EXPERIMENTS IN AUTOMATIC MEETING TRANSCRIPTION USING JRTK

Hua Yu, Cortis Clark, Robert Malkin, Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA
Email: {hyu,cortis,malkin,ahw}@cs.cmu.edu

ABSTRACT

In this paper we describe our early exploration of automatic recognition of conversational speech in meetings for use in automatic summarizers and browsers to produce meeting minutes effectively and rapidly. To achieve optimal performance we started from two different baseline English recognizers adapted to meeting conditions and tested resulting performance. The data were found to be highly disfluent (conversational human to human speech), noisy (due to lapel microphones and environment), and overlapped with background noise, resulting in error rates comparable so far to those on the CallHome conversational database (40-50% WER). A meeting browser is presented that allows the user to search and skim through highlights from a meeting efficiently despite the recognition errors.

1. INTRODUCTION

Meetings, seminars, lectures and discussions represent verbal forms of information exchange that frequently need to be retrieved and reviewed later on. Human-produced minutes typically provide a means for such retrieval, but are costly to produce and tend to be distorted by the personal bias of the minute taker or reporter. To allow for rapid access to the main points and positions in human conversational discussions and presentations we are developing a meeting browser which records, transcribes and compiles highlights from a meeting or discussion into a condensed summary. The early experiments described here report on the particular problem of recognizing conversational speech in meetings and on the user interface of a meeting browser for later presentation.

We have recorded discussions of three or more participants. To minimize interference with normal styles of speech, we have ruled out the use of close talking microphones and recorded meetings with lapel microphones on two or more speakers. The resulting speech was found to be highly disfluent, similar to spoken telephone conversations as in the Switchboard and CallHome databases, and include many rare words and/or unusual language. The signal quality is further degraded by crosstalk between speakers and reverberation and echo due to the use of the omnidirectional lapel microphones.

2. MEETING TRANSCRIPTION EXPERIMENTS

The experiments we designed were intended to show which of our existing speech recognition systems is best suited to the meeting transcription task. We discuss the data used for testing in Section 2.1. Section 2.2 contains details on the systems tested, and Section 2.3 details the results of our experiments.

2.1. Testing Data

The data used for our experiments are collected during internal group meetings. We gave lapel microphones to three of ten speakers, and recorded the signals on those three channels. Each meeting was approximately one hour in length, for a total of three hours of speech on which to adapt and test.

The microphones were given to two female speakers (fdmg and flsl) and one male speaker (maxl). Since the microphones were not unidirectional, there was significant channel mixing. Thus, we calculate word error based only on the words spoken by the owner of the channel; that is, we test and evaluate only those sections where the channel owner is speaking.

2.2. System Specifications

Two recognizers were used in the experiment; a dictation system (*WSJ*) and a spontaneous speech system (*ESST*) [2], [4], both built using the Janus Recognition Toolkit [1]. Incorporated into our continuous HMM system are techniques such as linear discriminant analysis (*LDA*) for feature space dimension reduction, vocal tract length normalization (*VTLN*) for speaker normalization, cepstral mean normalization (*CMN*) for channel normalization, and wide-context phone modeling (polyphone modeling). The main acoustic

features used in the ESST system are 24-order plp coefficients; in the WSJ system, 48-order mel-spectra are used. Table 1 shows various distinguishing features of these systems.

Feature	*System*	
	ESST	WSJ
front end	PLP (24)	mel-spectra (48)
speech style	spontaneous	read
training data	26.5 hrs	83 hrs
#codebooks	1500	3000
#distributions	7000	3000
WER	20%	9%
WER test set	ESST test	1994 official WSJ test

Table 1: Distinguishing system features

2.2.1. Language Modeling

Language modeling for the meeting domain is difficult due to the extreme lack of definition in the task. Meetings between humans can vary widely in topic, so rather than try to collect data with which to build language models, we relied on existing language models from various tasks. Specifically, we used language models built from 2.7 million words from the Broadcast News corpus (BN), 2.9 million words from the Switchboard corpus (SWB), and 300 thousand words from our English Spontaneous Scheduling Task (ESST) corpus. We tested these language models on the meeting transcripts; perplexity results are shown in Table 2.

The unusually high perplexities for the models trained on BN and WSJ data are due to noise words, which constitute 16% of our testing tokens. The BN and WSJ models, which are poor in noise words, thus make a significant number of predictions after backing off to the unigram distribution, causing high perplexity. Another contributing factor is the presence of many false starts and interruptions in the test data, a characteristic feature of the meeting domain.

We decided to use the SWB language model for our experiment, due to it's lower perplexity on the test data. We lacked enough data to create a reasonably sized cross-validation set, ruling out the possibility of using interpolated models. In our first experiments, we also used a closed vocabulary. That is, every word in the target transcript was included in the system vocabulary.

