DOUBLE THE TROUBLE: HANDLING NOISE AND REVERBERATION IN FAR-FIELD AUTOMATIC SPEECH RECOGNITION

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

Far-field microphone speech signals cause high error rates for automatic speech recognition systems, due to room reverberation and lower signal-to-noise ratios. We have observed large increases in speech recognition word error rates when using a far-field (3-6 feet) microphone in a conference room, in comparison with recordings from close-talking microphones. In an earlier paper, we showed improvements in far-field speech recognition performance using a long-term log spectral subtraction method to combat reverberation. This method is based on a principle similar to cepstral mean subtraction but uses a much longer analysis window (e.g., 1 s) in order to deal with reverberation. Here we show that a combination of short-term noise filtering and long-term log spectral subtraction can further reduce recognition word error rates.

1. INTRODUCTION

When speech is recorded in a room with a far-field microphone, the received signal is distorted by acoustic reverberation, since it consists of not only the direct path signal but also delayed and filtered versions of the signal (due to multiple reflections). Since reverberation has a long time characteristic in comparison to the typical speech analysis frame, this distortion is manifested as a temporal smearing of the short-term spectra that are used as the basis for speech recognition features. Additionally, the lower level of received speech energy at a far-field microphone makes the ambient additive noise much more significant than it would be for a closer microphone.

When stereo information is available (i.e., both near and far microphones), there are many methods available to improve the ASR accuracy for the far-field signal. Similarly, when comparable training data is available (or can be synthesized given known information about the target environment, as in [1]), existing methods can yield greatly improved word error rates. However, in many plausible scenarios, it will not be feasible to obtain training data that is comparable to the test data (either through collection or transformation) or to have a close-talking microphone. For this reason, we have been focusing on improving recognition for the mismatched case in which the training data is relatively clean (low noise and reverberation), and the test data has a realistic level of both aspects of acoustic environmental degradation.

Previously [2], we reported results using a long-term log spectral subtraction method that is based on a purely convolutional model, not incorporating additive noise. Modeling reverberation as a fixed linear time-invariant filter, the received signal spectrum $X(\omega)$ is equal to the product of the speech spectrum $S(\omega)$ and the filter spectrum $C(\omega)$. In practice, processing is based on a short-term Fourier transform $X(n,\omega)$ 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 [3]: $X(n,\omega) \approx S(n,\omega)\,C(\omega)$. Taking the logs of the magnitudes of both sides, we find that $\log |X(n,\omega)| \approx \log |C(\omega)| + \log |S(n,\omega)|$, and thus in theory we can approximately remove $|C(\omega)|$, along with any constant portion of the log speech spectrum, by subtracting the time average over n of $\log |X(n,\omega)|$ from $\log |X(n,\omega)|$. We are calling this approach long-term log spectral (mean) subtraction. 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 a telephone 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 hundreds of milliseconds (or more), so we used much longer analysis windows.

The authors of [4] found that long-term log spectral subtraction improved ASR performance, testing on simulated reverberant test data. In [2] we tested on actual far-field test data and found that the approach often produced substantial performance improvements. Furthermore, in that work we

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

found that the use of an analysis window longer than 32 ms (i.e., longer than typically used for cepstral mean subtraction) was necessary for optimal performance. [1]

The error rates for far-field microphone speech were still far worse than were obtained from the corresponding close microphones. There could be many sources of this increased error, but the most obvious deficiency in our convolutional model was the lack of an additive noise term. A compensating step in our enhancement-based approach is to apply a noise reduction module prior to log spectral mean subtraction. While not necessarily optimal (since each compensatory step is imperfect), the concatenation of two existing approaches to handle the combined sources of error was a reasonable first step. Figure 1 shows our model for the room acoustic effects followed by the compensatory steps explored here.

2. METHODS

2.1. Noise reduction implementation

We used a noise reduction algorithm developed for an Aurora 2 front end proposal, a joint effort between ICSI, OGI, and Qualcomm engineers, described separately in [5]. The algorithm performs Wiener filtering with typical engineering modifications such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor. We modified the algorithm to use a single noise spectral estimate for each utterance, which was calculated over all the frames judged to be nonspeech by the voice-activity detection component of the Qualcomm-ICSI-OGI front end. We invoked it independently for each utterance and used overlap-add resynthesis to create noise-reduced output waveforms, which we either gave directly to the ASR system or passed on to the mean subtraction algorithm.

