PROGRESS IN AUTOMATIC MEETING TRANSCRIPTION

Hua Yu, Michael Finke, Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: {hyu, finkem, ahw}@cs.cmu.edu

ABSTRACT

In this paper we report recent developments on the meeting transcription task, a large vocabulary conversational speech recognition task. Previous experiments showed this is a very challenging task, with about 50% word error rate (WER) using existing recognizers. The difficulty mostly comes from highly disfluent/conversational nature of meetings, and lack of domain specific training data. For the first problem, our SWB(Switchboard) system — a conversational telephone speech recognizer — was used to recognize wide-band meeting data; for the latter, we leveraged the large amount of Broadcast News (BN) data to build a robust system. This paper will especially focus on two experiments in the BN system development: model combination and HMM topology/duration modeling. Model combination can be done at various stages of recognition: post-processing schemes such as ROVER can lead to significant improvements; to reduce computation we tried model combination at acoustic score level. We will also show the importance of temporal constraints in decoding, present some HMM topology/duration modeling experiments. Finally, the meeting browser system and meeting room setup will be reviewed.

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. This task facilitates research in both speech recognition and automatic summarization / information extraction, as well as discourse modeling, because of the highly interactive nature of meetings.

Among all possible meeting scenarios, we especially targeted two different types: research group meetings and discussion-type TV news shows, where a host and several guests hold a discussion of current events. We collected several internal group meetings, recorded with lapel and some stand-microphones, as well as 27 hours of TV news shows (18 hours of Newshour, 9 hours of Crossfire) recorded directly from a TV set.

Our previous experiments [8], mostly on the group meeting data, showed it's quite challenging: we achieved about 50% WER with our WSJ(Wall Street Journal) system and ESST (English Spontaneous Scheduling Task) system, even after iterative unsupervised adaptation. The difficulty mostly comes from:

- speaking style mismatch: conversational, highly disfluent, cross-talk
- microphone mismatch and lack of domain-specific training data

In this paper, we first describe an experiment using the SWB system on the same data (Section 2), to address the speaking style mismatch. Section 3 presents the development of our BN system, which we hope can provide the required robustness for the meeting task. Along the way we will cover two interesting experiments in greater detail: model/system combination and HMM topology/duration modeling. Finally, the meeting browser interface and meeting room setup will be briefly reviewed.

2. EXPERIMENTS WITH SWB SYSTEM

As noted above, to account for the mismatch in speaking style, we used our SWB (Switchboard) system, one of the best performing systems in the 1997 Hub5 Evaluation. An interesting point is that by doing so, we introduced another type of mismatch: SWB is trained on 8KHz telephone speech, while the meeting data is 16KHz wideband speech. Thus no one would risk a prediction about the outcome when we started out.

We downsampled meeting data to 8KHz, under the risk of losing information contained in the higher frequency band, then fed it to the SWB recognizer. To our surprise, the result was an after-adaptation WER of 40%. This was better than both the WSJ and the ESST system (Table 1). (The vocabulary and language model are unmodified SWB models, with about 15k words, OOV rate is 3%.)

# Adapt Iterations	0	1	2
ESST[1]	67.4	57.5	55.2
WSJ[1]	54.8	49.6	49.9
SWB	47.0	42.3	41.6

Table 1: WER(%) on the group meeting data

We attribute this success to the matching of speech style. The conversational speech style is modeled in several components of the SWB system: acoustic model, language model and especially pronunciation lexicon. The latter modeled frequent pronunciation variants & common contractions probabilistically [9]. On average, it has about 2 pronunciation variants per word; and all frequent "phrases" like "KIND OF", "SORT OF", "AND A" are represented as compound words, in order both to give them accurate pronunciations and to benefit from having longer baseforms in decoding.

While it may not be easy to single out the component that contributed most, we tried a simple experiment to port the lexicon to WSJ system and achieved some success.

3. BN SYSTEM DEVELOPMENT

TV news shows (such as Newshour and Crossfire), while conversational in style, also bear some resemblance to broadcast news data, both in terms of wide topic coverage, and higher recording quality. Experiments with existing recognizers (WSJ/SWB) didn't give us a satisfactory result on this data. We decided to leverage the large amount of training data available in the Hub4 task, to build a robust recognizer for the meeting task.

Bootstrapping from the WSJ system, we followed the "standard" Janus training steps, and trained a relatively small VTL-normalized triphone system. It's roughly a semi-tied system, with 6000 distributions sharing 3000 codebooks, with each codebook having 24/32 Gaussians. We tried several frontends: standard cepstrum, mel-spectrum, and a slight variation of cepstrum. They all end up with a 42-dimensional feature after applying LDA. A simple word trigram language model is used throughout the experiments. We also compute a confidence score based on acoustic stability of the hypothesized word (by counting how many times it shows up when rescoring lattice with various language weights and word insertion penalties). The results reported here are first pass numbers (unless otherwise stated) on the Hub4 1996 development PE (partitioned evaluation) set, with a 20k vocabulary (OOV rate 2.1%).

