NEW DEVELOPMENTS IN AUTOMATIC MEETING TRANSCRIPTION

Hua Yu, Takashi Tomokiyo, Zhirong Wang and Alex Waibel

Interactive Systems Lab, Carnegie Mellon University, Pittsburgh, PA 15213
Email: {hyu, tomokiyo+, zhirong, ahw}@cs.cmu.edu

ABSTRACT

In this paper we report on new developments in the automatic meeting transcription task. Unlike other types of speech (such as those found in Broadcast News and Switchboard), meetings are unique in their richer dynamics of human-to-human interaction. An intuitive "fingernail" plot is proposed to visualize such turn-taking behavior. We will also show how recognition of short turns can be improved by building a language model tailored specifically for short turns. Out-Of-Vocabulary (OOV) words become a more salient problem in the meeting transcription task, as they are mostly topic words and proper names, lack of which not only causes Word Error Rate (WER) increase, but also limits further use of recognition hypotheses. We describe a prototype system which uses the Web as a source for vocabulary expansion, and present preliminary OOV retrieval results.

1. INTRODUCTION

As speech recognition research progresses from read speech (Wall Street Journal), to prepared speech (a major part of Broadcast News), to more natural forms of human interaction (Switchboard and meetings), every time we see exciting new problems arising, as well as re-surfacing of old problems. The challenge of meeting recognition is multi-fold. In previous works [11, 12], we have identified major issues such as:

- degraded recording conditions: the possibility of using a single omni-directional table microphone is left to future exploration. We chose clip-on lapel microphones over close-talking headsets to minimize interference with normal styles of speech. However, this comes with the cost of significant channel mixing.
- spontaneous speech style: cited the number one challenge in the Switchboard task, this imposes the same, if not greater, amount of difficulty in meetings.
- lack of training data: as meeting data collection is gaining steam in a number of locations, the situation is improving rapidly. However, we would like to argue that there will never be enough training data in an open domain scenario. Restricting meetings to, say, business meetings in the tobacco industry, will only lead to yet another domain-specific system.

In this paper, we focus on one of the new aspects in the meeting task — turn-taking behavior, and one of the old problems — dealing with OOV words.

The most distinguishable aspect of meetings is probably the dynamics of human interaction. In particular, turn-taking behavior can reveal a lot about a meeting. In Section 3 we give a preliminary study of this subject, and show how recognition can benefit from knowing the duration of turns.

OOV is a long existing problem coming from the fact that recognizers can recognize only a fixed vocabulary. In the past, we can get away by fixing the task domain and matching training and testing material. While under the open domain assumption, this issue persists regardless of how much the vocabulary size increases. It only becomes more pressing in meetings because most of the OOV words are either important topic specific words, or important proper names. Without them little sense can be made of recognition hypotheses. In Section 4 we present a method of dynamic vocabulary adaptation using the Web and experimental results.

The rest of the paper is organized as follows: Section 2 describes meeting data used for the experiments and its collection. Recognition system setup and results are presented in Section 5.

2. MEETING DATA

We recorded a number of internal group meetings (about 1 hour long each), of which 6 are used for experiments in this paper. Clip-on lapel microphones rather than headset are used to avoid being obtrusive. Different from earlier recordings, we are now using a multi-channel sound card which supports simultaneous recording of 8 speakers. This simplifies the recording setup. It also eliminates the need of synchronizing multiple channels. However, the microphone mixup problem still exists as before. Quite often one can hear multiple voices in a single channel.

The test set consists of two meetings in their entireties, and a randomly chosen 10% portion each for the rest of the meetings (roughly 30,000 word tokens).

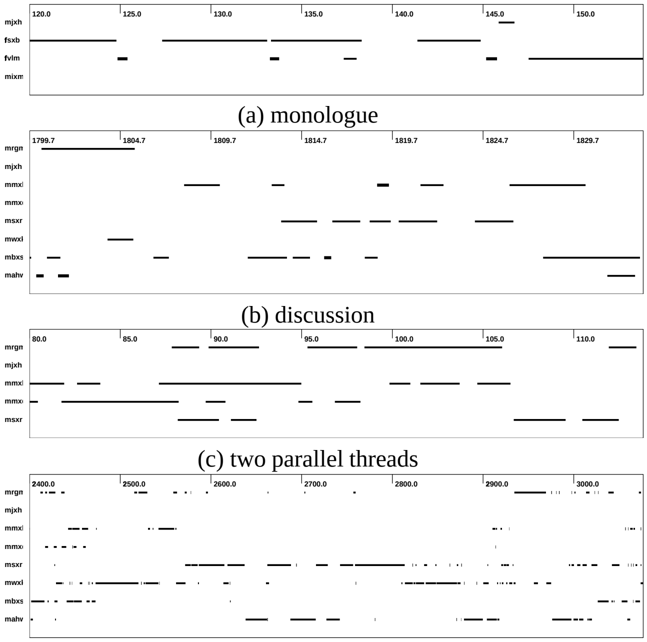
3. SHORT-TURN LANGUAGE MODELING

3.1. Turn-Taking Behavior

Interestingly, even without knowing the words spoken, one can already tell a lot about the dynamics of a meeting by just looking at the "who speaks when" results. For example, one can identify the dominant speaker(s), as well as the type of the meeting (whether it is a discussion or presentation).