2.2. Log spectral subtraction implementation

The log spectral subtraction was implemented by a standalone program which read in waveforms, performed spectral analysis and subtraction, and produced output waveforms by overlap-add resynthesis. The output waveforms were then given to the ASR system. Besides allowing the algorithm to be used with existing ASR software, the use of resynthesis is also a way of dealing with the mismatch in time scale between the long-term analysis window and the much shorter analysis windows used in ASR front ends.

Spectral analysis was performed using a Hanning-windowed N-point DFT stepped by N/4 samples. In all experiments we used a 1.024-second analysis window (N=8192 points at our 8000 Hz sampling rate). 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 22 and next 22 frames (thus the mean was calculated over a total of 12.288 s of input data). The mean calculated around each frame was then subtracted from the log magnitude spectrum of that frame, and the result was then re-combined with the original phase spectrum of that frame.

In order to simplify resynthesis and ensure an integer number of analysis frames, we padded each long-time frame with duplicate data samples at the beginning and end. After resynthesis the extra samples were discarded.

2.3. Training and test corpora

The experiments were carried out on connected digit strings. The training set consisted of 4220 male and 4220 female utterances from the TIDIGITS [6] training set, downsampled to 8000 Hz. This set was the same as the Aurora clean training set described in [7] except that we omitted the G.712 telephone bandwidth filtering.

For test data, a large subset of TIDIGITS digit strings (7704 words) was read by native English speakers seated around a conference table in a room we are using for recording natural meetings. Simultaneous recordings were made with close-talking mics and with a table-mounted PZM mic that was 3-6 feet from each of the talkers. This test set was collected as part of our Meeting Recorder project [8] and we will refer to it as Meeting Recorder Digits. The far-field waveforms have a significant amount of background noise—NIST's stnr tool [9] reports a 9.0 dB average SNR.

The TIDIGITS and Meeting Recorder Digits corpora are segmented into single utterances. To improve the mean estimates over what could be obtained from relatively short utterances, we concatenated all utterances from the same speaker into one long vector of samples prior to using the mean subtraction algorithm, and then split the resynthesized output back into single-utterance files for use by the ASR system. 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 could potentially improve performance for the mean subtraction method.

In the tests presented here, we applied the noise reduction and log spectral subtraction algorithms to the training set as well as to the test set.

2.4. ASR system

We used the Aurora reference system described in [7], which is a Gaussian-mixture-based HMM recognizer (HTK) configured to use word-level digit models with 16 states per

[1] However, in recent experiments we have seen that we can greatly improve the performance of the 32 ms window with further processing.

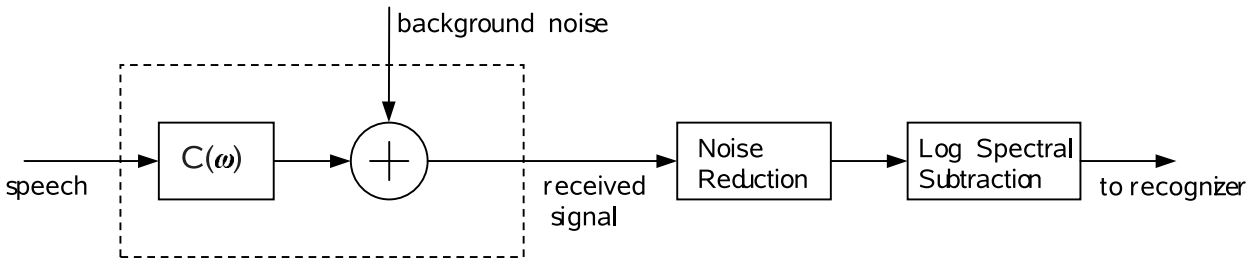

Fig. 1. **Model of acoustic degradation in the far-field signal and our compensatory processing. The speech signal is filtered by the "room response" $C(\omega)$ between the talker and the microphone. Background noise also arrives at the microphone. The resulting received signal is processed by noise removal followed by long-term log spectral subtraction.**

	Near	Far
Baseline	4.1%	26.3%
Noise reduction	3.6%	24.8%
Log spectral subtraction	3.1%	8.2%
Noise r. + Log spectral s.	2.7%	7.2%

Table 1. **Word-error rate results for the Aurora reference system combined with noise reduction and long-term log spectral subtraction algorithms. For comparison, this system has 1.0% WER when training and testing on G.712-filtered TIDIGITS [7].**

word and three Gaussians per state. (Pauses are modeled using fewer states and six Gaussian per state.) The reference system front end uses mel frequency cepstral coefficients with log frame energy and first- and second-order delta features.