[1] previous results, cf. [8]

3.1. Model Combination

Consistent with results in Hub4 evaluation, WER can be significantly improved by combining different systems using ROVER ([10]). In our case, WER is reduced from 35.0% to 32.0%, about 9% relative gain, by combining 4 systems retrained with different frontends mentioned above. More amazingly, in the "oracle" case, i.e. if we can combine the 4 different hypothesis optimally, the WER dropped to 23.4%. This is very attractive, while at the same time seems quite achievable since only a 4-way choice is involved at each word position.

The obvious starting point was to improve the voting scheme. ROVER is able to do simple majority voting, or use side-information like a confidence score for each word. Our best result was obtained by using average confidence scores. We can also envision more elaborate schemes such as using N-best hypothesis from each system to populate the voting pool.

Combining systems at the post-processing stage (the way ROVER does) can be costly since we need to run 4 decoding/rescoring passes. A desirable goal is to have a single system and to run a single decoding pass. To this end we tried a simple model combination approach. Since the 4 systems only differ in their frontends (and of course, acoustic models), and Janus has built-in support for combining acoustic scores from multiple streams, we combined acoustic scores from different systems in a linear fashion:

$$Overall\_acoustic\_score = \sum_{all\ streams} (w_i * Acoustic\_score_i)$$

where w_i is the weight for the i-th stream. Since the scores are in the log domain, this is essentially the same as the log-linear model combination approach in [1].

In the experiment below we combined 2 systems, with WER of 32.3% and 34.2% respectively. The model combination approach gives 31.3%, while ROVER with confidence measurement gives 30.8%, out of the perfect-ROVER WER of 24.9% (Table 2).

System	WER(%)
scB	32.3
MSPEC	34.2
ROVER(Oracle)	24.8
ROVER with CM	30.8
2-stream combination	31.3

Table 2: Model Combination Experiment. scB and MSPEC denotes 2 acoustic models with different frontends. Weights for each model are empirically determined. CM stands for Confidence Measurement. WER is measured on a subset of dev96pe data.

Thus model combination at the acoustic score level didn't outperform ROVER – model combination at the post-processing phase. We feel there's much more to explore: what's the exact nature of the between-system differences (are they really different or is it just some random noise due to perturbation), how to effectively combine them, etc. In the literature, Hazen[2] suggested that aggregation can be used to improve classifiers; Peskin[3] noted "jiggling" in adaptation can also smooth out different models; also another common observation with ROVER is that the more diverse participating systems are, the more win ROVER can provide. We believe an in-depth analysis is essential to a correct understanding of some speech techniques and can lead to better and more robust systems, since this issue is widespread in recognizer development and evaluation.

3.2. HMM Topology & Duration Modeling

Another interesting episode in BN system development is about HMM topology. So far 3-state left to right topology is the most commonly used for a phoneme, with each state having a forward transition and a self-loop. For some reason our initial topology allows very fast skipping: it can transit to the next phoneme directly from the beginning state (or the middle state). After switching back to a more conservative topology, which only allows skipping the end state, the WER went down 2% absolute. This is the result after retraining. Without retraining, i.e. simply decoding using the original acoustic model but with the "corrected" topology, we still get 1.5% absolute gain (Table 3).

System	WER (%)
Old Topo	34.3
New Topo (retrained)	32.1
New Topo (without retraining)	32.7
Minimum Duration Modeling	31.8

Table 3: Topology Experiments

Our explanation for the above results is that a more restrictive topology enforces a certain trajectory that a phoneme must go through, without which decoding could become too flexible and easily confused. But because training (forced alignment) is guided by reference texts, thus more restrictive, we didn't have as much problem in training as we would in decoding. Nontheless this posed an interesting dilemma: why did the unmatched testing setup turn out better than the matched case? The old topology basically subsumes the "correct" topology, thus it has higher training set likelihood than the "correct" counterpart. It seems that there're some important factors largely unaccounted for in the traditional framework.

While we currently don't know how to pursue the topology argument further, we suspect duration might be a factor there: it can also provide some guidance during decoding (for a smoothed, more reasonable hypothesis). After reviewing some of the earlier duration modeling work ([4, 5, 6]), we decided to take a slightly different approach. Instead of assigning probabilities to all possible durations of a context dependent phoneme, i.e. modelling with a multinomial distribution, we chose to simply enforce a minimum/maximum duration constraint. This prevents the occurance of extremely short or long phonemes commonly seen in recognition errors, and has several advantages:

- avoids the scaling problem of combining duration score with acoustic score
- allows easy incorporation of the durational constraint into decoding: a phoneme can only exit after consuming a minimum number of frames, and it must exit when the maximum duration is reached.
- simplifies models: only 2 numbers are needed for each triphone: minimum/maximum duration. The hope is to capture 80% of the possible gain with 20% of the effort.

As a first step, we used only the minimum duration constraint. In the training phase we went through the entire training corpus to gather duration information for each triphone. Then a decision tree was grown to cluster all triphones, so that for each leaf node we can robustly estimate a distinct minimum duration. The minimum duration of a leaf node is taken as the n-th percentage cutoff point of its duration distribution/histogram (with n typically being 3 or 5). At decoding time the minimum duration constraint is enforced by using different topologies.