In Figure 1(a), we can see that out of four participants, two are actively involved during the time period shown. While one is doing most of the talking, another is paying close attention, giving brief feedback along the way; then without much overlapping with speaker 1, takes a turn for about 10 seconds. We can assume it's more likely an answer to a question, rather than jumping in to air his/her own opinion, due to the lack of crosstalk.

Figure 1(b) looks more "random": many short turns, quite some overlapping, but not so heavily. It's a group discussion. We contrast it with the the third, where one can observe more "randomness": two or more people talking simultaneously for as long as 20 seconds. This is indeed pure chatting at the beginning



(a) monologue

(b) discussion

(c) two parallel threads

(d) “fingernail” plot of long term speaker activities

Figure 1: Turn Taking Plots
(Time marks at the top of each plot are in seconds.
Speaker names are shown to the left of each plot.)

of a meeting, with two independent threads going on the same time around the table.

The last plot is on a much larger time scale, showing speaker activities over 800 seconds (13 minutes). As expected, little overlapping is observed at this scale. It’s easy to identify both the major speakers and the “minor” speakers. Questions regarding people’s different interests at a meeting can also be answered. Different phases of a meeting are also clearly identifiable, such as where transition take place and when discussions happen.

The “fingernail” turn-taking plot is very useful in providing an intuitive overview of a meeting. The dynamics of a meeting along the time line is immediately available, which can be very hard to obtain from reading transcripts.[1] The simple plot shown here can be further enhanced with richer annotations, such as dialogue act and emotion [2].

3.2. Short-Turn Language Modeling

Short turns make up a significant portion in the meetings. In our case a little more than 30% of all turns have a duration less than 0.7 seconds. They provide further cues for understanding meeting dynamics. They are usually brief feedback that a listener gives to a speaker, such as “right,” “yes,” “mhm,” “no,” “I don’t think so.” Short turns carry the important information of whether or not the listener agrees with the speaker. Due to the short and isolated nature of such events, recognition accuracy is worse than average. However, analysis reveals short turns have a very constrained grammar.

In our experiments, short turns are defined as those with duration less than 0.7 seconds. Table 1 lists the most frequently seen turns sorted by frequency.

	Frequency		Frequency
yeah	0.304	cool	0.004
mhm	0.239	ah	0.004
okay	0.066	what	0.004
right	0.053	so	0.004
yes	0.026	good	0.004
oh	0.021	but	0.004
no	0.019	yep	0.003
hm	0.008	sorry	0.003
huh	0.006	i don’t know	0.003
yeah +noise+	0.006	+noise+ okay	0.003
+noise+ yeah	0.006	you know	0.002
well	0.005	who	0.002
really	0.005	the	0.002
oh yeah	0.005	that’s right	0.002
exactly	0.005	no no	0.002
+noise+ +noise+	0.005	i know	0.002
oh okay	0.004	i	0.002

Table 1: Most Frequent Short Turns

By building a language model tailored specifically for short turns, we expect to see a drop in word error rate. Table 2 shows recognition results on short-turn data with different language models Three different language models are tried: a regular trigram model trained on the Broadcast News corpus (BN), a smaller model trained on short turns extracted from Switchboard+Callhome+Meeting data (SCM-S), and a very small language model trained on short turns in the meeting data only (M-S). Details of the recognition system can be found in Section 5.

Condition	LM Training	Perplexity	WER(%)
First pass	BN	35.1	51.1
First pass	SCM-S	8.7	43.3
First pass	M-S	8.2	40.8
MLLR Adapted	M-S	8.2	38.9

Table 2: WER(%) on Short Turns (Lexicon: 20k, OOV=1.04%)

Of course for real recognition, one need to identify short turns beforehand. With the setup where each speaker wears a separate microphone, this may not be difficult as we can segment the audio into turns by simply comparing loudness across channels.

4. DYNAMIC VOCABULARY ADAPTATION

Statistics from the meeting corpus reveals that although OOV token rate is in a reasonable range of 1.1%–2.0%, OOV type rate is as high as 4.2%–8.0%. As most function words are already present in the vocabulary, it’s not difficult to imagine that most[2] OOV words are either novel keywords, such as technical terms and jargons, or proper names, such as names of participants. These are the last words one would like to be missing. It not only causes a significant WER increase, but also limits the potential of any post-processing of recognition hypotheses (such as keyword

[1] Reading a meeting transcript is found to be very hard and time-consuming, giving all the crosstalk and disfluencies. How to produce a meaningful transcript is definitely another major challenge for the meeting task.

