: the source of the adaptation data is matched to the test."I+N adapted" means adapted to *ICSI+NIST* training data. **+**: there was no training data for LDC, so ICSI data was used. *****: recognition on CMU was best with models adapted to ICSI-only, and SDM and MDM results are identical since only 1 microphone was available. Since the CMU and LDC dev data were mismatched to the eval data for IHM (lapel vs. headset), they were given less consideration in making the overall design decisions.

	All	**ICSI**	**CMU**	**LDC**	**NIST**
Individual Mics					
		Headset	Lapel	Lapel	Headset
Unadapted	31.5	20.9	43.6	33.7	28.5
SM Adapted	29.8	17.9	41.4	32.9+	27.6
I+N Adapted	30.3	17.4	43.0	34.0	27.5
Single Distant Mic					
Unadapted	53.6	43.0	60.8	62.9	52.3
SM Adapted	48.5	35.5	60.6	56.0	49.0
I+N Adapted	46.9	34.3	59.0*	54.3	46.9
Multiple Distant Mics (Delay-Summed)					
Unadapted	50.1	35.2	60.7	61.5	49.9
I+N Adapted	43.8	28.4	59.0*	52.3	44.0

Table 5: Overall WERs on RT-04 devset with reference and automatic segmentations, and with post-processing to eliminate cross-talk for IHM. The distant mic signals are delay-summed.

	Ref Seg	**Auto Seg**	**Auto+postproc**
IHM	30.3	36.8	36.1
MDM	42.9	43.8	N/A

Table 6: Results on the RT-04 evaluation set. "H" marks headset, "L" lapel mic conditions.

	All	**ICSI**	**CMU**	**LDC**	**NIST**
Individual Mics					
Dev IHM	36.1	20.5	50.2 L	43.8 L	30.1
RT-04s IHM	34.8	24.2	40.3 H	44.7 H	27.1
Distant Mics					
Dev MDM	43.8	28.4	59.1	52.3	44.0
RT-04S MDM	47.0	20.5	56.4	51.2	41.5
RT-04S SDM	51.3	30.4	56.4	52.2	56.2

Table 7: Results with full recognition system on RT-04 evaluation set.

System	**MDM**	**IHM**	**CTS**
5xRT	47.0	34.8	24.1
Full	44.9	32.7	22.2

tion, even though the per-source WERs are all lower, the overall WER is not, due to the fact that the more difficult sources (CMU and LDC) contribute a larger portion of the test set.

After having developed and tuned the system based on our 5xRT recognition architecture, we ported our current full (20xRT) CTS evaluation system to the Meeting domain. The full system adds a second decoding path using within-word PLP and cross-word MFCC models, lattice regeneration and model readaptation, and a final system combination of three different acoustic models. Table 7 shows overall results for IHM, MDM, and, for reference, 2003 CTS recognition. We see almost identical absolute error reductions on the three test sets, although the relative improvement is somewhat smaller on Meetings (around 5%, compared to 8% for CTS).

5. Conclusions and Future Work

We have shown how a combination of model adaptation, pre- and post-processing techniques can be effective in retargeting a conversational telephone speech recognizer to the meeting recognition task. The severe acoustic mismatch for distant microphones especially was alleviated by a combination of discriminative model adaptation and signal enhancement through noise filtering and array processing. Combined with LM adaptation, we achieved relative improvements of 9% and 22%, respectively, for individual and distant mic conditions. The system gave excellent results in the Spring 2004 NIST evaluation.

Still, many challenges remain. Automatic speech segmentation remains a problem, leading to significant degradation compared to a manual segmentation, which we hope to remedy with the use of novel acoustic features. Meetings also provide fertile ground for future work in areas such as acoustic robustness, speaker-dependent modeling, and language and dialog modeling.

6. Acknowledgments

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-7), and by the Swiss National Science Foundation (through NCCR's IM2 project). We also thank Ramana Gadde, Jing Zheng, and Wen Wang at SRI for advice and assistance.

7. References

[1] http://www.nist.gov/speech/test_beds/mr_proj/

[2] http://nist.gov/speech/tests/rt/

[3] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, Qualcomm-ICSI-OGI features for ASR, ICSLP 2002.

[4] I. Bulyko, M. Ostendorf, and A. Stolcke, Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures, HLT 2003, pp. 7-9.

[5] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke, Voicing Feature Integration in SRI's Decipher LVCSR System. ICASSP 2004, Montreal.

[6] H. Jin, S. Matsoukas, R. Schwartz and F. Kubala, Fast Robust Inverse Transform SAT and Multi-stage Adaptation, Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp. 105-109, Lansdowne, VA, 1998.

[7] N. Kumar, Investigation of Silicon-Auditory Models and Generalisation of Linear Discriminant Analysis for Improved Speech Recognition. PhD thesis, John Hopkins University, 1997.

[8] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations, ICASSP 2003, Hong Kong.

[9] L. Neumeyer and M. Weintraub, Probabilistic Optimum Filtering for Robust Speech Recognition, Proc. ICASSP, Adelaide, Australia, pp. I417-I420, 1994.

[10] D. Povey and P. C. Woodland, Large-scale MMIE Training for Conversational Telephone Speech Recongition, Proc. NIST Speech Transcription Workshop, College Park, MD, 2000.

[11] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen, Maximum Likelihood Discriminant Feature Spaces. ICASSP 2000, pp. 1747-1750.

[12] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, Prosodic Knowledge Sources for Automatic Speech Recognition. ICASSP 2003, pp. 208-211, Hong Kong.

[13] A. Waibel, M. Bett, F. Metze, K. Reis, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, Advances in Automatic Meeting Record Creation and Access, ICASSP 2001.