Preliminary experiments didn't post as much gain as we had hoped for. We had a total of 0.3% absolute gain over the "correct" topology by doing minimum duration modeling. In the future we can try more elaborate schemes, for example, making duration models dependent on speech rate.

3.3. Partitioning Strategy

All experiments above are conducted under the partitioned evaluation (PE) scenario: speaker adaptation and VTLN warping factor estimation are all done on a per utterance basis, which is clearly suboptimal. This is only because we don't have a tool to deal with continuous speech stream. Following the Hub4 trend, we implemented the LIMSI style partitioning scheme [7]: first classify incoming data into speech/music/silence category, throw away the non-speech data; do an initial segmentation, with parameter set to over-generating segments; assuming each segment as a cluster by its own, estimate a Gaussian mixture model for each cluster; then iteratively (viterbi) reestimate and cluster these mixture models, until the likelihood penalized by number of clusters and number of segments no longer increases. The result is a segmentation with "speaker" labeling.

Unlike its ad hoc counterparts, the LIMSI approach is quite elegant in that it uses a couple of global parameters to control the whole process. Each of them has a clear interpretation. This partitioning scheme works pretty well for the Newshour data (over 90% in terms of cluster purity and best-cluster coverage of a speaker). We plan to migrate to UE (unpartitioned evaluation) style recognition in the near future.

3.4. Results on TV News Show

Decoding those TV news show data with the BN system gave us much better WER compared to existing recognizers from other domain (WSJ/SWB), as shown in Table 4. Our observation is that Newshour data is fairly well behaved while Crossfire, as its name suggests, involved more heated discussion, crosstalk, and shorter turns. The result for meeting data remains pretty high, with WER in the 40% to 50% range.

Show type	WER(1st pass)	WER after adaptation
Newshour	26.9	26.3
Crossfire	36.0	34.6

Table 4: Decoding the News Show data with the BN System (same 20K vocab, BN language model as before)

4. MEETING BROWSER & MEETING ROOM

To assist efficient reviewing and browsing meetings, recognizer output is fed to an automatic summarizer based on Maximal Marginal Relevance (MMR) criteria, and then streamed into the meeting browser system [11]. The meeting browser interface can display meeting transcriptions, time-aligned to corresponding audio and video data. The user can choose to search, browse, or annotate the meeting.

Other than offline browsing, we're also developing an online meeting room demo, where realtime (or close to realtime) speech recognition, speaker identification, people tracking, people identification, face/gaze tracking, etc. are put together to make a live meeting scenario, so that we know the number of participants, who they are, who's talking (to whom), etc. We hope by extracting and fusing additional cues we can better capture/understand the meeting dynamics and structural information.

5. CONCLUSION

Both results on group meeting data and discussion-type news show data have shown significant improvements in automatic meeting transcription. We've reported preliminary experiments on model combination and HMM duration/topology modeling. As noted before, there're much more to be explored in the future.

6. ACKNOWLEDGEMENT

We would like to thank Rita Singh for providing us with various resources, and all members of Interactive System Lab for their help and support, especially Michael Bett, Takashi Tomokiyo, Donghoon Van Uytsel, Rob Malkin, Yue Pan, and Klaus Zechner. This research is sponsored by the ISX Corpporation under DARPA Subcontract PO97047. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of ISX, DARPA, or the U.S. government.

7. REFERENCES

[1] Peter Beyerlein, Discriminative Model Combination, ASRU97, Santa Barbaba

[2] Timothy J. Hazen, et al. Using Aggregation to Improve the Performance of Mixture Gaussian Acoustic Models, ICASSP98, Seattle

[3] Barbara Peskin, et al. Improvements in Recognition of Conversational Telephone Speech, ICASSP99, Phoenix

[4] Matthew A. Siegler, et al. On the Effects of Speech Rates in Large Vocabulary Speech Recognition Systems, ICASSP95

[5] Michael D. Monkowski, et al. Context Dependent Phonetic Duration Models for Decoding Conversational Speech, ICASSP95

[6] Anastasios Anastasakos, et al. Duration Modeling in Large Vocabulary Speech Recognition, ICASSP95

[7] Jean-Luc Gauvain, et al. Partitioning and Transcription of Broadcast News Data, ICSLP98, Sydney

[8] Hua Yu, et al. Experiments in Automatic Meeting Transcription using JRTk, Proc. ICASSP '98, Seattle, USA, May 1998

[9] Michael Finke, et al. Flexible Transcription Alignment, Proc. ASRU '97, Santa Barbaba, USA, Dec. 1997

[10] Jon G. Fiscus, A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), Draft. 1997 LVCSR DARPA HUB-5E Workshop, May 13-15, 1997

[11] Alex Waibel, et al. Meeting Browser: Tracking and Summarising Meetings, Proc. DARPA Broadcast News Workshop 1998