[2] Our informal estimation is above 90%

search or summarization). To ameliorate this problem, we use the Web as a backend knowledge source for dynamically expanding the recognition vocabulary.

The architecture of the prototype system is shown in Figure 2. The first pass hypothesis from a speech recognizer is filtered and topic words are extracted. The extracted words are then passed as a query to a Web search engine. From the retrieved documents, we extract a list of words that are not previously in the vocabulary and generate pronunciations for them. Currently the Festival speech synthesis system is used to give the pronunciation [3]. The entire process is configured to run automatically.

The idea of using the Internet as a potential source for language modeling is not new. Berger et al. has explored the possibility of using the Web for dynamic language model adaptation [1]. Here we perceive the OOV problem as a major barrier, for the simple reason that little can be done if a word is not known to the recognizer.

Our approach differs from that in [7], where a large amount of text is pre-collected, stored in site, and retrieved using algorithms of one's choice. Here we have to rely on the Web and existing search engines. This is again part of the open-domain challenge: one couldn't afford to store a corpus having the same coverage as the Web itself, especially since the Web is being updated constantly.

4.1. Search Strategy

As stated before, OOV words in the meeting data fall under two categories: topical words of the meeting, and proper nouns / jargons that are specific to the attendees. Our search strategy, therefore, has two components: generic Web searching and site-specific Web crawling. The first is expected to retrieve topic words, while the second will recover person/place/project names, as well as local jargons.

4.1.1. Generic Web Searching

Our task is quite different from what most search engines are designed for. First, we are not interested in pinpointing a specific Web page, but rather a number of pages pertaining to a certain topic. Second, queries must be generated automatically, which usually means the query will be long. However, it should not be too long to be accepted by search engines.

We compared different search engines and chose Infoseek [6] for the following reasons:

- Accepts long queries;
- Doesn't require exact match, i.e. the engine always returns some documents even if no exact match is found;
- Sorts result by reasonable relevance score.

To extract words relevant to the meeting topic, we use the mutual information statistic:

$$MI(w_i, t_j) = p(w_i, t_j) \log \frac{p(w_i, t_j)}{p(w_i)p(t_j)}$$

where w_i is a certain word and t_j is one of the topics. $p(w_i, t_j)$ is the joint probability, $p(w_i)$ and $p(t_j)$ are the respective marginals. This statistic is the element of the mutual information between random variables W and T:

$$I(W,T) = \sum_{w_i \in W} \sum_{t_j \in T} [p(w_i, t_j) \log \frac{p(w_i, t_j)}{p(w_i)p(t_j)}]$$

This mutual-information-based metric is widely used in topic classification [10]. It is also called *information gain* to distinguish it from *pointwise mutual information*[9, 4, 10], which is not normalized by $p(w_i, t_j)$.

The mutual information criterion is very successful in selecting keywords. Examples of the top 20 words extracted for two of the meetings (m010, m013) are:

units **mobility** map show **coa multi-modal**
visualization **queries** talking point
scenario **blob** screen display table unit
chart attributes **visualizations** morale

data scenario **cpof** maya demos **greybeards**
exercise **palmpilot** visualization **recognizer**
demo grammar handwriting show probably
conops ward map problem **dat-recorders**

respectively, where boldface indicates that word, or one of its inflections, is out of vocabulary. The set of selected words usually gives a good clue about the meeting topic: one can quickly guess that m010 is a project meeting about mutli-modal visualization for some military situation.

To train topic-independent word distributions, an 80 MB corpus from the Topic Detection and Tracking (TDT) project is also used, where many documents are annotated with topic labels.

During testing, every meeting is assumed to have a topic of its own. We extract the top N words according to the mutual information criterion (N=20 in our cases), and use them as the query.

4.1.2. Site-Specific Web Crawling

A Web-crawler is implemented to collect Web pages in a certain domain. For our experiment, we started from our group homepage (http://www.is.cs.cmu.edu) and collected all pages under a depth of 8. Links that lead to a page in other domains are not followed. Words occuring more than 3 times are collected.

The crawler has the following added features:

- Extracting text from Postscript and PDF files. Thus papers and technical reports can be effectively used;
- Language classification: every word in a foreign page is an OOV word to an English dictionary, and we did encounter a lot of foreign pages, unfortunately. A naive Bayes language classifier is created for this purpose, which uses character-based 4-gram model with Kneser-Ney smoothing[8]. The classifier works fairly well in excluding foreign pages.

