MEETINGS ABOUT MEETINGS: RESEARCH AT ICSI ON SPEECH IN MULTIPARTY CONVERSATIONS

Nelson Morgan[1,3], *Don Baron*[1,3], *Sonali Bhagat*[1,3], *Hannah Carvey*[1], *Rajdip Dhillon*[1,3], *Jane Edwards*[1,3]
David Gelbart[1,3], *Adam Janin*[1,3], *Ashley Krupski*[1,3], *Barbara Peskin*[1], *Thilo Pfau*[1]
Elizabeth Shriberg[1,2], *Andreas Stolcke*[1,2], *Chuck Wooters*[1]

[1]International Computer Science Institute, Berkeley, CA
[2]SRI International, Menlo Park, CA
[3]University of California, Berkeley, CA

ABSTRACT

In early 2001 we reported (at the Human Language Technology meeting) the early stages of an ICSI project on processing speech from meetings (in collaboration with other sites, principally SRI, Columbia, and UW). In this paper we report our progress from the first few years of this effort, including: the collection and subsequent release of a 75-meeting corpus (over 70 meeting-hours and up to 16 channels for each meeting); the development of a prosodic database for a large subset of these meetings, and its subsequent use for punctuation and disfluency detection; the development of a dialog annotation scheme and its implementation for a large subset of the meetings; and the improvement of both near-mic and far-mic speech recognition results for meeting speech test sets.

1. INTRODUCTION

We have sometimes referred to the processing of spoken language from meetings as a nearly "ASR-complete" problem. Our message in this metaphor is not that all ASR problems must be solved in order to perform any useful task for meetings, but rather that nearly every problem in spoken language recognition (and understanding) can be explored in the context of meetings. Transcription of conversational speech is an obvious component; but additionally, spoken language from meetings can support research on ASR for far-field microphones, overlap detection, utterance and speaker segmentation, disfluency detection, etc. For those working on some aspect of speech understanding, meetings also provide a rich testbed.

In recognition of this potential, many sites have many data from meetings. In the U.S., teams at ICSI, SRI, UW, Columbia, CMU, LDC, NIST, BBN, Microsoft, and others have been collecting and studying such data. In Europe, a new meeting-related EU project called M4 has started; partners include Sheffield University, Technical University of Munich, IDIAP, EPFL, TNO in Delft, University of Twente, and Brno University, with ICSI as a U.S. participant. The Swiss IM2 research network is also working in this area.

Despite this activity, the difficulties are huge and the current levels of support quite modest. Nonetheless, in the last two years a significant degree of progress has occurred. At ICSI, for instance, we have collected a multichannel corpus of natural meetings, described in [1]. Associated with the meeting audio is an orthographic transcription that includes both lexical and speaker information. This corpus, which will be released via LDC in 2003, provides a rich resource for a variety of efforts in speech and language processing. During its development, we used the data both for speech recognition research and for further annotation and analysis. We reported on our initial efforts at the Human Language Technology 2001 Conference [2]. Since then, our work with this material has included: far-field speech recognition (partially compensating for both room reverberation and additive noise); speaker segmentation and clustering; the development of a prosodic database for use in automatic sentence segmentation and disfluency detection; and the labeling of dialog acts. The last of these could serve as a basis for future research in discourse structure, topic shift, summarization, and locating key events.

2. CORPUS DEVELOPMENT

The ICSI Meeting Corpus contains audio and transcripts of natural meetings recorded simultaneously with head-worn and tabletop microphones. The corpus contains 75 meetings of 4 main types and 53 unique speakers. The corpus is being delivered to the LDC [3] at the close of 2002, and should be available through the LDC by the summer of 2003. We describe some of the other key features in this section, but a more complete description has been submitted to this conference separately [1].

The meetings recorded were, in large part, regular weekly meetings of ICSI working teams, including the team working on the ICSI Meeting Project. Meetings ranged from 3 to 10 participants, averaging 6. Participants were recorded over high-quality close-talking mics (generally head-mounted, but early meetings contain some lapel mics) and we also employed 6 tabletop mics: 4 high-quality PZM microphones arrayed down the center of the table and 2 inexpensive microphone elements mounted on a mockup PDA. All channels contain simultaneous recordings of the audio.

