EVALUATING LONG-TERM SPECTRAL SUBTRACTION FOR REVERBERANT ASR

David Gelbart and Nelson Morgan

International Computer Science Institute,
and the EECS Department at the University of California at Berkeley
{gelbart,morgan}@icsi.berkeley.edu

ABSTRACT

Even a modest degree of room reverberation can greatly increase the difficulty of Automatic Speech Recognition. We have observed large increases in speech recognition word error rates when using a far-field (3-6 feet) mic in a conference room, in comparison with recordings from head-mounted mics. In this paper, we describe experiments with a proposed remedy based on the subtraction of an estimate of the log spectrum from a long-term (e.g., 2 s) analysis window, followed by overlap-add resynthesis. Since the technique is essentially one of enhancement, the processed signal it generates can be used as input for complete speech recognition systems. Here we report results with both HTK and the SRI Hub-5 recognizer. For simpler recognizer configurations and/or moderate-sized training, the improvements are huge, while moderate improvements are still observed for more complex configurations under a number of conditions.

1. INTRODUCTION

When speech is recorded in a room, the effects of reverberation (along with air absorption and mic response) create a channel response that distorts the speech spectrum.

If the channel is assumed to be a linear time-invariant system, the received signal spectrum X(ω) is equal to the product of the speech spectrum S(ω) and the channel spectrum C(ω). In practice, processing is based on a short-term Fourier transform X(n, ω) where n is the time index around which a windowed DFT is taken. If the analysis window is long and smooth enough then the product property still approximately holds: X(n, ω) $\approx$ S(n, ω) C(ω) ([1]). (Here the channel has been assumed not to vary over time.) Taking the logs of both sides, we find that log X(n, ω) $\approx$ log C(ω) + log S(n, ω), and thus in theory we could approximately remove C(ω), along with any constant portion of the speech spectrum, by subtracting the time average over n of log X(n, ω) from log X(n, ω). This logic is also the basis for cepstral mean subtraction, which is commonly used to counteract the effects of a time-invariant coloration such as fixed channel frequency response. However, in the latter case, the relevant time constants can be measured in milliseconds so that a short-term (e.g., 20 ms) analysis window can be used. For room reverberation the typical time constants are closer to a second, so that much longer analysis windows are required.

Using these concepts, the authors of [2] implemented a speech recognition front end based on mean subtraction using a long-term (2 second) spectral analysis window. Simulating reverberation using a fixed channel response, they found that subtracting the mean of the log magnitude spectrum improved ASR performance.

Reducing the effects of reverberation under realistic conditions may be quite different. For instance, speakers may move, and additive noise may complicate the degradation. Additionally, we would like to understand the interaction between the proposed approach and the complexity of the recognition system, both for training and test. For these reasons, we have experimented with artificial and natural reverberation, with simple and complex recognizers, and with small and large training sets.

2. METHODS

2.1. Mean subtraction implementation

In order to use the mean subtraction method with existing ASR systems, we wrote a separate program which produced resynthesized audio after performing spectral analysis and mean subtraction. The resynthesized audio was then given to ASR systems as a regular audio file.

Spectral analysis was performed using a Hanning-windowed N-point DFT stepped by N/4 samples. Except when otherwise stated, we used a 2.048-second analysis window, which corresponds to 16384 points at the sampling rate of 8000 Hz that we used for all our experiments.

As in [2] we chose to normalize the log magnitude spectrum while leaving the phase spectrum unchanged. Thus

This work was supported by the Natural Sciences and Engineering Research Council of Canada,the German Ministry for Education and Research, DARPA, and Qualcomm.

following spectral analysis the spectra were separated into phase spectra and magnitude spectra. For each analysis frame, the arithmetic mean of the log magnitude spectrum was calculated by averaging the log magnitude spectra of that frame and the previous W and next W frames. (W was 10 except when otherwise stated.) The mean calculated for each frame was then subtracted from it, and each resulting log magnitude spectrum was re-combined with the original phase spectrum. Resynthesis was then performed.

In order to simplify resynthesis and ensure an integer number of analysis frames, we duplicated data samples at the beginning and end of the data to add pad samples. After resynthesis the extra samples were discarded.

2.2. Test corpora

The experiments were carried out on connected digit strings. For evaluations on clean speech, a 9918-word test set from the TIDIGITS connected digits corpus ([3]) was used (downsampled to 8000 Hz and filtered to telephone bandwidth as described in [4]). For some tests, the data was artificially reverberated with a fixed impulse response, corresponding to an RT60 of 0.5 seconds and a direct-to-reverberant energy ratio of 0 dB.[1] In other cases, a large subset of TIDIGITS digit strings (7704 words) was read by speakers in a room we are using for recording natural meetings, and was recorded with headset mics and with a PZM table-mounted mic that was 3-6 feet from each of the talkers. This test set was collected as part of our Meeting Recorder project ([5]) and we will refer to it as Meeting Recorder Digits.

