# Issues in Meeting Transcription – The ISL Meeting Transcription System

*Florian Metze, Qin Jin, Christian Fügen, Kornel Laskowski, Yue Pan, and Tanja Schultz*

Interactive Systems Labs
Universität Karlsruhe (TH), Carnegie Mellon University
{metze|fuegen}@ira.uka.de, {qjin|kornel|ypan|tanja}@cs.cmu.edu

## Abstract

This paper describes the Interactive Systems Lab's Meeting transcription system, which performs segmentation, speaker clustering as well as transcriptions of conversational meeting speech. The system described here was evaluated in NIST's RT-04S "Meeting" speech evaluation.

This paper compares the performance of our Broadcast News and the most recent Switchboard system on the Meeting data and compares both with a newly-trained meeting recognizer. Furthermore we investigate the effects of automatic segmentation on adaptation. Our best meeting system achieves 44.5% on the MDM condition in NIST's RT-04S evaluation.

## 1. Introduction

With the various advances in both hard– and software achieved over time, automatic speech recognition is becoming a viable modality of man-to-machine communication in many situations. Also, it is becoming the backbone for implicit services, that do not require a user, possible engaged in human-to-human communication in a "Meeting" room, to issue a dedicated request for assistance to a machine, but could instead play the role of a context-aware "information butler". In many of these situations however, it is impractical for potential users of speech recognition systems to wear high-quality personal microphones, so speech recognition is hindered not only by the usually colloquial type of speech, where people talk to each other instead of into a microphone, and they also talk at the same time, but also by the difficult recording conditions.

Even though many of these meetings will happen in rooms, not all of these rooms will have dedicated microphone arrays, the most popular technique to deal with the effects of multiple speakers speaking at the same time. Another possible scenario is to use a number of standard microphones distributed on the meeting table, as this set-up is often in place for telephone conferences anyway. Several sites have collected data and NIST conducted the RT-04S evaluation of "Meeting" type speech as part of the "Rich Transcription" series of evaluations.

In this paper, we present the Interactive Systems Lab's most recent speech-to-text system for "Meeting"-type speech, which has evolved significantly over previous versions [1] and which was evaluated in NIST's RT-04S "Meeting" evaluation[1] in the speaker segmentation, and single-distant channel (SDM), multiple-distant (MDM), and personal microphone channel (IPM) speech-to-text conditions.

[1] http://www.nist.gov/speech/tests/rt/rt2004/spring/ This site also contains further information about the data used in the experiments presented

## 2. "Meeting" Data

The experiments presented in this paper were conducted on "Meeting" data which has just recently become available to the research community through LDC. Some of these data sets comprise parallel recordings of both personal (head-set or lapel) microphones and room microphones, which were placed on a conference table which the meeting participant were seated around.

### 2.1. Training Data

All the acoustic data used in this work is in 16kHz, 16bit quality. For training, we merged this corpus with 180h of existing Broadcast News data from the 1996 and 1997 training sets.

| Corpus | Duration | Meetings | Speakers | Channels |
|---|---|---|---|---|
| CMU | 11h | 21 | 93 | 0 |
| ICSI | 72h | 75 | 455 | 4HQ+2LQ |
| NIST | 13h | 15 | 77 | 7 |

Table 1: Meeting training data: all data sets contain a variable number of personal microphone recordings (lapel/ head-set) in addition to the above number of distant microphone recordings

A comprehensive description of each data set with recording conditions and transcription conventions can be found in the literature [2, 3, 4, 5, 6]. Parts of the data have already been used in experiments on segmentation and distant speech recognition [7]. Note that we did not work on the "PDA" low quality data in the ICSI portion.

### 2.2. Development and Test Data

In addition to the data described above, 10-minute excerpts of 8 meetings, two per site, were transcribed for development purposes and another 8 11-minute excerpts of different meetings were used for testing. Each meeting has between 3 and 10 participants while the number of distant channels recorded in parallel varied between 1 (CMU data) and 10 (some LDC meetings).

