Meeting Recorder

Adam Janin
International Computer Science Institute
1947 Center Street
Suite 600
Berkeley, CA 94704-1105
510-666-2977
janin@icsi.berkeley.edu

Abstract

Despite recent advances in automatic speech recognition (ASR) technology, successful ASR applications tend to be limited to telephony, command-and-control functions, head-worn microphones, small vocabularies, and/or simple grammars. Recording and processing spontaneous, human-to-human conversations in natural settings is a far more challenging task.

Face-to-face meetings are not only challenging because of acoustic uncertainty, but they also contain very rich content. In contrast to typical ASR corpora, meetings contain a vast amount of overlapping speech, interrupts, false starts, laughter, and other interesting linguistic, acoustic, and prosodic features.

To address some of these issues, we have begun research on the recording and processing of meetings. In this paper, we will describe the Meeting Recorder project, including possible applications, corpus collection, and progress to date.

1 Why Record Meetings?

1.1 Challenging Speech Recognition

Firstly, natural human-to-human meetings provide a new, very challenging domain for research into automatic speech recognition (ASR). Although it is similar in spirit to other conversational corpora (such as Switchboard [1]), meetings pose an even greater problem for ASR [2].

The acoustic environment in a real meeting tends to be quite variable. Background noise such as fans, door slams, and paper rustling can all contribute to the acoustic background. Reverberation and echo can also be a significant problem in real rooms with desktop microphones.

Many sections of real meetings contain large amounts of overlap – people end each other's sentences, interrupt, encourage ("uh huh"), laugh, and so on. Not only do ASR systems perform very poorly on overlapped speech (the "cocktail party" problem), but overlap also makes traditional segmentation into speaker-disjoint regions somewhat problematic. For more details on segmentation, see section 4.1, below.

As part of the corpus collection process at ICSI (discussed in more detail in section 4, below), we record real meetings that take place regardless of the Meeting Recorder project. As such, they appear to be quite natural. They contain all the artifacts of spontaneous speech seen in other corpora, including false starts, corrections, filled pauses, and disfluencies. Furthermore, since most of the recordings are of regular weekly meetings, the participants often know each other quite well. In many cases, this leads to highly informal conversational styles, with many interrupts, asides, back channel communications, jokes, etc. As compared to prepared speech or read speech, this style of speech is quite difficult for traditional ASR systems to recognize.

Because of all these difficulties, we expect very poor initial results for meetings. As such, the corpus provides an excellent arena for research into robust ASR. See section 5 for some very preliminary results.

1.1.1 Digits

In order to provide a simpler task, and to isolate acoustic issues from other effects, we also record digit strings as part of the corpus. At the start or end of each meeting, the participants are given a sheet of paper with about 20 lines of digits. Each line contains from one to ten digits spelled out in English (so that "zero" and "oh" can be distinguished). They are asked to read them in order, with a short pause between lines. The digits strings themselves are taken directly from an existing digit corpus called Aurora [3], although the recording environment in Aurora is quite dissimilar.

The digits task is much easier than the general task for a variety of reasons. There is no overlap, since each participant reads the digits separately. Because it is read speech, there are fewer disfluencies and pauses. Finally, the number of words in the vocabulary is merely 11, rather than the 65000 that are typical of a dictation task.

1.2 Meetings are Interesting

Meetings are also interesting in their own right. In addition to the acoustic characteristics described above, there is also interest in extracting prosodic information[1], perhaps to detect salient parts of the meeting. Sorting by importance may aid information retrieval. It would also be extremely useful to detect specific parts of a meeting such as summary information, progress reports, action items, and so on. Finally, social structures in meetings may be derivable from detailed information present in the corpora - who speaks the most? Who interrupts whom? When does laughter occur?

2 SpeechCorder

A primary goal of Meeting Recorder is to provide a system that can search the audio record of a meeting using spoken or written queries, either in real-time during a meeting or offline. Combined with a note taking application, such a system could replace written notes and tape recordings with a much more useful and interactive tool. A related project underway at ICSI, called SpeechCorder, aims to provide such a device on a portable computer. In this section, we will discuss only the information retrieval application itself. Details on the hardware and software of SpeechCorder can be found in [4].