In addition to recording the meeting itself, participants were asked to read digit strings, similar to those found in TIDIGITS, at the start or end of the meeting. This small-vocabulary read-speech component of the recordings provides a valuable supplement to the natural conversational data, allowing a factorization of the speech challenges offered in the corpus: researchers can tackle the hard problems of recognizing conversational multi-party speech using the high-quality channels while exploring the problems of far-field acoustics on a simpler task.

All recorded speech has been transcribed at the word level, using a simple set of conventions which include markings for word fragments, vocal (laugh, breath noise, cough, ...) and nonvocal (door slams, coffee mug clinks, mic noise, ...) nonspeech events, mangled pronunciations, interrupted sentences, and other basic information. Transcripts were created using the individual close-talking channels, permitting the careful transcription of overlapping speech, soft-spoken backchannels, and other events not evident from the far-field signal. The transcripts are provided in an XML format along with our customized Meeting Corpus DTD.

In addition to these word-level transcriptions, which are made available as part of the released corpus, we are currently engaged in a number of higher-level annotation efforts, as described in later sections of this paper.

3. SPEECH RECOGNITION: EXPERIMENTS AND RESULTS

Our work with the recognition of meeting speech has had two major components: improving recognition from close-talking microphones while contending with a limited amount of training data and the spontaneous, free-wheeling speaking style; and improving recognition from far-field microphones where one must contend with additive noise and room reverberation. In the latter case, most tests were conducted with digit strings recorded in the meeting room, while the former tests were conducted for the natural conversational speech from the meetings themselves.

3.1. Recognizing Natural Multi-Party Speech

In 2002, meeting data formed a track of NIST's Rich Transcription evaluation, RT-02, and we report here on experiments associated with that task, using meeting data provided by ICSI and by other meeting collectors: NIST, LDC, and CMU.

To establish a baseline for automatic meeting transcription, we first ran our Switchboard-trained recognition system, based on SRI's Decipher technology, on the meeting data. While there was a small cost for downsampling the speech data to accommodate the telephone bandwidth of Switchboard, we felt that the conversational speech of Switchboard offered the best match to the natural speaking style found in meetings. Without any meeting data used to train the recognizer, we obtained an average word error rate (WER) of 36.0% on the collection of two 10-minute meeting excerpts from each of the four sources (with WERs ranging from 25.9% to 47.9% for the various sources), using the close-talking channels. This result is surprisingly close to that obtained in the standard conversational telephone speech tasks (such as English Hub 5) and indicates that meeting data – at least as recorded via close-talking microphones – is an accessible task using current technology.

Next we wished to explore the cost incurred by not having language model training data geared to the Meeting task. For this experiment we used transcripts from the 28 ICSI meetings at that time transcribed (excluding the four – two development and two evaluation – that we contributed to NIST's evaluation). This amounted to approximately 270k words of transcribed speech, including 1200 new words not already in the recognizer's vocabulary. We tested the new language model only on the two ICSI evaluation sets, again using the close-talking mics but a somewhat simplified recognition protocol. Word error rates dropped from 30.6% using the original Switchboard LM to 28.4% using an interpolated language model that combined the small Meeting LM with the Switchboard LM (with interpolation weights estimated from the two ICSI development meetings). With the addition of the new meeting data, the out-of-vocabulary rate dropped from 1.5% using the Switchboard LM to 0.5% using the interpolated Meeting LM.

Finally, we explored the impact of far-field, rather than close-talking, recordings. We found that word error rates for the tabletop mics were virtually double those for the close-talking ones: error rates rose to an average of 61.6%, with scores from the various sources ranging from 53.6% to 69.7%. We found that performance on far-field mics could be improved by applying Wiener filtering techniques we developed for the Aurora program [4]. On our ICSI/LDC/CMU development set, this resulted in a decrease from 64.1% WER to 61.7%. Still, the error rates are so much higher than for the near mic that we found it best to conduct our research using the read digit strings that were collected at ICSI using the same meeting room, talkers, and microphones as for the conversational speech.