For the distant microphone conditions, crosstalk regions are labeled in the reference and these are excluded from scoring. Also, some of the close-talking recordings contain a significant amount of energy from non-primary speakers.

## 3. Baseline Experiments

Initial experiments on distant speech were performed using existing systems for the Broadcast News and Conversational Telephony Speech domains. Experiments were run using ISL's Janus toolkit and the Ibis decoder [8, 9].

Our first experiments were run with a speech recognizer trained on BN96 training data [10], which has 2000 codebooks, 6000 distributions, a 42-dimensional feature space based on MFCCs after LDA and global STC transforms [11] with utterance-based CMS. The tri-gram language model was trained on BN96. First-pass decoding WER is 68.4% or 62.8% with VTLN, using both model-space and feature-space MLLR reaches 59.9%.

Experiments with the "Switchboard" recognizer were conducted with a simplified, 3-pass version of ISL's system described in [12]. This systems reaches a WER of 25.0% on the RT-03S "Switchboard" test set. For these experiments, speech was downsampled and passed through a telephony filter. A first-pass decoding using completely unadapted models without even VTLN on a single distant channel results in a word error rate of 64.2%, a system adapted with both model-space and feature-space MLLR reaches 56.4% WER.

Using cross-adaptation between the two systems (which use different language models, dictionaries, and phone sets), it was possible to reduce the error rate to 52.3%, using the Switchboard system for the final pass. All the above experiments were run with manual speaker segmentation and clustering and show performance comparable to previous systems [13].

# 4. Automatic Segmentation

Speaker segmentation and clustering consists of identifying who spoke when in a long meeting conversation. Given a meeting audio, ideally, it will discover how many people are involved in the meeting, and output clusters with each cluster corresponding to an unique speaker.

The speaker and clustering system used for speech recognition ("T2") bases on the acoustic segmentation software CMUseg_0.5 [14]. We removed the classification and clustering components and just used it as a segmenter. A hierachical, agglomerative clustering algorithm is then used to group the segments into clusters. Therefore, we first trained a Tied Gaussian Mixture Model (TGMM) based on the entire speech segments. The GMM for each segment is generated by adapting the TGMM on the segment. The Generalized Likelihood Ratio (GLR) distance is computed between any two segments. At each clustering step, the two closest segments, which have the smallest distance, are mergered. Bayesian Information Criterion (BIC) is used as a stopping criterion for clustering [15].

The speaker segmentation and clustering system for the MDM condition [15] contains two extra steps over the T2 system: unification across multiple channels and speaker turn detection in long segments. The speech recognition experiments throughout this paper uses the T2 system instead of the MDM system, since we found that unification and turn detection resulted in frequent speaker changes and therefore a high fraction of very short utterances which were detrimental to speech recognition performance. The T2 segmentation is computed on the most central channel per meeting only.

| Dataset | Segmentation | |
|---|---|---|
| | T2 | MDM |
| development set | 50.26% | 29.59% |
| evaluation set | 52.54% | 28.17% |

Table 2: Speaker diarization error for the T2 and MDM segmentation

For the IPM case, only segmentation is necessary. But in difference to the SDM/MDM case, mis-segmented parts, with no speech from the speaker of that microphones are counted as insertion errors and lost segments as deletion errors in the later recognition results. Therefore, a good segmentation is essential. So we used a completely different algorithm, which relies in contrast to the other segmentations on activity detection instead of speech detection.

For activity detection in personal microphone audio, each of $N$ channels is first segmented into 300ms non-overlapping frames and preemphasized using a high-pass filter ($1-z^{-1}$). We then compute all $\frac{N\cdot(N+1)}{2}$ crosscorrelations $\phi_{i,j}$ for each pair of channels $\{i,j\}$ and compute $N$ quantities $\Xi_i = \sum_{i\neq j} \frac{\max \phi_{ij}}{\phi_{ii}(0)}$. We declare the frame as speech for channel $i$ if $\Xi_i > 0$. Smoothing is applied independently for each channel over single frame dropouts and padding is added to the beginning and end of each hypothesized speech interval. A more detailed description can be found in [16].