2.1 Why Not Written Notes?

- **Unreadable Handwriting**: Often, written notes are illegible. This becomes even more of a problem as time passes. Notes that were comprehensible when written become indecipherable after a few weeks.
- **Detracts From Listening**: It can be very difficult to take notes and pay close attention to the meeting at the same time. Many people find that they can take notes or they can listen closely, but not both.
- **Incomplete Record**: It is almost impossible to record a meeting in its entirety. Frequently, important points and data are missing from the written record.
- **Hard to Search**: Written notes provide no support for searching and indexing. Finding particular information is very difficult.

2.2 Why Not A Tape Recorder?

- **Hard to Annotate**: With written notes, adding arbitrary annotations (underlining, circling, diagrams, doodles) is trivial. It is much more difficult to provide synchronized annotations with a tape recorder.
- **Playback Can Be Disruptive**: One cannot play back a tape recording during a meeting without disturbing the other participants. Also, one cannot play back and record simultaneously.
- **Too Much Data**: Recording an entire meeting via tape produces a very large amount of extraneous data.
- **Hard to Search**: As with written notes, a tape recording of a meeting cannot easily be indexed and searched.

2.3 Hypothetical Meeting Recording Example

Figure 1 shows a hypothetical Meeting Recorder screen shot. The image demonstrates some of the potential user-interface ideas for correction and annotation of the transcript.

As participants of the meeting speak, a possibly inaccurate transcript appears in real-time. Speaker change boundaries are detected and shown (in this case, just by numbering).

Since the speech recognizer often can produce a measure of how confident it is with its guess, the user-interface could highlight potentially poor transcriptions.

When the recognizer makes an error, the user can quickly correct it by clicking on the incorrect word. The recognizer generates a popup box of alternatives from which the user can select the correct word.

Also shown is a text annotation entered by the user.

[1] The non-lexical parts speech such as pitch, speaking rate, intonation, etc.

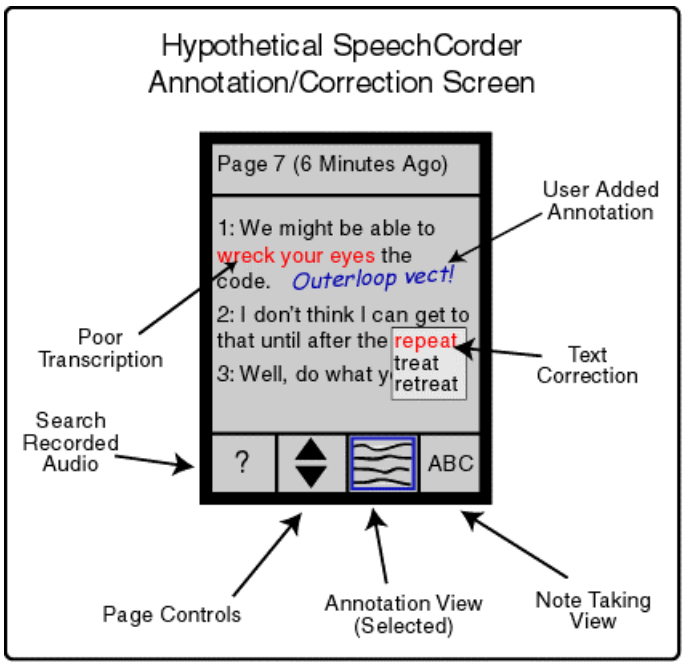

Figure 1 – Hypothetical SpeechCorder Screen

3 SpeechCorder Requirements

For SpeechCorder to be useful, it must meet a number of challenging requirements. It must:

1. Record meetings in natural settings.
2. Support multiple speakers.
3. Work stand-alone.
4. Support indexing and searching.

3.1 Record Meetings in Natural Settings

One of the primary design goals of SpeechCorder is to record real meetings in real settings. We would also like to include the ability to record impromptu meetings. The speech recognition must therefore work in uncontrolled acoustic environments, as described above. The vocabulary must be large enough to cover the domain of the meeting, and it must be robust to foreign accents.

