AUTOMATICALLY TRANSCRIBING MEETINGS USING DISTANT MICROPHONES

Florian Metze, Christian Fügen

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

Yue Pan, Alex Waibel

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

ABSTRACT

In this paper, we describe our efforts to develop acoustic models suitable for distant microphone automatic speech recognition. Our goal is to investigate, how the performance of a system trained on a combination of close-talking and distant microphone data can be optimized, while assuming as little information about the configuration of (multiple) distant microphones as possible, to avoid guesstimates and lengthy calibration runs.

We evaluated our system in NIST's RT-04S "Meeting" speech-to-text evaluation, where speech data was recorded at several sites with a varying number of different table-top microphones, but not with microphone arrays. Body-mounted microphones provide baseline numbers for distant ASR performance and allow for comparisons of meeting speech with other spontaneous speech data.

1. INTRODUCTION

An important effort in current speech research is focused on the processing of speech from natural multi-party interaction, aka "Meetings", which presents a number of new challenges in terms of style (highly interactive), segmentation (overlapping) as well as difficult recording condition(s). Data gathered during meetings provides an interesting testbed for work on robust automatic speech recognition, speaker detection, segmentation and tracking, discourse modeling, and many more. Ideally, automatic systems working on these tasks operate on data recorded from distant microphones, freeing users from the need to wear body-mounted microphones. As specialized microphone arrays will not be available in many cases, research should investigate speech recorded through room microphones, which could for example be built into hands-free telephone sets or other mobile units.

In this paper, we present the current Interactive Systems Lab's speech-to-text system for "Meeting"-type speech, which was evaluated in NIST's RT-04S "Meeting" evaluation [1, 2, 3]. The focus of this paper is on the rationale behind some of the design decisions and the experiments with the core speech-to-text system for (multiple) distant microphones.

2. THE "MEETING" SCENARIO AND DATA

"Meeting" data used in this work mainly consists of group meetings in a professional or research environment, where participants were usually seated around a table. As the meetings occured naturally, they contain spontaneous effects and sloppy speech, although the amount varies among the four collection sites CMU, ICSI, LDC, and NIST. Recordings were done with individual and distant microphones.

2.1. Training Data

Training data was available from three sites in 16kHz, 16bit quality, see table 1. The CMU data was recorded with lapel microphones, while the other groups used head-sets. Although the layout differed between sites, the distant microphones were generally of table-top, omni-directional type roughly distributed along an axis on the middle of the conference table. The NIST data contains directional microphones as well. No training data was collected at LDC, we also disregard the "Mock-PDA" data from ICSI.

Corpus	Duration	# Meetings	# Speakers	# Channels
CMU	11h	21	93	0
ICSI	72h	75	455	4
NIST	13h	15	77	7

Table 1. Meeting training data: all data sets contain recordings of individual speakers with personal microphones in addition to the above number of distant microphone recordings.

Pointers to these corpora as well as descriptions of their properties are available on the RT-04S web-site [1], the data is available through LDC. For training our recognizer, we merged these corpora with 180h of Broadcast News data from the 1996 and 1997 training sets. For language modeling, we also added the transcriptions for 360h of Switchboard data from phases I, II, "Cellphone" and "C-Tran".

2.2. Development and Test Data

Three evaluation conditions were defined for RT-04S:

MDM Multiple Distant Microphones (primary)

SDM Single Distant Microphone (optional)

IHM Individual Head-set Microphone (required contrast)

The same meeting can therefore be processed several times using different amounts of information. Development data for the RT-04S evaluation consisted of 10-minute excerpts of eight meetings, two per site. Eight 11-minute excerpts of different meetings (two per site) were used for evaluation. Each meeting has between three and ten participants while the number of distant channels varied between one (CMU) and ten (some LDC meetings).

For the distant microphone conditions, crosstalk regions, roughly three quarters of the data, are labeled in the reference transcriptions and excluded from scoring. The respective manual segmentation was derived from these transcriptions and the resulting segments only contain non-crosstalk regions. The SDM condition can be derived from the MDM condition by disregarding all but one "central" distant channel for every meeting.

3. SYSTEM DESIGN

3.1. Automatic Segmentation and Clustering

Speaker segmentation and clustering consists of identifying who spoke when in a long meeting conversation. Ideally, the process will discover how many people are involved in the meeting, and output clusters corresponding to an unique speaker each. This information is needed for speaker adaptation in multi-pass decoding as well as higher-level processing. This paper presents results on the RT-04S development data using manual and automatic segmentation.

