REVEREBERATION REDUCTION FOR IMPROVED SPEECH RECOGNITION

Ivan Tashev

Microsoft Research

One Microsoft Way
Redmond, WA 98052, USA
Email: ivantash@microsoft.com

Daniel Allred

Georgia Institute of Technology
School of ECE
777 Atlantic Drive NW
Atlanta, GA 30332, USA
Email: djallred@ece.gatech.edu

Abstract

In this paper we present a dereverberation algorithm for improving automatic speech recognition (ASR) results with minimal CPU overhead. As the reverberation tail hurts ASR the most, late reverberation is reduced via gain-based spectral subtraction. We use a multi-band decay model with an efficient method to update it in real-time. In reverberant environments the multi-channel version of the proposed algorithm reduces word error rates (WER) up to one half of the way between those of a microphone array only and a close-talk microphone. The four channel implementation requires less than 2% of the CPU power of a modern computer.

Introduction

The need to present clean sound inputs to today's speech recognition engines has fostered huge amounts of research into areas of noise suppression, microphone array processing, acoustic echo cancellation and methods for reducing the effects of acoustic reverberation.

Reducing reverberation through deconvolution (inverse filtering) is one of the most common approaches. The main problem is that the channel must be known or very well estimated for successful deconvolution. The estimation is done in the cepstral domain [1] or on envelope levels [2]. Multi-channel variants use the redundancy of the channel signals [3] and frequently work in the cepstral domain [4].

Blind dereverberation methods seek to estimate the input(s) to the system without explicitly computing a deconvolution or inverse filter. Most of them employ probabilistic and statistically based models [5].

Dereverberation via suppression and enhancement is similar to noise suppression. These algorithms either try to suppress the reverberation, enhance the direct-path speech, or both. There is no channel estimation and there is no signal estimation, either. Usual techniques are long-term cepstral mean subtraction [6], pitch enhancement [7], LPC analysis [8] in single or multi-channel implementation.

The most common issues with the preceding methods are slow reaction when reverberation changes, robustness to noise, and computational requirements.

Modeling and assumptions

We convoluted clean speech signal with a typical room response function and processed it trough our ASR engine, cutting the length of the response function after some point. The results are shown on Figure 1. The early reverberation practically has no effect on the ASR results, most probably due to cepstral mean subtraction (CMS) in the ASR engine front end. The CMS compensates for the constant part of the input channel response and removes the early reverberation. The reverberation has noticeable effect on WER between 50 ms and RT_{30}. In this time interval the reverberation behaves more as non-stationary, uncorrelated decaying noise $\Re(f)$:

$$Y(f) = X(f) + \Re(f) \qquad (1)$$

We assume that the reverberation energy in this time interval decays exponentially and is the same in every point of the room (i.e. it is diffuse). Our decay model is frequency dependent:

$$S_{\Re_n}(f) = \sum_{i=0}^{n-N} \alpha(f) S_{X_i}(f).e^{-\frac{iT}{\tau(f)}} = \alpha(f) S_{Y_{n-N}}(f) e^{-\frac{NT}{\tau(f)}}, \qquad (2)$$

where T is the frame duration, n is the current frame number, N is the number of frames where we do not want to suppress the reverberation (~50 ms/T), $\alpha(f)$ is proportional to signal-to-reverberation-ratio (SRR) and $\tau(f)$ is the decay time constant.

Model parameters estimation

Estimation of two decay parameters per frequency bin (α and τ) would consume too much CPU time and would need longer time for converging. Therefore we estimate the decay ratio and time constant in L frequency subbands, separated by cosine-shaped, 50% overlapping weight windows with logarithmically increasing width towards higher frequencies. The parameter estimation happens when we have a pure reverberation process: after the end of the word and only if the pause to the next word is longer than RT_{60}. We use a Gaussian probabilistic based speech/non-speech classifier [9]. We keep the energy in each subband for the last K=RT_{60}/T frames and interpolate with $S(k) = A\exp(-kT/\tilde{\tau}) + B, k \in [N, K]$. We have as unknowns A, B and $\tilde{\tau}$ and because K>3 we

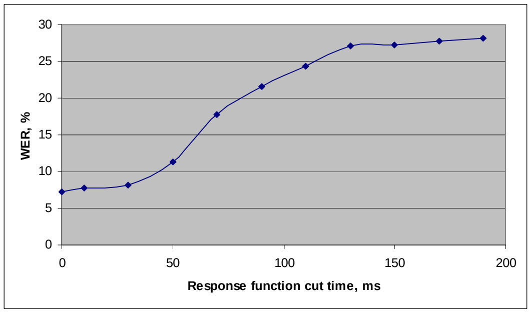

Figure 1. Late reverberation effect on ASR.