3.2 Support Multiple Speakers

Obviously, meetings have multiple participants. The speech recognition system should not only transcribe the speech of different people, perhaps talking at once, but it also should note when the speaker changes. It would also be convenient if it could record not only the speaker change, but also the identity of the speaker. This would allow searching by speaker as well as by the audio contents of the record.

Speech recognition systems can improve their accuracy by adapting to a particular user. Allowing the system to store and exchange user profiles among meeting participants could dramatically improve the quality of the transcript.

3.3 Work Stand-alone

If we want to support impromptu meetings in uninstrumented environments, it is necessary for SpeechCorder to be portable. Although it is certainly possible to use a hand-held computer as a terminal using a wireless network, we feel that a self-contained solution is better in the long run. The terminal/main-frame model has more components to fail - the terminal, the network, the wireless link, the mainframe, and the infrastructure. There is also the question privacy, which is much easier to address with a self-contained unit. The PC revolution has also shown us the utility of personal compute power.

3.4 Support Indexing and Searching

As described in section 2 above, we plan to provide for searching of the record both by text content and by speaker. Since the actual audio will be stored, the user can play back the portion of the meeting that matches his criteria. We will support both textual and spoken queries.

Note that, since the audio record is stored, the transcript need not be perfect. It need only be good enough so that queries match where users expect them to match. For referring to the actual content, the user can play back the audio. In addition, the speech recognizer can output not just the most likely transcript, but also a list of the top few most likely hypotheses (called N-best lists in speech recognition). Queries can be made against these N-best lists, rather than just the best hypothesis. Finally, it turns out that the recognizer does much worse with so-called function words (such as “the”, “a”, “of”, “an”) as opposed to content words. However, for text retrieval, systems usually ignore function words. For all of the above reasons, it is perfectly satisfactory for the recognizer to be imperfect. In fact, we expect word-error rates of up to 40% to be acceptable [5].

3.5 Other Issues

There are many other issues related to the design and use of the Meeting Recorder. The following paragraphs outline a few.

3.5.1 Will People Object?

Will people object to being recorded? This has proved to be a real problem both with video records and our pilot audio recording. Although audio records seem to be less of a problem, issues of privacy must still be addressed. Simple “muting” is probably insufficient, since many participants may have SpeechCorders. For our corpus collection effort, we allow participants to “bleep out” any section of the record that they do not wish to be made public. For a deployed application, something similar will probably be required.

3.5.2 Problems with Data Collection

The data collection phase is necessarily somewhat artificial, since the participants in the corpus collection phase wear head-mounted microphones, and are aware that they are being recorded. Some participants also use wired microphones that are connected to a jack under the table, while others use wireless microphones. It remains an open question on how the constraints of data collection affect real meetings. Does the speaking style change? Are people equally mobile? Although we have not done detailed measurements, it appears that people adapt very quickly to the setup, and do not change their behavior based on it.

3.5.3 Cross Domain Training

We cannot hope to cover the vocabulary and language models of all possible meetings. What commonalities can be found? How much data will we need to “cover” a new domain? Can we start with existing domain? Do textual corpora help?

4 Corpus Collection

In order for a speech recognition system to work, it must be trained on transcribed audio in the domain in which it will be used. To form a baseline and to control for factors such as noise, reverberation, multiple speakers at a time, and so on, our test subjects during corpus collection wear head-mounted (near-field) microphones. We simultaneously record the meeting using desktop omni-directional microphones and PDA mounted microphones. This allows us to generate very good acoustic data for training, in addition to the degraded acoustics of a desktop microphone. It also provides an easier method to mark speaker identity (using the output from the head-mounted microphones).

We are in the process of collecting the first such corpus. The domain consists of meetings at the International Computer Science Institute (ICSI), including Meeting Recorder, speech recognition, networks, and natural language processing meetings. We have instrumented a meeting room at ICSI with the following equipment:

- Five wireless head-mounted microphones.
- Four wired head-mounted microphones.
- Four PZM omni-directional desktop microphones.
- Two mockup PDA microphones.
- High quality multichannel audio recording equipment.

The equipment records all the channels concurrently and frame-synchronously. This allows direct comparison between microphone signals without additional correction for time skew. Since we record all 15 channels without mixing, the data requirements are somewhat higher than typical audio applications. Even compressed, the audio takes about 500 MBytes / hour.