Our system uses CMUseg_0.5 and a hierarchical, agglomerative clustering algorithm [4, 5]; in this work we use a common segmentation for SDM and MDM conditions.

3.2. Language Model Training

Language Model	Overall	CMU	ICSI	LDC	NIST
SWB-3G	54.8	65.0	47.1	57.4	54.3
Meeting-3G	53.4	64.9	41.3	60.7	53.4
Merged-3G	52.4	63.7	42.6	55.9	53.4
3-fold Interpolated	51.6	63.7	41.5	55.8	51.4

Table 2. Language Model development: word error rate in % on "SDM" condition using baseline Switchboard acoustic models.

Language models were trained in analogy to our RT-03S Switchboard system [6], see table 2. We trained a standard 3-gram LM and a 5-gram LM with ∼800 automatically deduced classes on a mixture of the Switchboard and Meeting transcriptions, as we considered these to be similar in style. We also trained a 4-gram Broadcast News LM. All LMs were computed over a vocabulary of ∼47k words, which resulted in an OOV rate of 0.6% on the development set. Distant speech decodings were run with the merged 3-gram LM. Confusion Network generation/ combination passes use a context-dependent interpolation of all three LMs, which was also directly used in the IHM decodings. The perplexity on the development set of the 3-fold interpolated LM was 112. We did not add data downloaded from the web or adapt the models to meeting or site, although they were very different in topic and style.

3.3. Acoustic Model Training

The 16kHz recognizers used in these experiments work in a 42-dimensional feature space based on MFCCs with CMS and CVN applied on a per-utterance basis. We use a ± 7 frames context window before applying separate LDA and global STC transforms [7]. No specific noise-filtering has been employed for distant data.

Our first experiments were run with a 2k codebooks, 6k distribution, 100k diagonal Gaussians system trained on BN96 training data only. Initial word error rate on Meeting data ("SDM" condition, i.e. one, central channel only; manual segmentation) is 62.8% with VTLN, using both model-space and feature-space MLLR we reach 59.9%.

Experiments with the "Switchboard" recognizer were conducted with a simplified, 3-pass version of ISL's system described in [6], which reaches a word error rate of 25.0% on the RT-03S "Switchboard" test set. For the Meeting experiments, speech was down-sampled and passed through a telephony filter. A first-pass decoding using completely unadapted models results in a word error rate of 64.2%, a VTLN system adapted with both model-space and feature-space MLLR reaches 56.4% word error rate.

Using cross-adaptation between the two systems, it was possible to reduce the error rate to 52.3%, using the Switchboard system for the final pass.

As our Switchboard system had been trained on ∼360h of telephony speech only and the combination of BN and Meeting data would yield ∼300h of close-talking or BN speech plus about the same amount of in-domain distant speech, we decided to re-train a 16 kHz system from scratch.

Training / Test (%WER)	Pooled	BN96/97 (180h)	ICSI (75h)	CMU (11h)	NIST (13h)
CMU	72.3	71.9	70.6	71.9	74.0
ICSI	60.2	62.2	59.9	63.0	67.2
LDC	67.9	68.2	69.1	71.8	76.6
NIST	71.4	72.7	75.2	72.9	75.8
Overall	66.7	67.5	67.2	68.9	72.6

Table 3. Results of training a "SDM" system on the different data sets: pooling BN and Meeting data improves robustness.

To see if merging the data was indeed a viable approach, we trained simple systems of equal size on different portions of close-talking data and tested these on the central channel of the distant Meeting development test. Results are summarized in table 3. It is interesting to note that the "CMU" system performs significantly better on the distant data than the "NIST" system with also little training data. We attribute this effect to the use of lapel microphones, which capture more room acoustics.

Two extra iterations of Viterbi training of the "ICSI"-trained system on all four high-quality channels of the ICSI distant microphone data resulted in a word error rate of 62.5%, an improvement of 5% absolute. Employing feature space normalization (constrained MLLR) [8] and VTLN during testing only reaches 58.6%. Alternatively we performed a combination of channel-adaptive (CAT) and speaker-adaptive (SAT) training also using constrained MLLR [9], by estimating a separate normalization matrix for every speaker and every recording channel. This resulted in a word error rate of 54.5%, which is a 8% absolute (13% relative) gain. Performing SAT alone on the close-talking data 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 assigned by the Merge-and-Split algorithm while also re-training STC. Re-running the close-talking and distant speech training with these extra parameters, while also adding the NIST distance data to the second step reduced the error rate by an extra 3.5% absolute, and the best performance was delivered by a system using newly trained models alone; no further improvement was possible using cross-adaptation with SWB models.