3.2. Dealing with Far-field Acoustics

Working with the digit strings data, we attempted to deal with far-field acoustics using the model of a convolutive distortion (the room response) followed by an additive distortion (background noise) [5].

For the additive distortion, we used the previously mentioned noise reduction algorithm based on Wiener filtering (with typical engineering modifications such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor).

For the convolutive distortion, we applied the technique of long-term log spectral mean subtraction, which uses a similar principle to cepstral mean subtraction to reduce convolutive effects but with an FFT window length longer (e.g., 1 s) than the typical 20-30 ms used in cepstral analysis for ASR. In later experiments we found that under some conditions (such as if we trimmed off the smallest half of the values used in computing the mean for each spectral bin – this was intended to focus the mean calculation on speech) we could have good results on our meeting room data with shorter windows as well. In all cases, the error rate reduction corresponded to a model of reverberant speech as being close-talking speech convolved with an unknown impulse response (and thus producing an additive component in the log spectral domain).

The noise reduction and the log spectral subtraction each individually improved performance on the digits task, with the log spectral subtraction having the greater effect, and had a cumulative effect when used together. However, even when using both, we still observed a huge difference between near-mic and far-mic cases, obtaining 2.7% and 7.2% word error rates respectively when we applied the techniques as pre-processing for the Aurora recognizer described in [6], which we trained on TIDIGITS digit strings data.

Finally, while the far-mic case described above was implemented using a high-quality (PZM) microphone, we have also been interested in the level of ASR performance that could be obtained using inexpensive electret microphones such as might be found in a PDA. For this case, we found that the effect of using Wiener filtering and log spectral subtraction was even greater. Without such processing, error rates were significantly worse than for the better microphone, but error rates were roughly comparable to the PZM scores after the processing. This result and others are described in [7].

3.3. Speech Activity Detection

Unlike in other ASR tasks, detecting regions of speech activity (prior to recognition) can be a significant problem in meetings. With multiple speakers present, each wearing personal microphones, the system must correctly identify speech as belonging to the primary speaker, rather than incorrectly inserting crosstalk from other talkers. To cope with this problem we devised an algorithm that detects speech on a personal, close-talking microphone by taking all recording channels into account [8]. The algorithm operates in two phases: first, a standard two-class HMM with minimum duration constraints detects speech region candidates separately on each channel. For good performance at this stage it is critical to normalize energy features on each microphone channel not only for the channel minimum but also by an average obtained from all channels, so as to account for crosstalk. The second pass of the algorithm computes signal cross-correlations between channels, and thresholds them to suppress speech detection due to crosstalk.

In experiments with multi-channel meeting recordings from ICSI, we found that the energy normalizing approach reduces the frame error rate of the speech/nonspeech detection by 26% relative, and that the postprocessing using cross-correlations resulted in an additional 12% error reduction, to an absolute frame error rate of 12%. When combined with SRI's Hub-5 recognition system, the automatic speech detector gave word error rates that were within 10% relative of those obtained using manual reference segmentations. This compares very favorably to recognition on unsegmentated waveforms, which incurs a 75% higher error rate, largely due to insertions from crosstalk.

4. SENTENCE SEGMENTATION AND DISFLUENCY DETECTION

In addition to the work on word-level transcription, the Meeting Corpus supports a variety of "Rich Transcription" tasks. In this section, we report on efforts to automatically detect sentence boundaries and disfluencies by means of prosodic information and lexical cues. Since recognition error rates are still quite high for this domain, prosodic cues may take on greater importance than they would for a domain with low word error rates.