Corpus	*Size (words)*	*Perplexity*	*Conversational*
BN	2.7MW	915.2	No
WSJ	2.1MW	1257.0	No
SWB	2.9MW	171.2	Yes
ESST	300KW	246.1	Yes

Table 2: Language Model Perplexities on Meeting Data

2.2.2. Acoustic Modeling

The primary task for acoustic modeling in this experiment is the adaptation of existing acoustic models to the test data. Both recognizers were trained on speech recorded in a clean environment with close-talking microphones; the environment in which the testing data were collected represents a significant mismatch for these models. We thus employed VTLN and CMN to adapt the signal (for speaker and channel, respectively), and Maximum Likelihood Linear Regression (*MLLR*) [5] to adapt the model. Our adaptation strategy is defined below.

1. *MLLR* MLLR is widely known as an effective technique for adaptation. In our system, we employed a regression tree, constructed using the acoustic similarity criterion, to define regression classes. The tree is pruned to a degree which allows for each leaf to have sufficient adaptation data. For each leaf node we calculate a linear transformation in order to maximize the likelihood of the adaptation data. Thus, the number of transformations is determined automatically.

2. *Iterative batch-mode unsupervised adaptation* The quality of adaptation depends directly on the quality of the hypotheses on which the alignments are based. We thus iterate the adaptation procedure, incrementally improving both the acoustic models and the hypotheses they produce at each iteration. We found significant gains in the first and second iterations, after which the gains reach an asymptote.

3. *Adaptation with confidence measures* Confidence measures were used to automatically select the best candidates for adaptation. Our method was based on lattice rescoring. If, in rescoring the lattice with a variety of language model weights and insertion penalties, a word appears in every possible top-1 hypothesis, acoustic stability is indicated. Such acoustic stability often identifies a good candidate for adaptation. Using only these words in the adaptation procedure produces 1-2% gains in word accuracy over blind adaptation.

4. *Guided Adaptation* We also performed guided adaptation experiments for the ESST recognizer. That is, we adapted using transcripts instead of decoder output. This result shows how much gain in word accuracy is possible using MLLR. These results are summarized in Table 5.

5. *CMN* As noted in [3], there are many ways to employ CMN. We found that global CMN (CMN over a set of utterances) outperforms using CMN on a per-utterance basis. This is probably an effect of data fragmentation in the per-utterance case.

Speaker	*Adaptation Iterations*			
	0	*1*	*2*	*Adaptation Gain*
maxl	37.0	48.2	51.3	22%
fdmg	40.7	46.9	49.7	15%
flsl	20.8	32.3	33.3	16%
Total	32.6	42.5	44.8	

Table 3: ESST word accuracy rates

Speaker	*Adaptation Iterations*			
	0	*1*	*2*	*Adaptation Gain*
maxl	48.3	54.7	54.8	12%
fdmg	51.6	56.2	55.1	9%
flsl	36.2	40.5	40.4	7%
Total	45.2	50.4	50.1	

Table 4: WSJ word accuracy rates

2.3. Results

As expected, MLLR yields considerable improvements for both the ESST system (Table 3) and the WSJ system (Table 4). This adaptation technique allowed two very different acoustic models to attain comparable performance on the testing data. This result underscores the importance of employing sound adaptation techniques for recognizers expected to function in a variety of acoustic environments, as would be the case for a meeting recognition system.

We were somewhat surprised that the WSJ system outperformed the ESST system. We felt that since the ESST system had a much better match in speaking style to the target domain, it should attain higher accuracy than the WSJ system, trained on read speech. The WSJ system, though, is much stronger than the ESST system in one key area. It was trained on over three times as much speech data, resulting in better polyphone coverage. Thus, words like "Japanese," "recognition," and "analysis" often were mishypothesized by the ESST system as strings of shorter, similar-sounding words.

3. THE MEETING BROWSER INTERFACE

As noted in Section 1 above, we also require an interface with which to view and browse transcribed meetings. The interface we have created for this task is our Meeting Browser system, pictured in Figure 1.

Speaker	*Supervised Adaptation*
maxl	62.1
fdmg	61.0
flsl	48.3
Total	57.2

Table 5: ESST word accuracy with Supervised Adaptation

The Meeting Browser interface displays meeting transcriptions, time-aligned to the corresponding sound files. The user can select all or a portion of these sound files for playback; text highlighting occurs in sync with the sound playback.

The Meeting Browser is built around information streams. Transcribed meeting text is just one such stream; the interface can accept streams from virtually any source which produces text output. These streams are fully editable and searchable, allowing humans to annotate and correct recognizer output as well as add new streams manually.

Since ultimately, the usefulness of a meeting transcription system is bounded by the usability of the interface, we feel that the flexibility present in the Meeting Browser is extremely important in user acceptance of the meeting recording and transcription process.

4. CONCLUSIONS AND FUTURE WORK

We have described our preliminary experiments in automatic meeting transcription as well as the interface we have designed for viewing and browsing transcripts. Early transcription experiments have underscored the importance of quality adaptation methods, and have shown that using existing acoustic models for new tasks is not an unreasonable course of action.

While MLLR helps a great deal in overcoming mismatches between training and testing data, there is still apparently a significant gap between unsupervised and supervised adaptation. Our simple confidence metric helps somewhat; it is possible that more sophisticated methods for confidence annotation will further increase the efficacy of unsupervised adaptation.

