Reducing the Effect of Room Acoustics on Human-Computer Interaction

David Gelbart

International Computer Science Institute (ICSI)
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
(510) 666-2990

<gelbart@icsi.berkeley.edu>

Reducing the Effect of Room Acoustics on Human-Computer Interaction

David Gelbart
International Computer Science Institute (ICSI), Berkeley, CA

Abstract

Hands-free use of speech recognition (i.e., not requiring a microphone worn or held by the user) introduces the technical challenges of room acoustics and background noise. In this paper, I will start by describing a possible hands-free application from the SmartKom dialogue system project. I will then describe the acoustical issues involved and the possible technical approaches to dealing with them. I will then discuss one such approach that we have been working with, long-term log spectral subtraction, and give experimental results examining its usefulness for interactive applications.

SmartKom

The SmartKom project [Smartkom; Wahlster] is intended to create multimodal dialogue systems that combine the use of speech and gesture (both by the system and the user). Figure 1 illustrates the SmartKom Home system concept: a portal to information services such as television for home use. The 'face' of the system is the blue character on the left, named Smartakus, who communicates with the user with both synthesized speech and animated gestures. The use of an animated character is intended to make the system easier for novice or naive users. User queries are spoken (for example, "What movies are on TV tonight?") and the user can use gestures while they are speaking to point to parts of the display.

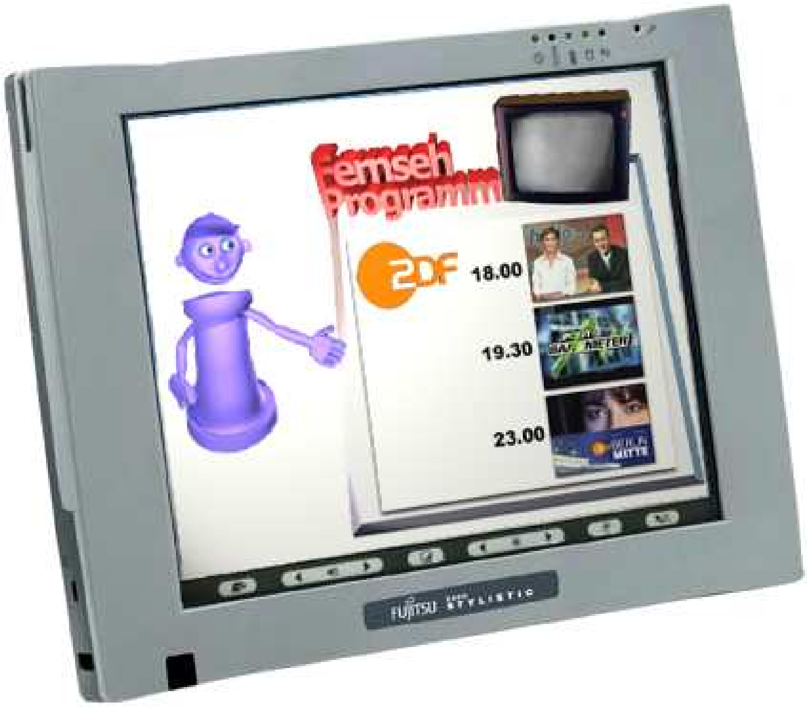

Figure 1: SmartKom Home.

Room Acoustics and Hands-Free Speech Recognition

The system shown in Figure 1 is envisioned to be usable hands-free without requiring the user to carry or wear a microphone. Making hands-free recognition more reliable is a major research problem. This is because, compared to recordings made near the user's mouth, recordings made from more distant microphones have higher levels of background noise relative to the speech level and are more affected by reverberation due to echos off reflective surfaces such as walls. Even a modest degree of reverberation, which would present little or no difficulty to a human listener, can substantially decrease the performance of automatic speech recognition systems. The greater distance that sound travels to the microphone also results in higher frequencies in speech being weakened relative to lower frequencies. I will call the combined effect of reverberation and other effects such as this the 'room response' from the user to the microphone. It is common to model the room response as a filter applied to the original speech which affects the spectrum of the speech in a consistent way.

There are various approaches which could be tried to improve recognition performance in this situation:

- single microphone signal processing to reduce the effects of noise, reverberation, and spectral distortion
- microphone array signal processing to the same purpose
- using noisy or reverberant training data to create the speech models [Stahl]
- using adaptation methods to adjust the speech models
- using methods of representing the signal that are less sensitive to the distortions caused by reverberation, such as the modulation spectrogram [Kingsbury]

Several of these are discussed in [Omologo], which also has a useful discussion of room acoustics. The remainder of this paper will focus on the first approach, which is not to say that I feel it is the single best approach. I think it is likely that the best performance will come from combining approaches.

Long-term Log Spectral Subtraction

Carlos Avendano [Avendano] developed a method of room response compensation which estimates the room response using the assumption that it does not change and so can be estimated by measuring the unchanging part of the speech spectrum. The estimated room response is then removed from the speech spectrum by subtraction (in the logarithmic magnitude domain). Since speech itself has average characteristics, the room response estimate also contains information about the speech signal. Because of this and other artifacts of this method, we train the recognizer on processed speech.

This method is similar to the common technique of cepstral mean subtraction, which is used to compensate for microphone or telephone channel effects. Reverberation effects have a longer extent in time, so to deal with reverberation Avendano's method calculates the speech spectrum using longer periods of speech than is normally done for cepstral mean subtraction. I will call his method 'long-term log spectral subtraction' to emphasize this.