# 5. Training

## 5.1. Acoustic Model Training

As a first step, labels (time-alignments) were written for the close-talking part of the four data sets (BN, CMU, ICSI, NIST) with the BN-based system mentioned above. We then re-trained the BN system with 2k models on the separate data sets.

| Set | BN96/97 | CMU | ICSI | NIST | All |
|---|---|---|---|---|---|
| WER | 67.5% | 68.9% | 67.2% | N/A | 66.7%% |

Table 3: Re-training on the different data sets (2k codebooks, 6k distributions, 100k Gaussians)

Two extra iterations of viterbi training of the "ICSI"-trained system on all channels of the ICSI distant microphone data resulted in a WER of 62.5%. As this step basically trains the full model on the same data several times, we alternatively performed a combination of speaker-adaptive and channel-adaptive (SAT/CAT) training using constrained MLLR [17], by estimating a normalization matrix for every speaker and every recording channel. This resulted in a word error rate of 54.5%, when testing this system with VTLN and normalization matrices estimated on the "ICSI" system. Employing feature space normalization during testing only reaches 58.6%, while performing SAT on the close-talking data alone did not significantly decrease word error rate. Estimating the adaptation parameters of the SAT/CAT system on the previously best hypotheses (52.3% of the SWB system) yields an error rate of 51.4% with roughly a third of the parameters.

As a next step, we re-trained the context decision tree on the combined data sets, increased the model complexity to 6k codebooks, 24k distributions, ∼300k Gaussians while also retraining the STC transform. Re-running the training with these extra parameters, while also adding the NIST distance data reduced the error rate by an extra 3.5% absolute.

The experiments reported so far were run and scored on a pre-release of the official RT-04S development data set, which could not accomodate the Multiple Distant Microphone (MDM) condition. Due to changes to both transcripts and data[2], absolute numbers cannot be compared before and after this point;

[2] Also published on the RT-04S web site

due to recent errata, future numbers will also be slightly off, quantitative assessments of different methods' merits as presented here should however be unaffected and valid.

### 5.2. Language Model Training

Language models were trained in analogy to the Switchboard system. We trained a simple 3-gram LM and a 5-gram LM with ∼800 automatically introduced classes on a mixture of the Switchboard and Meeting transcriptions and also a 4-gram BN LM. All LMs were computed over a vocabulary of ∼47k words with an OOV rate of 0.6% on the development set. For the first decoding passes only the 3-gram LM was used, later decoding and CNC passes uses a 3-fold context dependent interpolation of all three LMs. The perplexity on the development set of the 3-fold interpolated LM was 112.

## 6. Tests

The same meetings were processed by the recognizer in several conditions. Here, the same acoustic and language models were used in a similar manner for each condition to allow comparisons of the respective task's "difficulty".

All tests use a dictionary extended with vocabulary from the meeting domain and the simple language model described above unless stated otherwise. All models use ∼300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks in a 42-dimensional feature space trained as described above. Consensus lattice processing (CLP) [18] and confusion network combination (CNC) was also performed in later stages.