Further, we noted that our simple energy-based segmentation method tended to cause overfragmented data, leading to recognition errors at the beginning and end of utterances. Improving our segmentation methods should result in increased word accuracy rates.

Future work in meeting transcription will incorporate new methods to deal with these problems, as well as an expansion from meeting transcription to general meeting tracking and summarization, hopefully without the need for lapel microphones. We plan to combine the many sources of information present in a meeting setting, including speaker localization and channel separation using microphone arrays; face and gaze tracking to model who is speaking to whom; lip reading to aid speech recognition; and automatic summarization procedures, in order to produce an accurate summary of the events of a meeting with minimal human effort or supervision.

All of this information can be included in the streams

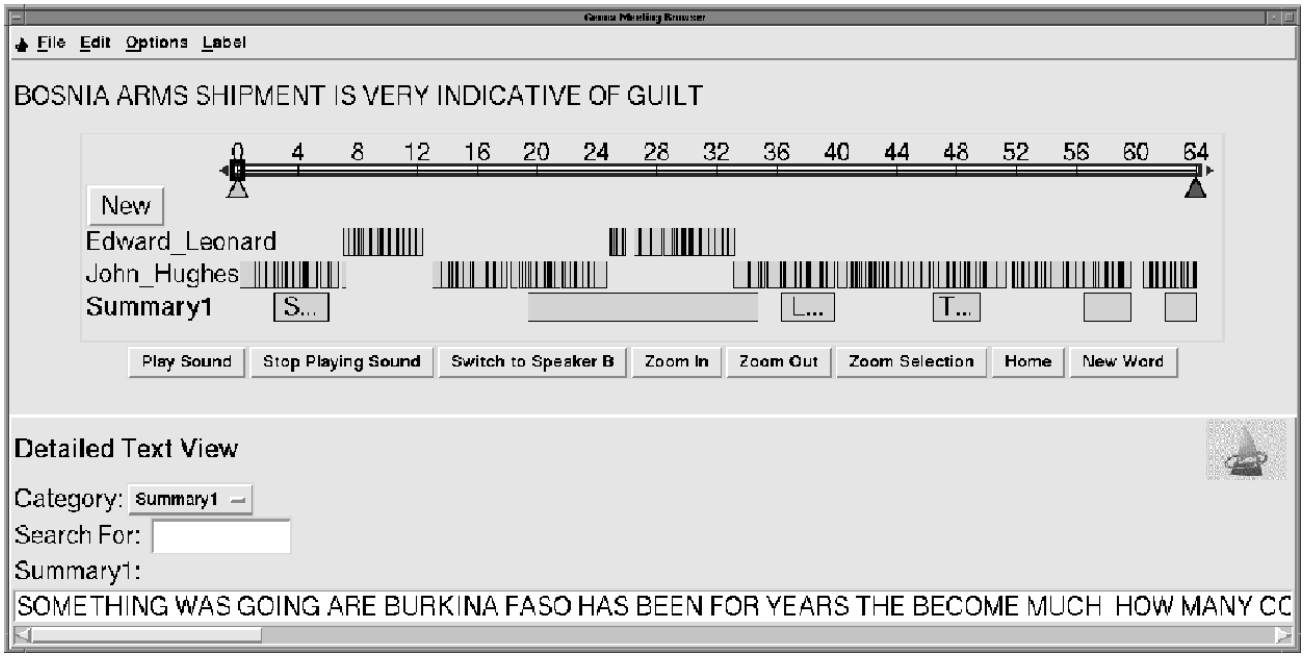

Figure 1: The Meeting Browser Interface

passed to the Meeting Browser interface. This interface is being extended in numerous ways to increase usability and user acceptance, including security features to restrict access to portions of some streams and incorporating multimodal repair facilities [6] into the interface. We are also exploring ways to produce and include information describing the topical and discourse structure of a meeting, as well as multimedia presentations of such structures.

5. ACKNOWLEDGEMENTS

This research is sponsored by the Defense Advanced Research Projects Agency under the Genoa project, subcontracted through the ISX Corporation under Contract No. P097047. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, ISX or any other party.

6. REFERENCES

[1] Finke, Michael, et.al. “The JanusRTk Switchboard/Callhome 1997 Evaluation System”. *Proceedings of the LVCSR Hub5-e Workshop*, Baltimore, USA, 1997.

[2] Zeppenfeld, Torsten, et. al. “Recognition of Conversational Telephone Speech Using the Janus Speech Engine”. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Germany, 1997.

[3] Westphal, Martin. “The Use of Cepstral Means in Conversational Speech Recognition”. *Proceedings of Eurospeech Conference*, Greece, 1997.

[4] Zhan, Puming. “Speaker Normalization and Speaker Adaptation - a Combination for Conversational Speech Recognition”. *Proceedings of Eurospeech Conference*, Greece, 1997.

[5] Leggetter, C.J., and Woolland, P.C. “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs”. *Computer Speech and Language*, 9:171-186, 1995.

[6] Suhm, Bernhard, et. al. “Interactive Recovery from Speech Recognition Errors in Speech User Interfaces”. *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.