The method assumes that the room response is approximately constant. In fact people are rarely perfectly still, and the room response from them to a microphone will change as they move, but we have found this method to be useful in realistic data collected with seated speakers. It is less likely to work if the speaker is walking around a room.

Experimental Results

In experiments at ICSI last year, we found long-term log spectral subtraction to be useful for offline (non-interactive) recognition [Gelbart01]. In that work we collected several utterances worth of data to estimate the room response before perfoming the subtraction and starting the recognizer. I will now present some of those results and then discuss the use of this method in an interactive system.

For our experiments we used the Aurora reference system described in [Hirsch], which is based on the HTK speech recognizer configured to recognize digits. For training data we used four hours of data from the TIDIGITS connected digits corpus, which was collected by close-talking microphone in a quiet environment. To test it, we used connected digits strings (a total of 7704 words) which were read by native English speakers seated around a conference table in a room we are using for recording natural meetings. Simultaneously recordings were made with close-talking microphones and with a table-mounted microphone that was 3-6 feet from each speaker.

Table 1 shows the original system's performance, measured by word error rate (WER). (Word error rate is the fraction of words omitted, changed, or falsely inserted by the recognizer out of the total number of spoken words.) The first column ("Near") refers to the close-talking microphones and the second column ("Far") refers to the table-mounted microphone.

Near microphone	Far microphone
4.1%	26.3%

Table 1: Baseline results.

Table 2 on the next page shows the WER results when the long-term log spectral subtraction method was applied to the training and test data. The room response was estimated using 7.168 seconds of data. (This was done separately for each speaker, and if the same speaker re-appeared during different recording sessions we did the estimate separately each time.) Performance improved dramatically for the far test data. It also improved significantly for the near test data. Reverberation may not be an issue with the near microphone, but since the original system did not include any kind of channel compensation (such as cepstral mean subtraction) to compensate for differences in microphone type, etc., between the training and test data, it's likely that the method is helping in this regard.

Near microphone	Far microphone
3.0%	8.2%

Table 2: Results with long-term log spectral subtraction.

In an interactive application, collecting several utterances worth of data to estimate the room response before beginning processing is not feasible—utterances need to be recognized as soon as the user has finished speaking them. Therefore, to investigate the use of long-term log spectral subtraction in interactive applications I modified the algorithm to process the current utterance using a room response estimate calculated from whatever utterances the user has spoken thus far, so that the current utterance be can passed immediately to the recognizer.

The question now is whether the method will still perform well for the first few utterances, where only a few seconds of data are available to estimate the room response. (The average utterance length in the test set was 1.3 seconds.) Table 3 shows the new far microphone WER broken down by whether utterances were the first, second, third, etc. utterance (indicated in the first column) spoken by that speaker in that recording session. The second column gives the total number of words and the third column gives the WER.

Utterance number in session	Total words	Far WER
1-2	679	12.8%
3-4	692	8.4%
5-6	702	8.4%
7-8	670	9.3%
9-10	736	7.5%
11-12	692	7.4%
13-14	661	7.6%
15-16	679	9.3%
17-18	695	8.1%
19-20	697	6.3%
21+	801	4.7%

Table 3: Far microphone results with past-and-present-only long-term log spectral subtraction.

The results in Table 3 show that after the first two utterances the mean subtraction is performing well. This is encouraging for the use of it in an interactive system. Incidentally, the especially good result in the last row showed up in the experiment in Table 2 as well (only 6 of the 17 speakers supplied more than 20 utterances in any recording session; perhaps they were easier to recognize than average).

Conclusions

The SmartKom project is aiming at natural human-computer interaction through voice I/O (as well as images and gesture). Making interactive systems like this usable in hands-free way is a current research goal in automatic speech recognition. I have reviewed the difficulties involved and presented results from my research group's work using long-term log spectral subtraction to combat these difficulties. The results show that long-term log spectral subtraction can be useful even in an interactive system. However, the hands-free error rate remained much higher than the error rate from the close-talking microphones. Background noise is likely to have been a factor in this, and my research group has begun to investigate combining long-term log spectral subtraction with noise suppression techniques [Gelbart02].

References

[Avendano] C. Avendano, S. Tibrewala and H. Hermansky, “Multiresolution Channel Normalization for ASR in Reverberant Environments”, in *EUROSPEECH*, Rhodes, 1997.

[Gelbart01] D. Gelbart and N. Morgan, “Evaluating Long-term Spectral Subtraction for Reverberant ASR”, in *ASRU*, Madonna di Campiglio, 2001, with a correction at http://www.icsi.berkeley.edu/Speech/papers.html.

[Gelbart02] D. Gelbart and N. Morgan, “Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition”, submitted to *ICSLP*, Denver, 2002.

[Hirsch] 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, 2000.

[Kingsbury] Brian Kingsbury, Nelson Morgan and Steven Greenberg, “Robust speech recognition using the modulation spectrogram”, *Speech Communication*, vol. 25, pp. 117-132, 1998.

[Omologo] M. Omologo, M. Matassoni, and P. Svaizer, "Environmental Conditions and Acoustic Transduction in Hands-free Speech Recognition", *Speech Communication*, vol. 25, pp. 75-95, 1998 .

[SmartKom] http://www.smartkom.com

[Stahl] V. Stahl, A. Fischer, and R. Bippus, “Acoustic synthesis of training data for speech recognition in living room environments”, in *ICASSP*, Salt Lake City, 2001.

[Wahlster] W. Wahlster, N. Reithinger, and A. Blocher, "SmartKom: Multimodal Communication with a Life-Like Character", in *EUROSPEECH*, Aalborg, 2001.