### 6.1. Individual Personal Microphone (IPM) Condition

For the IPM condition we used a reduced version of our Switchboard system, extended by some close talking Meeting Systems. So the following acoustic modells were tested:

**PLAIN** Merge-and-split training followed by Viterbi (2i) on the Close-talking data, no VTLN

**SAT** ≡ PLAIN, but trained with VTLN

**Tree6.8ms** Our Tree6 Switchboard acoustic [12], decoded with 8ms frame shift

**Tree150.8ms** Our Tree150 Switchboard acoustic [12], cross-adapted on Tree6, decoded with 8ms frame shift

**SAT.8ms** cross-adapted on Tree6, decoded with 8ms frame shift

| Models | Segmentation | |
|---|---|---|
| | Manual | IPM-SEG |
| PLAIN | 39.6% | 43.6% |
| SAT | 33.8% | 38.8% |
| Tree6.8ms | 30.8% | 35.0% |
| Tree150.8ms | 29.9% | 34.2% |
| SAT.8ms | 30.2% | 35.3% |
| CNC | 28.0% | 32.7% |

Table 4: Results on the RT-04S development set, IPM condition, CNC is between the last three passes

When comparing the CNC results of both segmentations, it can be seen in table 4 that one of the main problems of the IPM condition is the segmentation. The problem lies mainly in the number of deletion errors, which increases from 9.8% to 14.7%.

### 6.2. Single Distant Microphone (SDM) Condition

The following acoustic models were tested on the SDM microphone condition:

**PLAIN** Merge-and-Split training followed by Viterbi (2i) on the Close-talking data only, no VTLN

**SAT/CAT** Extra 4i Viterbi training on the distant data, no VTLN

**SAT/CAT-VTLN** ≡ SAT/CAT, but trained with VTLN

| Models | Segmentation | |
|---|---|---|
| | Manual | SDM-SEG |
| PLAIN | 59.5% | 60.8% |
| SAT/CAT | 53.2% | 55.2% |
| SAT/CAT-VTLN | 48.9% | 53.1% |
| CNC | 47.8% | 51.5% |

Table 5: Results on the RT-04S development set, SDM condition, CNC is between the last two passes

### 6.3. Multiple Distant Microphone (MDM) Condition

The decoding and adaptation strategy for the MDM condition uses the same models as for the SDM case, but after every decoding step, CNC was performed over all available channels.

| Models | Segmentation | |
|---|---|---|
| | Manual | SDM-SEG |
| PLAIN | 53.4% (59.8%) | 54.4% (60.8%) |
| SAT/CAT | 46.6% (50.7%) | 48.5% (51.9%) |
| SAT/CAT-VTLN | 43.3% (47.7%) | 45.5% (51.5%) |
| Multi-pass CNC | 42.8% | 45.0% |

Table 6: Results on the RT-04S development set, MDM condition; the number in brackets is the performance of a single channel (#1) without CNC

### 6.4. RT04-S Evaluation Results

ISL's submissions to the "sttul" condition of the RT-04S Meeting evaluation reached a word error rate of 35.7% for the IHM condition, 49.5% for the SDM condition, and 45.2% for the MDM condition.

ISL's primary submission to NIST's RT-04S speech-to-text evaluation used a segmentation based on a single distant channel only. To investigate the influence of improved speaker segmentation and clustering on speech-to-text performance, the following table compares STT performance with the SDM segmentation with STT performance based on a MDM segmentation, i.e. a segmentation which uses information from multiple channels and reaches an improved segmentation score of 28.17% compared to 52.54%.

The distribution of errors across the different meetings in the test set and the meeting sites as well as their relation to number of channels and number of speaker clusters generated by the automatic segmentations are shown in table 8.

| Models | Segmentation | |
|---|---|---|
| | SDM-SEG | MDM-SEG |
| PLAIN | 55.4% | 53.7% |
| SAT/CAT | 49.9% | 48.1% |
| SAT/CAT-VTLN | 47.6% | 45.4% |
| Multi-pass CNC | 45.2% | 44.5% |

Table 7: Results on the RT-04S evaluation set, MDM condition; results with CNC of all available channels

| Meeting Site | # CHNS | # SPKS | SDM-SEG # S | SDM-SEG WER | MDM-SEG # S | MDM-SEG WER |
|---|---|---|---|---|---|---|
| CMU | 1 | 6/4 | 2/2 | 47.4% | 3/3 | 46.7% |
| ICSI | 4 (HQ) | 7/7 | 1/3 | 37.6% | 3/4 | 33.7% |
| LDC | 9/5 | 3/3 | 2/4 | 47.8% | 3/2 | 48.8% |
| NIST | 7 | 6/7 | 1/2 | 44.7% | 3/3 | 43.8% |