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, absolute error rate cannot be compared before and after this point; quantitative assessments of different methods' merits however are unaffected and valid.

4. RESULTS

4.1. Single Distant Microphone

Experimentation with adaptation and decoding with the above setup led to the following decoding strategy, where second- and third-pass models were adapted with model-space and feature-space MLLR using the hypothesis generated in the preceeding step. A single decoding pass takes less then 5 RTF on a 3GHz Pentium4 machine, memory consumption is typically 250Mb when ignoring the footprint of cached audio data.

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

SAT/CAT-noVTLN ≡ PLAIN with extra SAT/ CAT Viterbi (4i) training on the distant data, no VTLN

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

CNC Confusion Network Combination

Models	Segmentation	
	Manual	Automatic
PLAIN	59.5	60.8
SAT/CAT-noVTLN	53.2	55.2
SAT/CAT	48.9	53.1
CNC	47.8	51.5

Table 4. Decoding results (%WER) on the RT-04S development set, SDM condition, CNC is between the last two passes.

Confusion Networks [10] were generated from the union of different lattices, where confidences were computed separately on the individual lattices after pruning. Here, we are combining lattices from the last two decoding passes.

4.2. Individual Microphones

For comparison, we also report results for our close-talking system. For the IPM condition, we used a simplified 3-pass version of the Switchboard system [6] together with several close-talking Meeting models:

PLAIN ≡ first pass of SDM case

SAT/CAT ≡ third pass of SDM case

Tree6.8ms Tree6 Switchboard MLE-SAT AM, decoded with 8ms frame shift

Tree150.8ms Tree150 Switchboard MMIE-SAT AM, cross-adapted on Tree6, decoded with 8ms frame shift

SAT/CAT.8ms ≡ SAT/CAT, cross-adapted on Tree6, decoded with 8ms frame shift

Models	Segmentation	
	Manual	Automatic
PLAIN	39.6	43.6
SAT/CAT	33.8	38.8
Tree6.8ms	30.8	35.0
Tree150.8ms	29.9	34.2
SAT/CAT.8ms	30.2	35.3
CNC	28.0	32.7

Table 5. Decoding results (%WER) on the RT-04S development set, IPM condition

On the close-talking data, using three different acoustic models (one 16kHz, two 8kHz with different optimization criteria) and adapting these on each other leads to a large reduction in word error rate, which is again significantly reduced by Confusion Network Combination (here between last three passes).

4.3. Multiple Distant Microphone (MDM) Condition

The decoding and adaptation strategy for the MDM condition uses the same models and the same decoding setup as the SDM case, but after every decoding step, CNC was performed over all channels (one to eight, depending on site) processed in the last step.

Models	Segmentation	
	Manual	Automatic
PLAIN	53.4 (59.8)	54.4 (60.8)
SAT/CAT-noVTLN	46.6 (50.7)	48.5 (51.9)
SAT/CAT.8+10ms	43.3 (47.7)	45.5 (51.5)
CNC	42.8	45.0

Table 6. Decoding results (%WER) on the RT-04S development set, MDM condition; the number in brackets is the performance of a single channel without CNC.

Computing Confusion Networks at the initial 60% word error rate immediately reduces word error rate by more than 10% relative over the whole data set, which includes 25% data with only one channel (CMU). The possibility to adapt on this hypothesis leads to a gain of approximately 1.5% absolute in single-channel word error rate for the SAT/CAT pass. The gain is more pronounced for the automatic segmentation case. Final CNC is between last two passes and multiple channels including decodings with different frame rate.

4.4. Summary and Comparison

Segmentation for the IPM condition proved surprisingly difficult, as we observe a 14.7% deletion rate, which is nearly as high as the one for the SDM case (16.7%). Manual segmentation has a deletion rate of 9.8%, see table 7. In many meetings, significant amounts of speech from non-primary talkers can be found in the IHM recordings, which makes trade-off between insertions and deletions difficult to optimize without hand-tuning. The recognizer's performance broken down according to data collection site is shown in table 8.

To further improve system performance for the distant microphone case, we tried adapting our recognizer to whole meetings

Condition	SUB		DEL		INS		WER	
IHM	16.0	15.1	9.8	14.7	2.2	2.9	28.0	32.7
SDM	27.8	30.7	17.4	16.7	2.6	4.1	47.8	51.5
MDM	24.1	25.8	16.4	15.9	2.3	3.3	42.8	45.0

Table 7. Error distribution for the three conditions in the RT-04S "Meeting" task. Left number in column is with manual segmentation, right number is automatic segmentation.