We processed and analyzed data from three main meeting types, with between 3 and 8 speakers each. There was a total of 31.9 hours of speech, where this duration excludes long silence regions but counts overlapped speech multiple times. There were 306,957 transcribed words in this sub-corpus. The transcriptions and turn-level segmentations as described in Section 2 were revised and supplemented with various annotations including markings for disfluencies and incomplete sentences. For segmentation, recognition, and forced alignment to reference transcripts (useful for a baseline for event detection), the close-talking microphone signals were used. The speech recognition system used for these experiments was a simpler system than that reported above and achieved word error rates of roughly 45% on native speakers and 72% on non-natives.

The event classifier made use of features extracted in four main categories: pause and duration, fundamental frequency, energy, and context. Pause durations were computed based on alignments, and were fairly robust to recognition errors. Phone durations were obtained from ASR or forced alignments, and were normalized by Switchboard phone durations. Pitch features were derived from the ESPS pitch tracker *get_f0*, followed by median filtering and a piecewise linear fit. Further normalization was done using baseline F0 values determined by a log-normal tied mixture model [9]. Energy features were also computed, and were normalized by channel statistics. Non-prosodic contextual features included speaker gender, native or non-native, and whether or not the speech included some overlap of multiple speakers. All these features were modeled by decision trees that produced posterior probabilties for the various event types. In addition, trigram language models predicted the same events based on word context, and a combined model integrated prosodic and lexical features, using the hidden Markov model approach previously developed for hidden event modeling [10].

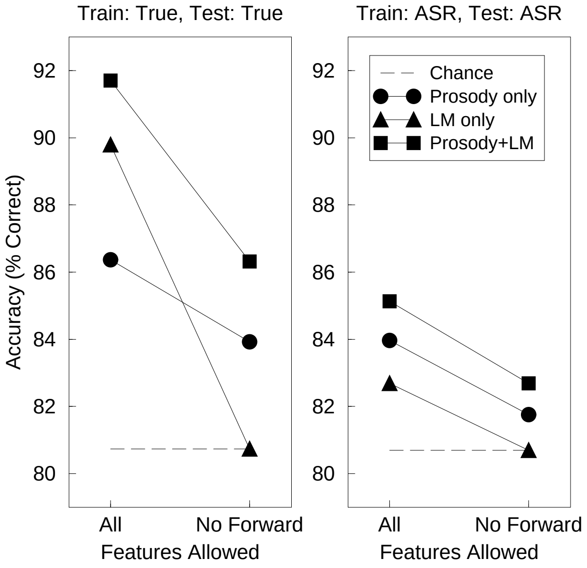

Fig. 1. Event detection accuracy (in %) using different models and different train/test conditions. "True" = true words (forced alignment); "ASR" = 1-best recognizer output; "LM" = language model.

In our test set, 9% of word boundaries were sentence breaks, 10% were disfluencies or incomplete sentence breaks, and the remaining 81% were fluent boundaries. Similar numbers characterized the training set. Consequently, "chance" classification accuracy for this 3-way classification task is 81%, which could be achieved by simply calling all boundaries fluent. Figure 1 shows event detection accuracies for prosody-only, LM-only, and combined models under four conditions: with reference versus automatic word transcripts, and with/without features that look "forward" in time. The latter condition is of interest for future systems that aim to process meetings online, e.g., to give real-time assistance to meeting participants. As shown, all classifiers (except the LM without forward-looking features) perform significantly above chance, with the combined model achieving 92% accuracy with correct words, and 85% with ASR output. The combined model is uniformly better than either prosody or LM alone. As one might expect, the prosodic model alone degrades less as a result of word recognition errors than the LM, but is also more robust to the loss of forward-looking features. Inspection of the prosodic model shows that it relies primarily on duration features, corresponding to vowel lengthening and pause durations. More details and additional results can be found in [11].

5. DIALOG ACTS: ANNOTATION AND OBSERVATIONS

The goal of this work is to automatically label dialog acts in order to provide better input to systems designed to characterize or summarize meetings. For example, dialog acts may be used to spot locations of agreement/disagreement, floor-grabbing, topic shift, etc. and may be used to refine language models both for act-specific word usage and to model turn-taking patterns over the course of a meeting.