4.2. Evaluation on OOV Retrieval

Preliminary experiments are promising. Retrieval results for meeting m010 are summarized in Table 3. The first two rows are cheating experiments. First, we tried to use the transcript as input for topic words extraction (therefore the query might actually contain some of the OOV words.) The retrieved documents (top 20 pages) contains 13% of the OOV types, about 26% of all OOV tokens. But when excluding OOV words from the transcript in the beginning, the gain virtually vanishes. Amazingly, hypothesis did extremely well: even with a WER above 40%, the OOV rate reduction is as good as that using the transcript. Site-specific Web crawling provides another major gain on top of Web searching. Combining the two, we have recovered 60% of the OOV words.

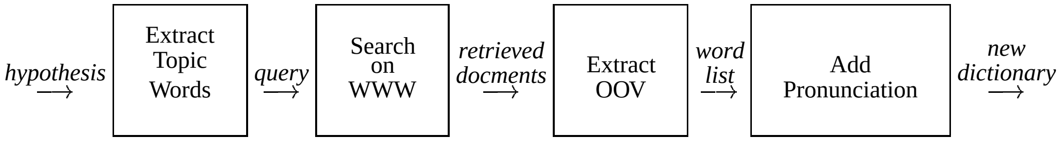


Figure 2: A Prototype System for Dynamic Dictionary Adaptation

Input	OOV recall (%)		new OOV rate (%)
	types	tokens	(start at 1.06)
transcript (with OOV)	13.16	26.14	0.78
transcript (excluding OOV)	2.63	1.14	1.04
hypothesis	15.79	25.57	0.79
site crawling	44.74	42.61	0.61
hypo+site	52.63	61.93	0.40

Table 3: OOV Retrieval Results on m010 (176 OOV words, 38 unique)

However, the Web retrieval performance is not always so favorable, the same as our daily experience with search engines. On the other hand, site-specific retrieval performance is quite stable.

5. RECOGNITION EXPERIMENTS

Our recognizer is trained on Broadcast News using JRTk [5]. The baseline system deploys a quinphone model with 6000 distributions sharing 2000 codebooks. There are about 105k Gaussians in the system. Vocal Tract Length Normalization(VTLN), cluster-based Cepstral Mean Normalization(CMN), and a 7-frame context window are used. LDA (Linear Discriminant Analysis) is applied to reduce feature dimensionality to 42, followed by an optional diagonalization transform (also called MLLT, Maximum Likelihood Linear Transform). A 40k vocabulary and trigram language model are used. The baseline language model is trained on the Broadcast News corpus.

While the recognition system achieves a first pass WER of 19.0% on all F-conditions of Broadcast News task, the WER on meeting data still remains quite high. We tried various training / adaptation / system-voting techniques, as well as language model interpolation with Switchboard/CallHome and a limited amount of meeting data. The best WER is so far 40.4%.

Stage	WER(%)
First pass (best)	43.1
MLLR Adapted	41.8
Rover of 5 hypotheses	40.4

Table 4: Meeting Recognition Results

6. CONCLUSION

We have presented our new developments in automatic meeting transcription. Obviously there are more to be deserved: automatic segmentation of meetings into turns, improving OOV retrieval performance, as well as adding Just-In-Time language modeling and ultimately conducting recognition experiments with all the added features.

7. ACKNOWLEDGEMENTS

We would like to thank Klaus Ries and other members of the Interactive Systems Labs for helpful discussions and active support.

8. REFERENCES

[1] Adam Berger and Robert Miller. Just-In-Time language modelling. In *Proc. ICASSP98*, 1998.

[2] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. Multimodal meeting tracker. In *Proc. RIAO2000*, 2000.

[3] Alan Black and Paul Taylor. Festival speech synthesis system: System documentation. Technical Report HCRC/TR-83, University of Edinburgh, 1997.

[4] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information and lexicography. In *Proceedings of ACL 27*, pages 76–83, Vancouver, Canada, 1989.

[5] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld. The JanusRTk Switchboard/Callhome 1997 evaluation system. In *Proceedings of LVCSR Hub5-e Workshop*, 1997.

[6] http://www.infoseek.com.

[7] Thomas Kemp and Alex Waibel. Reducing the oov rate in broadcast news speech recognition. In *Proc. ICSLP98*, 1998.

[8] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proc. ICASSP95*, 1995.

[9] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[10] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann, 1997.

[11] Hua Yu, Cortis Clark, Robert Malkin, and Alex Waibel. Experiments in automatic meeting transcription using JRTk. In *Proc. ICASSP98*, 1998.

[12] Hua Yu, Michael Finke, and Alex Waibel. Progress in automatic meeting transcription. In *Proc. Eurospeech99*, 1999.