The TIDIGITS corpus is divided into short utterances, many of which are shorter than the analysis window length of 2.048 seconds. Therefore, we concatenated all utterances from the same speaker into one long vector of samples and performed mean subtraction using that vector, and then split the resynthesized output back into single-utterance files for use by the ASR system. This was done for training and test data. For the Meeting Recorder Digits corpus the same speakers re-appeared during different recording sessions, but we only concatenated utterances that were produced by the same speaker during the same session. This concatenation of segmented utterances has the effect of removing some of the inter-utterance silence, which may have made things easier for the mean subtraction method.

2.3. ASR systems

In our tests we first used the Aurora evaluation system described in [4], which is a Gaussian-mixture-based HMM system, HTK, configured to use word-level digit models.

[1] Thanks to Jim West and Gary Elko, of Bell Labs, and Carlos Avendano for providing the impulse response, which was measured experimentally in a varechoic chamber.

	Clean (TI)	Artificial reverb (TI)	Near (MR)	Far (MR)
Baseline	1.0%	19.2%	6.6%	41.4%
Mean sub.	1.2%	3.6%	3.3%	7.9%

Table 1. Results with Aurora baseline system, testing on TIDIGITS, artificially reverberated TIDIGITS, and near and far mic Meeting Recorder Digits.

In later tests, we used SRI's Hub-5 ASR system, which is described in [6]. This incorporates both speaker adaptation and a broader range of context-dependence than the configuration of HTK that we were using. For simplicity's sake, though, we ran the recognizer in a single pass without any rescoring, using only within-word triphone acoustic models. The language model of the Hub-5 system was replaced with a loop over all digit words with equal probabilities. Gender detection was not used—the system was only trained and tested on male speakers, in order to minimize training time. We used the speaker adaptation selectively so that we could observe its effects, and in particular any interaction with the log spectral subtraction technique. Speaker adaptation was unsupervised and used maximum likelihood linear regression on a phone-loop model (over all phones found in digits) to adjust the means of the Gaussians in the acoustic model via three affine transformations. (This is the first adaptation stage in the SRI system; subsequent adaptation to recognition hypotheses was omitted.)

3. EXPERIMENTAL RESULTS

3.1. Core experiments with HTK

To train the Aurora evaluation system we used the Aurora clean training set, consisting of 4220 male and 4220 female utterances from the training portion of the TIDIGITS corpus, downsampled to 8000 Hz and filtered to telephone bandwidth. For mean subtraction tests we performed mean subtraction on the training set as well as the test data.

Table 1 contrasts the word error rate of the baseline system with the system's performance when mean subtraction was performed. The first column shows the results for the TIDIGITS test data. The second column shows the results for the same data with the artificial reverberation described above. The third and fourth columns show word error rate results for Meeting Recorder Digits data (7704 words) collected with headset mics and the tabletop mic respectively.

It can be seen that the mean subtraction causes a small increase in error rate in the first column, but provides dramatic drops in error rates in the second and fourth columns. Performance in the third column also improves, despite being based on speech from a close-talking mic.

Time	Near	Far
3.072 s	4.6%	11.2%
4.096 s	4.5%	10.3%
5.120 s	4.2%	9.3%
6.144 s	3.9%	8.9%
7.168 s	3.8%	8.7%
10.240 s	3.6%	8.0%
12.288 s	3.3%	7.9%
13.312 s	3.3%	7.9%
14.336 s	3.3%	7.9%
15.360 s	3.2%	7.8%
16.384 s	3.2%	7.9%
17.408 s	3.1%	8.0%

Table 2. Word error rate for headset (near) and tabletop (far) mics, given the length of time over which the log spectral mean was calculated.

3.2. Length of time used to estimate mean

Our mean subtraction implementation uses a sliding window of 2W+1 frames to estimate the mean, and in real-world conditions (as opposed to simulated reverberation) one might expect a trade-off between better ability to estimate a fixed channel response with a larger W and to track changes in the channel response with a smaller W. However, for the Meeting Recorder Digits we found a minimum best value of W (W=10, corresponding to 12.288 seconds of data) but no clear maximum. Perhaps this is because there was not much speaker movement during the digit readings. Table 2 shows our results.

3.3. Analysis window length

Similarly, we wanted to explore other window lengths to see how far our original choice was from optimum. In particular, we wanted to confirm our intuition that for the case of reverberation, the usual 20-30 ms analysis window was insufficient.

Table 3 shows our results. The first column gives analysis window length, and the second and third columns give word error rates for near and far mic Meeting Recorder Digits. For these experiments, we used 12.288 seconds of data to estimate the mean.

Clearly, shorter analysis windows are insufficient for this task, and window lengths in the 1-2 second range appear to work well.

3.4. Follow-up with the SRI Hub-5 system

While the results from the Aurora evaluation system showed the mean subtraction causing dramatic gains in performance, the baseline results were often poor. For this reason we

Analysis Window	Near	Far
0.032 s	7.0%	38.4%
0.256 s	3.4%	14.6%
0.512 s	3.0%	8.8%
1.024 s	3.0%	7.9%
2.048 s	3.3%	7.9%
4.096 s	3.8%	8.4%

Table 3. Word error rate for headset (near) and tabletop (far) mics, given a range of analysis window lengths.