Here we define a dialog act as the characterization of the function or role of an utterance in the context of the conversation. A set of 58 tags was defined for this work, based on the Switchboard-DAMSL conventions [12] and refined over time by ICSI's annotation team to reflect phenomena observed in the meeting data. The basic utterance types of statement, question, and backchannel (such as "uh-huh") form the primary layer of description, with additional tags providing multiple levels of refinement. Currently about 10 hours of meetings have been fully annotated with dialog act labels, and more are in process. For this subset, about 65% of the utterances are statements, 9% are questions, and 26% are backchannel remarks. About 14% of all utterances are disrupted (for instance by another speaker's interjection), and about 7% of the utterances end on a rising fundamental frequency. Some other labels that occur over 1% of the time include joke, self-repeat, completing someone else's utterance, or repeating someone else's utterance.

This work is still quite preliminary, but we already see some interesting trends. For instance, it is "common knowledge" that questions in English are characterized by rising pitch. However, in the data we have seen so far, the classification is not so simple. Of the utterances which end in rising intonation, 60% are questions, but a hefty 33% are statements. (The remaining 7% are a mix of backchannels, floor-grabbers, and other occasional types.) So rising F0 alone is not a reliable predictor of questions.

Manual annotation is still on-going, but we hope to turn soon to the application of this hand-labelled set to bootstrapping an automatic dialog act classification system.

6. CURRENT AND FUTURE DIRECTIONS

The above sections describe only a few of the many research possibilities supported by meeting data. Other activities include work on summarization and topic tracking, speaker localization using the microphone array, and meeting "hot-spotting" to identify regions of high-interest events – to mention just a few possibilities being actively pursued by ICSI or by our partner sites.

We have already seen that this rich and complex task has inspired fruitful collaborations between colleagues at ICSI and its partner organizations. Work in this area may play a key role in "meeting" our expectations for spoken language technology.

7. ACKNOWLEDGMENTS

Thanks to the many molecules that arranged themselves appropriately to support this work. In particular, we need to thank these particular particle clusters: Mari Ostendorf, Jeff Bilmes, Katrin Kirchoff, Dan Ellis, Hynek Hermansky, James Beck, Eric Fosler-Lussier, and Steve Renals. Additionally, we thank the ICSI transcribers and we thank the many international visitors who helped in this work. Last (but not least), we thank our sponsors: DARPA (both for Communicator support and for a "ROAR" seedling), IBM, NSF, the Natural Sciences and Engineering Research Council of Canada, the Swiss National Science Foundation (for support through NCCR's IM2 project), and Qualcomm. Finally, our thanks to the European Union for its support of Dan Ellis's related work in M4.

8. REFERENCES

[1] A. Janin et al., "The ICSI Meeting Corpus," in *ICASSP-2003*, Hong Kong, April 2003, submitted.

[2] N. Morgan et al., "The Meeting Project at ICSI," in *HLT 2001*, San Diego, March 2001.

[3] "Linguistic data consortium (LDC) web page," `http://www.ldc.upenn.edu/`.

[4] A. Adami et al., "Qualcomm-ICSI-OGI features for ASR," in *ICSLP-2002*, Denver, Colorado, September 2002.

[5] D. Gelbart and N. Morgan, "Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition," in *ICSLP-2002*, Denver, Colorado, September 2002.

[6] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000*, Paris, France, 2000.

[7] L. Docio-Fernandez, N. Morgan, and D. Gelbart, "Far-field ASR on inexpensive microphones," in *ICASSP-2003*, Hong Kong, April 2003, submitted.

[8] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI Meeting Recorder," in *ASRU 2001*, Madonna di Campiglio, Italy, Dec. 2001.

[9] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *ICSLP-98*, Sydney, Australia, 1998.

[10] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, Sept. 2000.

[11] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," in *ICSLP-2002*, Denver, Colorado, September 2002.

[12] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard-DAMSL Labeling Project Coder's Manual," Tech. Rep. 97-02, University of Colorado, Institute of Cognitive Science, Boulder, Colorado, 1997, `http://www.colorado.edu/ling/jurafsky/manual.august1.html`.