As of March 2001, we have collected about 40 hours of meetings, including about 4000 digit strings. We intend to continue collection up to about 100 hours. Additionally, meetings at the University of Washington, SRI, and NIST will begin shortly. It is our intention to combine all the efforts into one corpus, which will be distributed through standard methods (e.g. the Linguistic Data Consortium).

4.1 Transcription

The task of transcribing the audio from these meetings is somewhat daunting. A rule of thumb is that one hour of audio requires ten hours to transcribe for an experienced transcriptionist. Since we have multiple channels, speaker identity markings, and some discourse and prosodic markings, this estimate is somewhat low for Meeting Recorder.

We have completed about 10 hours of transcription in-house at ICSI. We are in the process of transcribing the rest of the data.

Although tools are available for transcription, we need to modify the tools and the procedures in order to obtain good quality transcripts that include detailed time markings, especially in overlapped regions. For example:

Speaker 1: We plan to, uh, what I'd like to do is rewrite that part ... ... yeah, the detector, for the whole ...			
Speaker 2:	all?		the detector?
Speaker 3:		uh huh	yes

We plan to only roughly indicate overlap for most of the corpus, simply because detailed marking of exactly when people start and stop talking in complex regions can be extremely time consuming for the transcriptionists. So we would transcribe the above as something like:

S1: We plan to, uh / S2: all?

S1: what I'd like to do is rewrite that part / S3: uh huh

S2: the detector? / S1 yeah, the detector, for the whole / S3: yes

We are in the process of extending the transcription tools to ease the effort of marking exact time boundaries. Once this is complete, we plan to mark a subset of the data in much higher detail.

To aid the transcriptionists, we have also written a tool that does initial segmentation based only on the audio signal. We take the signal from the head-worn microphones, and perform a procedure that distinguishes between speech, noise, and silence. From this, we generate a segmentation of the data that at least roughly marks breaks in the conversation, or where the primary speaker changes. Although the automatic tool makes many errors in the segmentation, it is much easier for the transcriptionists to correct the segmentation than to produce one from scratch.

5 Preliminary Results

We are only in the very initial stages of the project, and therefore have few concrete ASR results to report. We have run an existing recognizer (the SRI Switchboard recognizer) on a few meetings using the segmentation provided by the transcriptionists and using the near-field microphone signals. We ran the recognizer only on the segments that were marked by the transcriptionists as containing speech. A system trained on real meeting data would do better than this preliminary system. Using automatic segmentation would do worse. Regardless, the result for native English speakers was generally in the 40% word-error range. Some of the non-native native speakers had very large errors (e.g. greater than 100%), indicating that more work needs to be done with accents. The results are generally encouraging – for native speakers, they are roughly the same as a system trained on one version of the Switchboard corpus and tested on another.

6 Summary

We have begun work on recording natural meetings for use in speech recognition research, information retrieval, and meeting analysis. Corpus collection is also underway, with 40 hours collected and 10 hours transcribed. Preliminary results are encouraging, but much work remains to be done. More details and up-to-date information can be found on our web site [6].

Acknowledgements

We would like to thank all our collaborators on the project both at ICSI, and ARPA, IBM, SRI, University of California Berkeley, and the University of Washington.

References

[1] J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: A telephone speech corpus for research and development. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP92)*, volume 1, page 517—520, San Francisco, 1992.

[2] Hua Yu, Michael Finke, and Alex Waibel. Progress in automatic meeting transcription. In *6th European Conference on Speech Communication and Technology (Eurospeech-99)*, volume 2, pages 695—698, Budapest, September 1999.

[3] D. Pearce. Aurora Project: Experimental framework for the performance evaluation of distributed speech recognition front-ends. *ETSI working paper*, September 1998.

[4] A. Janin and N. Morgan. SpeechCorder, the portable meeting recorder. *Workshop on hands-free speech communication*, April 2001.

[5] D. Abberley, S. Renals, and G. Cook. Retrieval of broadcast news documents with the THISL system. In *Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP98)*, pages 3781—3784, 1998.

[6] Meeting Recorder web pages. http://www.icsi.berkeley.edu/real/meeting-recorder.html