3. EXPERIMENTAL RESULTS

Table 1 gives the results in terms of word error rate. The first and second columns show word error rate results for Meeting Recorder Digits data (7704 words) collected with close-talking ("near") mics and the tabletop ("far") mic respectively. The rows correspond, respectively, to the baseline system, noise reduction only, log spectral subtraction only, and noise reduction followed by log spectral subtraction.

Clearly, both noise reduction and log spectral subtraction individually improve performance for both near and far microphones, and log spectral subtraction is by far the most useful in the far case. Combining them gives a further improvement in performance.

The noise reduction alone improves far-field performance by similar absolute amounts whether done in combination with log spectral subtraction (1.0% absolute reduction) or alone (1.5% absolute reduction). While the SNR is moderately low for these data, it could still be that for the connected digits task the reverberation per se is the larger problem.

However, there were also limitations to the efficacy of the noise reduction. In the far mic case, for 49 out of the 2350 utterances the voice-activity detection classified none of the frames as nonspeech, and the noise reduction module defaulted to a noise spectral estimate of zero. In this case it performed no enhancement other than applying its DC offset compensation filter. This happened for 741 of the utterances in the near case, perhaps due to the occurence of breath and other noises in the near mic data.

4. DISCUSSION AND CONCLUSIONS

As noted in [2], we have found that long-term log spectral subtraction can help with ASR degradation due to room reverberation. However, the remaining error rates after log spectral subtraction are still quite significant, and are much greater than are observed for near-field microphone placements. It is likely that at least some of this error is due to additive noise. The experiments reported here show that this appears to be true, in that Wiener filtering, a sensible approach to reducing additive stationary noise, eliminates some of the errors. This error reduction is observed whether the log spectral mean subtraction is done or not. When these two techniques were used in combination, the ratio of word error rates for far and near conditions was reduced by over 50%. However, this resulting ratio was still 2.7, suggesting that we have a long way to go to bring far-field performance close to the near-field case.

It is likely that the simple concatenation of processing steps described here is suboptimal, and that a better system could be devised by jointly optimizing for both noise and reverberation. In the case of short-time convolutional effects this has been done, for instance, in J-RASTA [10]. Additionally, it may not be possible to completely compensate

for the room acoustic differences between near and far microphone performance when the "near" signal is not available, and when the "far" signal is not spatially enhanced with array techniques. However, we have also found in the past that modifying the feature extraction to yield features that are more invariant to reverberation and noise can help ASR performance [11]. In future work we intend to combine such techniques with the more explicit approach of dereverberation and denoising described here.

Finally, we are currently working to adapt these techniques to a near real-time application, a demonstrator for the SmartKom project [12]. In this case, we will need to make use of coarser estimates of room acoustics in the earlier parts of a session [13].

5. ACKNOWLEDGEMENTS

In addition to the sponsors acknowledged earlier, the authors would like to thank a number of helpful colleagues; in particular, the other researchers involved in developing the Qualcomm-ICSI-OGI front end. Hynek Hermansky and Carlos Avendano provided us with the core methodology that we extended in our recent work. Adam Janin and others involved with the Meeting Recorder project provided the Meeting Recorder Digits corpus. Birger Kollmeier, Michael Kleinschmidt, and Barry Chen provided signal processing insights.

6. REFERENCES

[1] V. Stahl, A. Fischer, and R. Bippus, "Acoustic synthesis of training data for speech recognition in living room environments," in *ICASSP*, Salt Lake City, Utah, 2001, vol. 1, pp. 21–4.

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

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

[4] C. Avendano, S. Tibrewala, and H. Hermansky, "Multiresolution Channel Normalization for ASR in Reverberant Environments," in *EUROSPEECH*, Rhodes, Greece, 1997, vol. 3, pp. 1107–1110.

[5] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *ICSLP*, Denver, Colorado, 2002, (to be submitted to Aurora special session).

[6] R. G. Leonard, "A Database for Speaker-Independent Digit Recognition," in *ICASSP*, San Diego, California, 1984, vol. 3, p. 42.11.

[7] 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.

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

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

[10] H. Hermansky and N. Morgan, "RASTA Processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–89, 1994.

[11] B. Kingsbury, *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments*, Ph.D. thesis, University of California at Berkeley, 1998.

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

[13] D. Gelbart, "Reducing the Effect of Room Acoustics on Human-Computer Interaction," in *Applied Voice Input/Output Society Conference (AVIOS)*, San Jose, California, 2002, in press.