THE ISL RT-04S MEETING TRANSCRIPTION SYSTEM

Florian Metze and Christian Fügen

Interactive Systems Labs
Universität Karlsruhe (TH)
Karlsruhe, Germany
{metze|fuegen}@ira.uka.de

Yue Pan, Tanja Schultz, and Hua Yu

Interactive Systems Labs
Carnegie Mellon University
Pittsburgh, PA
{ypan|tanja|hyu}@cs.cmu.edu

ABSTRACT

This paper describes the speech recognition (STT) part of the Interactive Systems Lab's 2004 Meeting transcription system, for the IPM (Individual Personal Microphone), SDM (Single Distant Microphone), and MDM (Multiple Distant Microphone) conditions; which was evaluated in NIST's RT-04S "Meeting" evaluation.

We compare the performance of our Broadcast News and the most recent Switchboard system on the Meeting data and compare both with the newly-trained meeting recognizer. Furthermore, we investigate the effects of automatic segmentation on adaptation. Our best (post-evaluation) Meeting system achieves a WER of 44.5% on the "MDM" condition.

1. INTRODUCTION

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]. The ISL system was submitted to the "ul" (unlimited run-time) condition.

2. "MEETING" DATA

The system described in this paper is trained on 16kHz/ 16bit quality audio and was newly trained using mostly the recently released "Meeting" training data. We used 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, whenever possible.

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

2.1. Training Data

Training data of all acoustic models in the ISL system consisted of the "Meeting" training data (see table 1) merged with 180h of existing Broadcast News data from the 1996 and 1997 training sets.

Corpus	Duration	Meetings	Speakers	Channels
CMU	11h	21	93	N/A
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]. Parts of the data have already been used in experiments on segmentation and distant speech recognition [6]. Note that we did not work on the "PDA" low quality data in the ICSI portion of the training data.

2.2. Development and Test Data

The decoding experiments described in this paper were conducted on the following data sets:

RT02 The RT-02 "Meeting" evaluation set, containining one distant channel only (80min)

Dev The official RT-04S development test set, derived from RT02, containing several distant channels (90min)

Eval The RT-04S evaluation set (90min)

For the Dev and Eval sets, the erratum concerning the selection of the best channel for the NIST data has not been applied, unless explicitely noted.

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, personal-microphone recordings contain a significant amount of cross-talk from non-primary speakers, particularly for the CMU meetings, which were recorded with consumer-grade equipment, to be as "real" as possible.

3. BASELINE EXPERIMENTS

All experiments described in this paper were run using ISL's Janus toolkit and the Ibis decoder [7, 8] in version 5.0, patch-level 013.

Our first experiments were run with a speech recognizer trained on BN96 training data, which has 2000 codebooks, 6000 distributions, a 42-dimensional feature space based on MFCCs after LDA and global STC transforms [9] with utterance-based CMS. The tri-gram language model was trained on BN96. This system performed better on in-house meeting data than our standard BN recognizer [10]. First-pass decoding WER on NIST data 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 [11]. 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 [12].

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. However in speech recognition, the goal of speaker segmentation and clustering is to serve speaker adaptation. Speaker adaptation concerns more about regression of speakers, not strict classification of speakers. So if two speakers sound reasonably indistinguishable, they can be considered as equal and grouped into one cluster.

The speaker segmentation and clustering system used for speech recognition ("T2") is based on CMUseg_0.5 [13]. Of this software package, we used the segmenter part and added a hierarchical, agglomerative clustering algorithm 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 segments with the smallest distance are merged. Bayesian Information Criterion (BIC) is used as a stopping criterion for clustering.

The speaker segmentation and clustering system for the MDM condition 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 use the T2 system instead of the MDM system, since unification and turn detection initially 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 (as defined before post-evaluation errata) per meeting only; also, segments longer than 15s were cut at positions where an initial quick transcription pass generated noise or silence tokens with a duration of more than 40ms.

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. Opposed to the SDM/MDM case however, mis-segmented parts, with no speech from the primary speaker of that microphone result in insertion errors and lost segments in deletion errors during STT scoring. To deal with this situation, we used a completely different algorithm, which, in contrast to the other segmentations, relies 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.

The ISL STT system used the following different segmentation systems, which did also perform speaker clustering for the distant-speech cases:

IPM Used for the IPM (personal/ close-talking) system, based on activity detection [14]

T2 A single-channel segmentation, also used for the evaluated MDM system, as the MDM segmentations available at the beginning of the evaluation period contained a high degree of short sentences, which were unsuitable for speech recognition

MDM A multi-channel segmentation based on a unification of the available single-channel segmentations

A more detailed description of these can be found elsewhere in these proceedings [15].

5. TRAINING

5.1. Acoustic Model Training

As a first step, we generated time-alignments and warping factors for the close-talking part of three 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	Merged
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); test on pre-release of RT-04S development data ($\approx$ RT-02 Meeting test data)

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%. Employing feature space normalization (constrained MLLR) [16] and VTLN during testing only reaches 58.6%. As an alternative to Viterbi training we performed a combination of speaker-adaptive and channel-adaptive (SAT/CAT) training also using constrained MLLR, 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.

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, $\sim$300k Gaussians while also re-training 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, and the best performance was delivered by a system using newly trained models alone.

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; 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 $\sim$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 $\sim$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

All tests use a dictionary extended with vocabulary from the meeting domain and the simple language model described above for decoding unless stated otherwise. All models use $\sim$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) [17] and confusion network combination (CNC) was also performed in later stages using the interpolated language model.

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 models were tested:

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

SAT $\equiv$ PLAIN, but trained with VTLN

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

Tree6.8ms Our Tree6 Switchboard acoustic [11], decoded with 8ms frame shift

Tree150.8ms Our Tree150 Switchboard acoustic [11], 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

Comparing CNC results of both segmentations in table 4, it is clear that segmentation is one of the IPM condition's main problems. This lies mainly in the number of deletion errors, which increases from 9.8% to 14.7%. Processing resulted in a real-time factor (RTF) of 173 on 3GHz Intel Pentium4 machines with Hyper-Threading enabled under Linux (2 jobs per processor).

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 $\equiv$ SAT/CAT, but trained with VTLN

Processing time was $\approx$84h, resulting in a real-time factor (RTF) of 56. Every single decoding pass runs with RTF < 20 (also for the MDM case).

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. Overall, processing resulted in a RTF of $\approx$259.

Models	Segmentation	
	Manual	SDM-SEG (T2)
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

Models	Segmentation	
	Manual	SDM-SEG (T2)
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 STT evaluation reached a word error rate of 35.7% for the IPM, 49.5% for the SDM, and 45.2% for the MDM condition. To investigate the influence of improved speaker segmentation and clustering on STT performance, the following table compares STT performance with the "T2" segmentation with that based on the submitted MDM segmentation, which uses information from multiple channels and reaches a segmentation score of 28.17% compared to 52.54%. However, this segmentation only became available for ASR experiments after the evaluation deadline.

Models	Segmentation	
	SDM-SEG (T2)	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. The unification and smoothing of the segmentation across channels results in lower WERs already for the non-adapted case

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

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). Different segmentation algorithms hypothesize a different number of speakers, which has a large influence on the performance of adaptation

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. Other approaches to channel combination (more suitable for systems with constrained realtime requirements) will also be investigated. 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. Yu, 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 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

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

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

[17] 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.