SUPPRESSION RULE FOR SPEECH RECOGNITION FRIENDLY NOISE SUPPRESSORS

Ivan Tashev, Jasha Droppo, Alex Acero

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Abstract: Audio signal enhancement often involves the application of a time-varying filter, or suppression rule, to the frequency-domain transform of a corrupted signal. Known approaches use rules derived under Gaussian models and interpret them as spectral estimators in a Bayesian statistical framework. While this mathematical approach provides rules that satisfy certain optimization criteria these rules are not optimal when the enhanced signal is for a speech recognition engine. In this paper we present the approach and the results for creation of a speech recognition friendly suppression rule. The described approach increases the average speech recognition rate in Aurora 2 tests from 52.47% to 77.69% while maintaining performance for low noise utterances.

Key words: Noise reduction, Audio enhancement, Speech recognition

Introduction

In this paper we address an important issue in audio signal processing, that of creation of noise robust speech recognition engines. Due to its ubiquity in applications of this nature, we concentrate on short-time spectral attenuation, a popular method of broadband noise reduction in which a time-varying filter, or suppression rule, is applied to the frequency-domain transform of a corrupted signal. We first address existing suppression rules derived under a Gaussian statistical model and interpreted in a Bayesian framework. We then derive a new speech recognition friendly suppression rule using multidimensional optimization methods.

To date, the most popular methods of broadband noise reduction involve the application of a time-varying filter to the frequency-domain transform of a noisy signal. Let $x_n = x(nT)$ in general represent values from a finite-duration analog signal sampled at a regular interval *T*, in which case a corrupted sequence may be represented by the additive observation model $y_n = x_n + d_n$, where y_n represents the observed signal at time index *n*, x_n is the original signal, and d_n is additive random noise, uncorrelated with the original signal. The goal of signal enhancement is then to form an estimate $\hat{x}_n$ of the underlying signal x_n based on the observed signal y_n. In many implementations where efficient on-line performance is required, the set of observations $\{y_n\}$ is filtered using the overlap-add method of short-time Fourier analysis and synthesis, in a manner known as short-time spectral attenuation. Taking the discrete Fourier transform on windowed intervals of length *N* yields *K* frequency bins per interval: $\boldsymbol{Y}_k = \boldsymbol{X}_k + \boldsymbol{D}_k$, where these quantities are complex. Noise reduction in this manner may be viewed as the application of a suppression rule, or nonnegative real-valued gain H_k, to each bin *k* of the observed signal spectrum $\boldsymbol{Y}_k$, in order to form an estimate $\hat{X}_k$ of the original signal spectrum: $\hat{X}_k = H_k \cdot Y_k$. This spectral estimate is then inverse-transformed to obtain the time-domain signal reconstruction.

Within such a framework, a simple Gaussian model often proves effective [1]. In this case the elements of $\{X_k\}$ and $\{D_k\}$ are modeled as independent, zero-mean, complex Gaussian random variables with variances $\lambda_x(k)$ and $\lambda_d(k)$, respectively: $X_k \sim \mathrm{N}_2(0, \lambda_x(k)\mathrm{I})$, $D_k \sim \mathrm{N}_2(0, \lambda_d(k)\mathrm{I})$.

A frequent goal in signal enhancement is to minimize the mean-square error of an estimator; within the framework of Bayesian risk theory, this MMSE criterion may be viewed as a squared-error cost function. Considering the corrupted signal model, the Bayes' rule, and the prior distributions defined above, the optimal suppression rule in an MMSE sense is $H_k = \dfrac{\lambda_x}{\lambda_x + \lambda_d}$, which is recognizable as the well-known Wiener filter [2]. Later McAulay and Malpass [3] derive a maximum-likelihood (ML) spectral amplitude estimator under the assumption of Gaussian noise and an original signal characterized by a deterministic waveform of unknown amplitude and phase. As an extension of the underlying model, Ephraim and Malah [4] derive a minimum mean-square error (MMSE) short-time spectral amplitude estimator based on the assumption that the Fourier expansion coefficients of the original signal and the noise may be modeled as statistically independent, zero-mean, Gaussian random variables. They introduce the *a priori* and *a posteriori* signal-to-noise ratios (SNR) as $\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}$ and $\gamma_k \triangleq \frac{|Y_k|^2}{\lambda_d(k)}$ respectively. Their suppression rule is a function of these two SNRs: $H_k = f(\xi_k, \lambda_k)$. The success of the Ephraim and Malah suppression rule is largely due to the authors' decision-directed approach for estimating the a priori SNR ξ_k. For a given audio frame *n*, the decision-directed *a priori* SNR estimate $\hat{\xi}_k$ is given by a geometric weighting of the SNR in the previous and current frames:

$$\hat{\xi}_k = \alpha \frac{\left|\hat{X}_k(n-1)\right|}{\lambda_d(n-1,k)} + (1-\alpha)\max\left[0, \gamma_k(n)-1\right], \alpha \in [0,1) \cdot$$

The spectral amplitude estimator given by Ephraim and Malah, while being optimal in an MMSE sense, requires the computation of exponential and Bessel functions. Wolfe and Godsil [1] derive the analytic form of three alternative suppression rules under the same model, each of which admits a more straightforward implementation. Otherwise, they are more or less close to the Ephraim and Malah suppression rule. Each one of these suppression rules is optimal in some sense, but none of them is optimized for best speech recognition results. Each of the suppression rules introduce distortions that damage speech features important for speech recognition.