solve the over-determined non-linear system of equations using Gaussian minimization with minimum mean square error as criterion. Here B is the noise floor, $\tilde{\tau}$ is the decay time constant and the SRR parameter is computed as $\tilde{\alpha} = A / S_{Y_{n-N}}$. Estimated momentary parameters are reflected in the decay model:

$$\begin{aligned} \tau_n(l) &= \tau_{n-1}(l) + \frac{\tau_A}{T}\left[\tilde{\tau}(l) - \tau_{n-1}(l)\right] \\ \alpha_n(l) &= \alpha_{n-1}(l) + \frac{\tau_A}{T}\left[\tilde{\alpha}(l) - \alpha_{n-1}(l)\right] \end{aligned} \qquad (3)$$

where τ_A is the adaptation time constant and l is the frequency subband. The values of the decay model parameters for all frequencies are computed using linear interpolation between the L estimated points.

Reverberation reduction

Based on the assumption that the reverberation in the time interval of interest already behaves as non-correlated noise we use spectral subtraction for optimal, in the sense of minimum mean square error, reverberation reduction:

$$\tilde{X}(f) = \left| \begin{array}{l} \dfrac{S_Y(f) - \beta S_{\Re}(f)}{S_Y(f)} Y(f), when S_Y(f) > S_{\Re}(f) \\ \qquad (1-\beta) Y(f), otherwise \end{array} \right. \qquad (4)$$

Here $S_{\Re}(f)$ is estimated according to (2) and $\beta \in [0,1]$ is used to set the suppressed portion of the reverberation.

Results

This algorithm was implemented as microphone array preprocessor. The multi-channel implementation uses the same decay model for all channels, the SRR is estimated separately for each channel. We recorded a control set of 2700 words in anechoic chamber, office, conference room and lecture room, varying the reverberation time, the noise level and the distance. The source speech was generated by a B&K mouth simulator. For recording were used a close-up microphone and a 196 mm linear microphone array with four cardioid transducers. The system works with 16 kHz sampling rate, 16 bits precision and 20 ms audio frames. The decay model was estimated in four frequency subbands. The latest ASR from Microsoft was used. The results are shown in Table 1.

Table 1. WER in % for different conditions

Conditions	***Dist.***	***MA***	***MA+DR***	***Close-up***
Chamber	1.5 m	4.1	4.8	3.3
Office	1.2 m	7.1	5.9	4.2
Conf. room	2.5 m	15.7	12.1	4.1
Lecture room	3.2 m	12.4	8.1	6.7

Comments and conclusions

Reverberation reduction is non-liner processing and introduces some WER degradation in the anechoic chamber conditions where is no reverberation to remove. In the presence of reverberation it reduces the WER for microphone array plus dereverberation (MA+DR) up to one half of the way between a microphone array only (MA) and close-up microphone (Close-up). The designed dereverberation algorithm is computationally efficient: the four channel implementation uses less than 2% CPU time on modern computers.

References

[1] D. Bees, M. Blostein, and P. Kabal, “Reverberant speech enhancement using cepstral processing”, in Proc. *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 977-980, 1991.

[2] J. Mourjopoulos and J.K. Hammond, “Modeling and enhancement of reverberant speech using an envelope convolution method”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1144-1147, Boston, MA, USA, 1983.

[3] J. Liu and H.S. Malvar, “Blind deconvolution of reverberated speech signals”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, pp. 3037-3040, Salt Lake City, UT, 2001.

[4] Subramaniam, A.P. Petropulu, and C. Wendt, “Cepstrum-based deconvolution for speech dereverberation”, *IEEE Trans. Speech and Audio Processing*, 4(5):392-396, Sept. 1996.

[5] H. Attias, J.C. Platt, A. Acero, and L. Deng, “Speech denoising and dereverberation using probabilistic models”, in Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 13*, MIT Press, Cambridge, MA, 2001.

[6] D. Gelbart and N. Morgan, “Double the trouble: Handling noise and reverberation in far-field automatic speech recognition”, in *Proc. of ICSLP*, Denver, Colorado, USA, Sept. 2002.

[7] M. Wu and D.L. Wang, “A one-microphone algorithm for reverberant speech enhancement”, in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 844-847, 2003.

[8] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering”, in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, pp. 3701-3704, 2001.

[9] J. Sohn et. al., ``Statistical Model Based Voice Activity Detector", IEEE Signal Processing Letters, Vol. 6, No. 1,pp. 1-3, Jan 1999.