Advances in Meeting Recognition

Alex Waibel[c,k] *Hua Yu*[c] *Martin Westphal*[k] *Hagen Soltau*[k]
Tanja Schultz[c,k] *Thomas Schaaf*[k] *Yue Pan*[c] *Florian Metze*[k] *Michael Bett*[c]
Interactive Systems Labs
[c]Carnegie Mellon University, NSH 2602, Pittsburgh, PA, 15217, USA
[k]Universität Karlsruhe, Fakultät für Informatik, 76128 Karlsruhe, Germany

1 Introduction

Speech recognition has advanced considerably, but has been limited almost entirely either to situations in which close speaking microphones are natural and acceptable (telephone, dictation, command&control, etc.) or in which high-quality recordings are ensured. Furthermore, most recognition applications involve controlled recording environments, in which the user turns the recognition event on and off and speaks cooperatively for the purpose of being recognized.

Unfortunately, the majority of situations in which humans speak with each other fall outside of these limitations. When we meet with others, we speak without turning on or off equipment, or we don't require precise positioning vis a vis the listener. Recognition of speech during human encounters, or "meeting recognition", therefore represents the ultimate frontier for speech recognition, as it forces 1.) robustness, 2.) knowledge of context, and 3.) integration in an environment and/or human experience.

2 Challenges

Over the last 3 years we have explored meeting recognition at the Interactive Systems Laboratories. Meeting recognition is performed as one of the components of a "meeting browser"; a search retrieval and summarization tool that provides information access to unrestricted human interactions and encounters. The system is capable of automatically constructing a searchable and browsable audiovisual database of meetings. The meetings can be described and indexed in somewhat unorthodox ways, including by what has been said (speech), but also by who said it (speaker&face ID), where (face, pose, gaze, and sound source tracking), how (emotion tracking), and why, and other meta-level descriptions such as the purpose and style of the interaction, the focus of attention, the relationships between the participants, to name a few (see [1, 2, 3, 4]).

The problem of speech recognition in unrestricted human meetings is formidable. Error rates for standard recognizers are 5-10 times higher than for dictation tasks. Our explorations based on LVCSR systems trained on BN, reveal that several types of mismatches are to blame [5]:

1. Mismatched and/or degraded recording conditions (remote, different microphone types),
2. Mismatched dictionaries and language models (typically ideosynchratic discussions highly specialized on a topic of interest for a small group and therefore very different from other existing tasks),
3. Mismatched speaking-style (informal, sloppy, multiple speakers talking in a conversational style instead of single speakers reading prepared text).

In the following sections, we describe experiments and improvements based on our Janus Speech Recognition Toolkit JRTk [6] applied to transcribing meeting speech robustly.

3 Experimental Setup

As a first step towards unrestricted human meetings each speaker is equipped with a clip-on lapel microphone for recording. By this choice interferences can be reduced but are not ruled out completely. Compared to a close-talking headset, there is significant channel cross-talk. Quite often one can hear multiple speakers on a single channel. Since meetings consist of highly specialized topics, we face the problem of a lack of training data. Large databases are hard to collect and can not be provided on demand. As a consequence we have focused on building LVCSR systems that are robust against mismatched conditions as described above. For the purpose of building a speech recognition engine on the meeting task, we combined a limited set of meeting data with English speech and text data from various sources, namely Wall Street Journal (WSJ), English Spontaneous Scheduling Task (ESST), Broadcast News (BN), Crossfire and Newshour TV news shows. The meeting data consists of a number of internal group meeting recordings (about one hour long each), of which fourteen are used for experiments in this paper. A subset of three meetings were chosen as the test set.

4 Speech Recognition Engine

To achieve robust performance over a range of different tasks, we trained our baseline system on Broadcast News (BN). The system deploys a quinphone model with 6000 distributions sharing 2000 codebooks. There are about 105K Gaussians in the system. Vocal Tract Length Normalization and cluster-based Cepstral Mean Normalization are used to compensate for speaker and channel variations. Linear Discriminant Analysis is applied to reduce feature dimensionality to 42, followed by a diagonalization transform (Maximum Likelihood Linear Transform). A 40k vocabulary and trigram language model are used. The baseline language model is trained on the BN corpus.

Our baseline system has been evaluated across the above mentioned tasks resulting in the word error rates shown in Table 1. While we achieve a first pass WER of 18.5% on all F-conditions and 9.6% on the F0-conditions in the Broadcast News task, the word error rate of 44.2% on meeting data is quite high, reflecting the challenges of this task. Results on the ESST system [7] are even worse with a WER of 54.1% which results from the fact that ESST is a highly specialized system trained on noise-free but spontaneous speech in the travel domain.

4.1 Acoustic and Language Model Adaptation

The BN acoustic models have been adapted to the meeting data thru Viterbi training, MLLR (Maximum Likelihood Linear Regression), and MAP (Maximum A Posteriori) adaptation. To improve the robustness towards the unseen channel conditions, speaking mode and training/test mismatch, we trained a system "BN+ESST" using a mixed training corpus. The comparison of the results indicate that the mixed system is more robust (44.2% $\rightarrow$ 42.2%), without loosing the good performance on the original BN test set (18.5% vs. 18.4%).