Speech recognition friendly suppression rule

A generalized noise suppressor can be represented as $\hat{X}_k = \arg(Y_k).\left[H_k(\xi_k, \gamma_k)\cdot\left|Y_k\right|\right]$, where H_k is the suppression rule. It can be parameterized as a square *L*x*L* matrix, where the working range of values for ξ_k and λ_k are presented as *L* discrete values. Intermediate values can be obtained by using interpolation methods.

In this case the derivation of the suppression rule is converted to optimization problem with L^2 optimization parameters – the values of H_k in the matrix. To optimize the parameters of the noise suppression rule, we maximize an objective function that is closely related to speech recognition accuracy. The objective function used in this paper is maximum mutual information (MMI), which is well known to be correlated with speech recognition accuracy [6].

To optimize the suppression rule parameters H_k with respect to the MMI objective function *F*, we use the Rprop algorithm [5]. At each iteration, this algorithm requires the value of the objective function at the current parameter values, and the gradient of the objective function with respect to each parameter at the current parameter values. It can be shown [7] that the desired gradients are equal to: $\frac{dF}{d\theta} = \sum_i \frac{dF}{dx_i}\frac{dx_i}{d\theta}$, where x_i are the speech recognition features for the current frame and θ is a parameter from H_k. The gradient $\frac{dx_i}{d\theta}$ of the features with respect to the noise suppression parameters was computed by using a discrete approximation to the derivative function. The gradient $\frac{dF}{dx_i}$ and the sum can be computed as demonstrated previously [7]. The optimization starting point was initialized with the Minimum Mean Square Error Spectral Power Estimator rule [1], chosen based on superior recognition results compared to other rules.

All of the experiments reported in this paper were performed under the Aurora 2 noise robust speech recognition framework [8]. The acoustic model is trained with 8,440 noise-free utterances, using the “complex” backend training scripts. The test set consists of the 20,020 utterances in Set A, which were constructed by artificially mixing 4,004 clean utterances with four recorded noise types at 0, 5, 10, 15 and 20 dB SNR. All audio files have an 8,000 Hz sampling rate.

Results

Here we discuss the average recognition rate, obtained using all test data and clean recognition rate, obtained by using only clean speech data. Without noise suppression, the average recognition rate was 52.47%, and the recognition rate on the clean signals was 99.53%. Table 1 presents the speech recognition results when noise suppression is employed. For noisy speech, the baseline (with the MMSE noise suppression rule) performs much better (74.90%) than the reference system (52.47%). But, for clean speech it has a much lower accuracy (96.88%) than the reference (99.53%). Further optimization of the noise suppression parameters reduces the average number of errors by 11% relative, and the number of clean condition errors has dropped by 69% relative. Figure 1 shows the initial suppression rule, the final suppression rule and the difference between the two.

Table 1. Speech recognition results

Iteration	Average	Clean
0	74.90	96.88
5	76.20	97.55
10	77.26	98.45
15	77.49	99.04
20	77.69	99.02
25	77.82	98.66

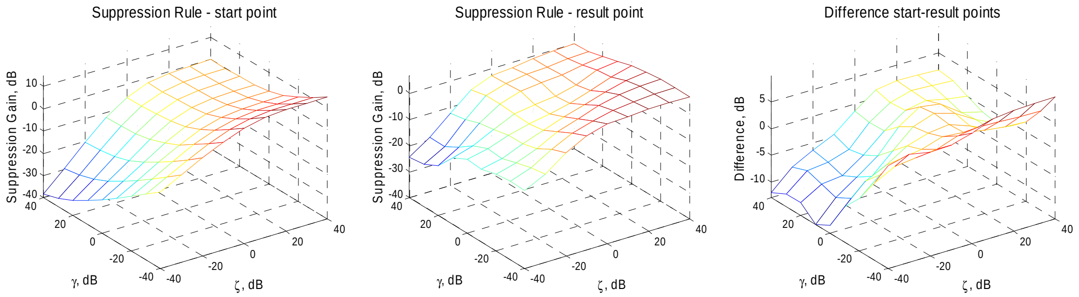

Figure 1. Initial and final suppression rules and the difference between them.

Conclusions

The presented approach for finding a suppression rule that is optimal from a speech recognition rate perspective substantially improves the speech recognition results for noisy speech, without serious degradation of the clean speech results. The suppression rule is computationally efficient and suitable for real-time applications. The speech recognition rate for Aurora 2 tests is increased from 52.47% to 77.69% while keeping the clean speech results degradation minimal (from 99.53% to 99.02 %).

Literature

[1] P. J. Wolfe and S. J. Godsill, “Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement,” In Proceedings of the IEEE Workshop on Statistical Signal Processing, pages 496-499, 2001.

[2] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications, Principles of Electrical Engineering Series. MIT Press, Cambridge, MA, 1949.

[3] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 2, pp. 137{145, 1980.

[4] Y. Ephraim, D. Malah. “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6, December 1984.

[5] M. Reidmiller and H. Braun, “A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm,” in IEEE International Conference on Neural Networks, 1993, vol. 1, pp. 586-91.

[6] V. Valtchev, Discriminative Methods in HMM-based Speech Recognition, Ph.D. thesis, CMU, 1987.

[7] J. Droppo, M. Mahajan, A. Gunawardana and A. Acero, “How to Train a Discriminative Front End with Stochastic Gradient Descent and Maximum Mutual Information,” in Proceedings IEEE Automatic Speech Recognition and Understanding Workshop, 2005.

[8] 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.