Table 8: Distribution of errors across the RT-04S Meeting evaluation set (MDM case, 2 meetings per site)

## 7. Conclusions

While these experiments, performed within the RT-04S evaluation framework, are non-exhaustive by far, the results presented in this paper demonstrate a significant improvement over previous "Meeting" speech recognition systems, particularly when using multiple distant microphones not arranged as a microphone array.

A closer analysis of system errors is currently being carried out, but it is clear that speaker segmentation and clustering plays a vital role in improving the performance of adaptation on this type of data; in the SDM case, VTLN works significantly less well with automatic segmentation than with manual segmentation, while CNC can compensate some of the loss. To further improve segmentation, we are therefore planning to use the present speech recognition system in multi-modal rooms, which could combine acoustic and visual evidence with context information, to improve segmentation and adaptation.

## 8. Acknowledgements

Part of this work has been funded by the European Union under IST projects No. IST-2000-28323 (FAME: "Facilitating Agents for Multi-cultural Exchange", `http://isl.ira.uka.de/fame`) and No. FP6-506909 (CHIL: "Computers in the Human Interaction Loop", `http://chil.server.de`).

## 9. References

[1] A. Waibel, H. Yue, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett, "Advances in Meeting Recognition," in *Proc. HLT-2001*. San Diego, CA: ISCA, 3 2001.

[2] S. Burger, V. MacLaren, and H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style," in *Proc. ICSLP-2002*. Denver, CO: ISCA, 9 2002.

[3] S. Burger and Z. Sloan, "The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions," in *Proc. ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004.

[4] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI Meeting Project: Resources and Research," in *Proc. ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004.

[5] S. Strassel and M. Glenn, "Shared Linguistic Resources for Human Language Technology in the Meeting Domain," in *Proc. ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004.

[6] V. Stanford and J. Garofolo, "Beyond Close-talk – Issues in Distant speech Acquistion, Conditioning Classification, and Recognition," in *Proc. ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004.

[7] L. Docio-Fernandez, D. Gelbart, and N. Morgan, "Far-field ASR on Inexpensive Microphones," in *Proc. Eurospeech-2003*. Geneva; Switzerland: ISCA, 9 2003.

[8] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," in *Proc. ICASSP 97*. München; Germany: IEEE, 4 1997.

[9] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Proc. ASRU 2001*. Madonna di Campiglio, Italy: IEEE, 12 2001.

[10] H. Yu and A. Waibel, "Streaming the Front-End of a Speech Recognizer," in *Proc. ICSLP-2000*. Beijing; China: ISCA, 10 2000.

[11] M. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing*, vol. Vol. 2, May 1999.

[12] H. Soltau, H. Yu, F. Metze, C. Fügen, Q. Jin, and S.-C. Jou, "The 2003 ISL Rich Transcription System for Conversational Telephony Speech," in *Proc. ICASSP 2004*. Montreal; Canada: IEEE, 2004.

[13] R. R. Gade, D. Gelbart, T. Pfau, A. Stolcke, and C. Wooters, "Experiments with Meeting Data," in *Proc. RT02 Workshop*. Vienne, VA: NIST, 5 2002.

[14] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," in *Proc. DARPA Speech Recognition Workshop*, 1997.

[15] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker Segmentation and Clustering in Meetings," in *Proc. ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004.

[16] K. Laskowski, Q. Jin, and T. Schultz, "Cross-correlation–based Multispeaker Speech Activity Detection," in *subm. Proc. ICSLP-2004*. Jeju; Korea: ISCA, 10 2004.

[17] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Cambridge University, Cambridge, UK, Tech. Rep., 1997.

[18] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.