To tackle the lack of training corpus, we investigated linear interpolation of the BN and the meeting (MT) language model. Based on a cross-validation test we calculated the optimal interpolation weight and achieved a perplexity reduction of 21.5% relative compared to the MT-LM and more than 50% relative compared to the BN-LM. The new language model gave a significant improvement decreasing the word error rate to 38.7%. Overall the error rate was reduced by 12.4% relative (44.2% $\rightarrow$ 38.7%) compared to the BN baseline system.

System WER on Different Tasks [%]	
BN (h4e98_1) F0-condition	9.6
BN (h4e98_1) all F-conditions	18.5
BN+ESST (h4e98_1) all F-conditions	18.4
Newshour	20.8
Crossfire	25.6
Improvements on Meeting Recognition	
Baseline ESST system	54.1
Baseline BN system	44.2
+ acoustic training BN+ESST	42.2
+ language model interpolation (14 meetings)	39.0
Baseline BN system	
+ acoustic MAP Adaptation (10h meeting data)	40.4
+ language model interpolation (14 meetings)	38.7

Table 1: **Recognition Results**

4.2 Model Combination based Acoustic Mapping (MAM)

For the experiments on meeting data reported above we have used comparable recording conditions as each speaker in the meeting has been wearing his or her own lapel microphone. Frequently however this assumption does not apply. We have also carried out experiments aimed at producing robust recognition when microphones are positioned at varying distances from the speaker. In this case data, specific for the microphone distance and SNR found in the test condition is unavailable. We therefore apply a new method, Model Combination based Acoustic Mapping (MAM) to the recognition of speech at different distances. MAM was originally proposed for recognition in different car noise environments, please refer to [8, 9] for details.

Word Error Rate [%]		Close	Distant
Baseline	Speaker 1	35.2	46.1
MAM	Speaker 1	-	38.2
Baseline	Speaker 2	13.8	31.2
MAM	Speaker 2	14.4	26.1

Table 2: Recognition results on Model Combination based Acoustic Mapping (MAM)

MAM estimates an acoustic mapping on the log-spectral domain in order to compensate for noise condition mismatches between training and test. During training, the generic acoustic models λ_k $(k = 1, 2, ..., n)$ and a variable noise model N are estimated. Then, model combination is applied to get new generic models $\hat{\lambda}_k = \lambda_k + N$, which correspond to noisy speech. During decoding of a given input x, the mapping process requires a classification as a first step. The score for each $class(model)$ is computed as $g_k(x) = P(k|x, \hat{\lambda}_k)$. In the second step x is reconstructed according to the calculated score, where μ refers to the mean vector: $\hat{x} = x + \sum_{k=1}^{n} g_k(x)(\mu_k - \hat{\mu}_k)$.

We applied MAM to data that was recorded simultaneously by an array of microphones positions at different distances from the speaker. Each speaker read several paragraphs of text from the Broadcast News corpus. The results of experiments with two speakers are summarized in Table 2. Experiments suggest that MAM effectively models the signal condition found in the test resulting in substantial performance improvements.

5 Conclusions

In this paper we have reviewed work on speech recognition systems applied to data from human-to-human interaction as encountered in meetings. The task is surprisingly challenging with error rates of 5-10 times higher than read speech (BN F0-condition). Our experiments using different training data, language modeling interpolation, adaptation and signal mapping yield more than 20% relative improvements in error rate.

6 Acknowledgment

We would like to thank Susanne Burger, Christian Fügen, Ralph Gross, Qin Jin, Victoria Maclaren, Robert Malkin, Laura Mayfield, John McDonough, Thomas Polzin, Klaus Ries, Ivica Rogina, and Klaus Zechner for their support.

References

[1] Klaus Ries, "Towards the Detection and Description of Textual Meaning Indicators in Spontaneous Conversations," in *Proceedings of the Eurospeech*, Budapest, Hungary, September 1999, vol. 3, pp. 1415–1418.

[2] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel, "Multimodal Meeting Tracker," in *Proceedings of RIAO2000*, Paris, France, April 2000.

[3] Rainer Stiefelhagen, Jie Yang, and Alex Waibel, "Simultaneous Tracking of Head Poses in a Panoramic View," in *International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, September 2000.

[4] Thomas S. Polzin and Alex Waibel, "Detecting Emotions in Speech," in *Proceedings of the CMC*, 1998.

[5] Hua Yu, Michael Finke, and Alex Waibel, "Progress in Automatic Meeting Transcription," in *Proceedings of the EUROSPEECH*, September 1999.

[6] Torsten Zeppenfeld, Michael Finke, Klaus Ries, and Alex Waibel, "Recognition of Conversational Telephone Speech Using the Janus Speech Engine," in *Proceedings of the ICASSP'97*, München, Germany, 1997.

[7] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, "Multilingual Speech Recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer-Verlag, 2000.

[8] Martin Westphal "Robust Continuous Speech Recognition in Changing Environments", University of Karlsruhe, Ph.D. thesis, 2000.

[9] Martin Westphal "Model-Combination-Based Acoustic Mapping", in *Proceedings of the ICASSP'01*, To appear in May 2001.