%WER	IHM		SDM		MDM	
Segm.:	Man.	Auto.	Man.	Auto.	Man.	Auto.
CMU	39.6	43.0	59.8	63.4	60.7	62.9
ICSI	16.2	20.4	32.5	36.5	27.5	30.1
LDC	28.9	33.3	52.9	56.3	48.1	48.9
NIST	28.2	35.0	57.0	60.7	44.5	47.9
Overall	28.0	32.7	47.8	51.5	42.8	45.0

Table 8. Word error rates for the individual sites making up the RT-04S development data. CMU is most difficult in all conditions, indicating it has spontaneous speech and only one distant channel. Channel combination significantly reduces word error rate for ICSI (which represents a large part of training data), LDC; and NIST.

(generally longer than 60 minutes) instead of only the evaluation part. Presumably due to the quality of the automatic segmentation, this did not lead to a gain in performance. A diagnostic experiment, in which we "filtered" automatic segmentation with the best-matching true speaker segments, so that each cluster would only be adapted on the speech of one speaker, did also not increase performance, as only little adaptation data survived after filtering. Unfortunately, the whole meetings have not yet been manually segmented.

In our training experiments, we achieved best results with acoustic models seeded with pooled close-talking data and then trained in a normalized feature space on parallel recordings of distant speech. For combining several distant channels during decoding, we achieved best results with Confusion Network Combination. A particular advantage of this approach over Array Processing (a simple form of delay&sum beamforming to compensate for time skew and sound travel delays) or Multi-Stream processing (evaluating acoustic models separately for each channel and using the averaged log-likelihood during beam-search) is the robustness of the gains, as no assumption on microphone type, recording location and relative position of speaker and microphone is necessary. The latter two approaches, in our experiments, did not significantly reduce word error rate without strong "educated guesses" about reasonable channels and parameters for combination.

5. CONCLUSION

ISL's primary "sttul" submissions to the NIST's RT-04S "Meeting" evaluation as presented in this paper gave excellent results and reached a word error rates of 35.7%, 49.8%, and 44.9% for the IHM, SDM, and MDM conditions respectively on the evaluation set.

The results demonstrate a significant improvement over previous "Meeting" systems, particularly when using multiple distant microphones not arranged as a microphone array. We are already using an improved version of the "SDM" SAT/CATno-VTLN system for realtime speaker-independent topic spotting around an "augmented table" with very good results. As keywords appear frequently and repeatedly in this application, cross-talk is not such a significant problem here. Distant speech "Meeting" recognition, and the problems it poses in the areas of segmentation and clustering, robust pre-processing, acoustic modeling, and channel combination, as well as language modeling and natural language processing however remains a challenging task for future research.

6. ACKNOWLEDGEMENTS

Part of this work has been funded by the European Union under IST project No. FP6-506909 "CHIL – Computers in the Human Interaction Loop", `http://chil.server.de`.

7. REFERENCES

[1] *Rich Transcription 2004 Spring Meeting Recognition Evaluation*. NIST, 2004, http://www.nist.gov/speech/tests/rt/rt2004/spring/.

[2] Florian Metze, Qin Jin, Christian Fügen, Kornel Laskowski, Yue Pan, and Tanja Schultz, "Issues in Meeting Transcription – The ISL Meeting Transcription System," in *Proc. INTERSPEECH2004-ICSLP*. 10 2004, ISCA.

[3] N. Mirghafori, A. Stolcke, C. Wooters, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, , and M. Ostendorf, "From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System," in *Proc. INTERSPEECH2004 – ICSLP*, Jeju Island; Korea, 10 2004, ISCA.

[4] Qin Jin, Kornel Laskowski, Tanja Schultz, and Alex Waibel, "Speaker Segmentation and Clustering in Meetings," in *Proc. ICASSP-2004 Meeting Recognition Workshop*, Montreal; Canada, 5 2004, NIST.

[5] Kornel Laskowski, Qin Jin, and Tanja Schultz, "Cross-correlation–based Multispeaker Speech Activity Detection," in *subm. Proc. ICSLP-2004*, Jeju; Korea, 10 2004, ISCA.

[6] Hagen Soltau, Hua Yu, Florian Metze, Christian Fügen, Qin Jin, and Szu-Chen Jou, "The 2003 ISL Rich Transcription System for Conversational Telephony Speech," in *Proc. ICASSP 2004*, Montreal; Canada, 2004, IEEE.

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

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

[9] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala, "Fast Robust Inverse Transform SAT and Multi-stage Adaptation," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA; USA, 1998.

[10] 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.