	Small training set	Large training set
Artificial reverb	16.3/9.7	9.2/4.1
Natural far mic	12.7/5.1	4.8/3.0

Table 4. Word error rates in percent for the baseline SRI recognizer without mean subtraction. Results without/with speaker adaptation are before/after the slashes. The test sets were artificially reverberated TIDIGITS and tabletop-mic Meeting Recorder Digits.

also experimented with SRI's Hub-5 ASR system, for which the number of parameters automatically increases when the amount of training data is increased, and which also incorporates other enhancements (e.g., speaker adaptation) that were not available in the Aurora configuration of HTK. [2] The same test sets and mean subtraction parameters were used as for the tests described in section 3.1 above.

Initially, we used the TIDIGITS training set, as was used for the HTK experiments; in this case, however, we used male speaker data (4235 utterances) from the original TIDIGITS set, rather than the Aurora version. We also tried training on a larger set consisting of 11.1 hours from the Switchboard and Callhome-English conversational speech corpora ([7], [8]) and 21.1 hours from the Macrophone read speech corpus ([9]), all from males.

The baseline word error rates for the two kinds of degradations are given in table 4. All of these results were significantly better than for the original experiments.[3]

Table 5 shows the percent reduction in WER resulting from the mean log spectral subtraction technique. Reductions in error rate (some of which are quite large) can be seen for all but one case.

[2] Unlike the Aurora baseline system, the SRI system uses short-term cepstral mean subtraction in its front end. Therefore any gains in performance due to long-term mean subtraction with the SRI system are in addition to the short-term cepstral mean subtraction.

[3] The results given here are not directly comparable, as they are performed on somewhat different, male-only, test and training sets. Nonetheless, using the baselines shown here might be closer to what would be relevant for a well-trained modern system.

	Small training set	Large training set
Artificial reverb	69/68	34/15
Natural far mic	50/18	6/-2

Table 5. Relative improvement in percent using the mean subtraction technique, without/with speaker adaptation.

4. DISCUSSION AND CONCLUSIONS

The results reported in this paper generally support the notion that long-term log spectral subtraction can help with ASR degradation due to room reverberation. We found that the subtraction could improve performance on data collected in realistic acoustic conditions, where the channel response may vary over time due to speaker movement, and that the use of the long analysis window was important.

However, our results also reinforce the common observation that it is easier to improve a poor system than a relatively good one. Speaker adaptation and the use of a large diverse training set (recorded over many different telephones with a wide variation in noise and channel characteristics) improves performance well enough for the Meeting Recorder digits that the subtraction technique provides no improvement, and in fact appears to hurt performance slightly. However, if the degradation matches the assumptions of the mean subtraction technique (purely artificial reverberation) or if only a smaller training set is available, the improvement can be large, even if adaptation is used. Furthermore, in many applications a system with many parameters might not be feasible, for instance due to memory constraints in a portable device. And if the number of parameters are too limited, the system might not be able to take advantage of much more training data, and results might be closer to what we have observed for the small training set.

Of course, we have only tested one "natural" room in this experiment, and it may well be that results would differ in general. In particular, rooms and mic/talker placements will yield both differing reverberation characteristics (e.g., decay time and direct-to-reverberant ratios), and also different SNRs. The NIST stnr tool ([10]) yields a SNR for the Meeting Recorder Digits (far mic) of 9.0 dB, which suggests that our purely convolutional degradation model might be a poor match to our test conditions. It is clear to us that we need to have the model incorporate additive noise as well. This is a goal for future work.

5. ACKNOWLEDGEMENTS

In addition to the sponsors acknowledged earlier, the authors would like to thank a number of helpful colleagues; in particular, Andreas Stolcke (and SRI in general) for all of the assistance with the SRI recognition software. Hynek Hermansky and Carlos Avendano provided us with the core methodology that we extended in this work. Stephane Dupont and Barry Chen helped us with the Aurora evaluation system. Adam Janin and others involved with the Meeting Recorder project provided the Meeting Recorder Digits corpus. Dan Ellis provided his signal processing insights.

6. REFERENCES

[1] C. Avendano, *Temporal Processing of Speech in a Time-Feature Space*, Ph.D. thesis, Oregon Graduate Institute, 1997.

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

[3] R. G. Leonard and G. Doddington, "A Database for Speaker-Independent Digit Recognition," in *ICASSP 1984*, San Diego, CA, USA, 1984.

[4] 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.

[5] "The ICSI Meeting Recorder Project," http://www.icsi.berkeley.edu/Speech/mr/.

[6] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 Conversational Speech Transcription System," in *NIST Speech Transcription Workshop*, College Park, Maryland, USA, 2000.

[7] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *ICASSP 1992*, San Francisco, CA, USA, 1992.

[8] Linguistic Data Consortium, "CALLHOME American English Speech," 1997.

[9] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: An American English telephone speech corpus for the Polyphone project," in *ICASSP 1994*, Adelaide, Australia, 1994.

[10] "NIST Speech Quality Assurance Package Version 2.3," http://www.nist.gov/speech/tools/.