Adaptive Noise Reduction of Speech Signals

Wenqing Jiang and Henrique Malvar

July 2000

Technical Report
MSR-TR-2000-86

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Adaptive Noise Reduction of Speech Signals

Wenqing Jiang and Henrique Malvar

Abstract

We propose a new adaptive speech noise removal algorithm based on a two-stage Wiener filtering. A first Wiener filter is used to produce a smoothed estimate of the *a priori* signal-to-noise ratio (SNR), aided by a classifier that separates *speech* from *noise* frames, and a second Wiener filter is used to generate the final output. Spectral analysis and synthesis is performed by a modulated complex lapped transform (MCLT). For noisy speech at a low 10 dB input SNR, for example, the proposed algorithm can achieve on average about 13 dB noise-to-mask ratio (NMR) reduction, or about 6 dB SNR improvement.

1 Introduction

Noise removal is a necessary preprocessing step for speech acquisition in computer telephony and other applications, such as speech-assisted human-computer interfaces. Office noise from fans and computers, as well as vehicle noise, not only degrades the subjective speech quality, but it also hinders performance of speech coding and recognition systems.

Many approaches have been reported in the literature for speech noise reduction, such as the short-time spectral amplitude estimator in [1, 2], the signal subspace approach in [3] and the human auditory system model-based approaches in [4] and [5]. In this paper, we focus our study on short-time spectrum attenuation techniques, which have been shown to be very effective and simple for low cost implementations [1, 2, 6].

A typical spectrum attenuation technique, assuming an additive uncorrelated noise model, consists of two basic steps [7]: (i) estimation of noise spectrum and (ii) filtering of the noisy speech to obtain the cleaned speech. In spectral subtraction systems, a noise spectral magnitude estimate is actually subtracted from the signal magnitude spectrum. That can lead to larger amounts of noise reduction. Both approaches are usually effective, but they can generate artifacts known as *musical noise*[1] [6], especially in spectral subtraction systems. Approaches to reduce musical noise include using sophisticated speech/noise classification mechanisms, such as the cepstral detector by Sovka et al. [8], the pitch-based detector by Tucker et al. [9], and the multiple features-based voice activity detector (VAD) in G.729 by Benyassine et al. [10].

[1] The residual noise composed of sinusoidal components with random frequencies that come and go in each short-time frame. It is caused by the mismatch between the noise spectrum estimation and the noise spectrum at each short-time frame.

In particular, the system in [10] improves the probability of correct *noise* frame classification for improved noise spectrum estimation, and smoothes the *a priori* SNR estimation over time, as in the minimum mean-square error short-time spectral magnitude estimator in [1,2]. Time smoothing is effective in reducing musical noise, but it leads to reverberation artifacts.

In this paper we propose a two-stage Wiener filter system for speech noise removal. For simplicity, we use an adaptive energy-based speech/noise classification technique similarto [11]. To reduce the classification error, specifically the error of misclassification of *speech* frames as *noise* frames, we smooth the initial energy-based classification result over time. That is justified by the observation that *speech* frames tend to cluster to each other in time. In other words, both the energy measure and classification results of neighboring frames are used to obtain the final classification result for each current frame, a context-adaptive classification idea that has been successfully used reducing reconstruction noise in picture coding [12].

Driven by the frame classifier, we use a Wiener filter to estimate the speech and noise spectra, or equivalently the *a priori* SNR. Another Wiener filter then generates a minimum-mean square estimate of the speech signal. This two-stage Wiener filtering approach is simple to implement and performs closely to the best systems reported to date, but with a lower level of musical tones.

2 System Outline

A simplified block diagram of our proposed system is shown in Figure 1. The input signal is first transformed on a frame-by-frame basis using a modulated complex lapped transform (MCLT). The MCLT is similar to a windowed Fourier transform frequency analyzer, but with slightly different center frequencies [13]. Frame classification and Wiener filtering, as described in the next sections, are performed in the magnitude MCLT domain. The filtered magnitude information is combined with the original phase information and inverse transformed via the IMCLT.

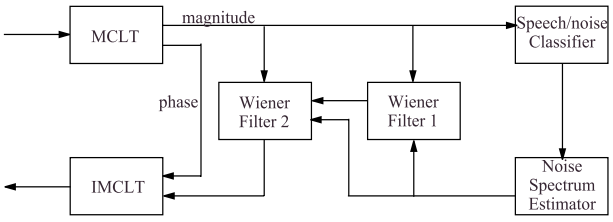

Figure 1: Basic block diagram of the proposed system.

Let x be the input signal, s the original speech signal and n the uncorrelated noise. We assume as usual an additive noise model, that is

$$x = s + n \tag{1}$$

Let $X(i,k)$ be the input spectrum of frame i at frequency bin k, computed via the MCLT:

$$X(i,k) = \sum_{n=0}^{2N-1} x(iN+n)p_a(n,k) \tag{2}$$

where N is the frame length and $p_a(n,k)$ is the MCLT analysis kernel [13].

3 Context-Adaptive Classification

Our classifier is based on an energy metric. The ith frame energy $E^2(i)$ is computed from the input spectrum as follows:

$$E^2(i) = \frac{1}{k_1 - k_0} \sum_{k=k_0}^{k_1} [|X(i,k)| - \bar{X}(i)]^2 \tag{3}$$

where the average frame magnitude $\bar{X}(i)$ is given by

$$\bar{X}(i) = \frac{1}{k_1 - k_0 + 1} \sum_{k=k_0}^{k_1} |X(i,k)| \tag{4}$$

We usually set $k_0 = 300N/f_s$ and $k_1 = 3000N/f_s$ (where f_s is the A/D sampling frequency). That choice is motivated by the fact that for human speech essentially all energy is concentrated in the 300Hz–3000Hz band.

Once the energy $E^2(i)$ is computed, We make an initial decision by hard thresholding: if $E(i) > T$ then frame i is classified as *speech*; otherwise, it is labeled as *noise*. Since speech is nonstationary, we adapt the threshold T from past frames by the simple rule

$$T = E_{min} + \delta(E_{max} - E_{min}) \tag{5}$$

where $E_{min} = min\{E(j)\}$, $E_{max} = max\{E(j)\}$ and $j = i - W_e, i + 1 - W_e, \cdots, i - 1$ with (W_e, δ) respectively the window size (number of past frames) and a relative thresholding constant. We can slow down adaptation of T by increasing the window size W_e, and we can make it more robust to large energy fluctuations in noise frames by increasing δ. Typical values in our experiments are $W_e = 20$ and $\delta = 0.3$.

A problem with this simple hard-thresholding technique is that it often misclassifies low energy *speech* frames (e.g. for unvoiced speech) as *noise* frames. To reduce this error, we propose the following smoothing rule: if the energies of the current frame and the past W_e frames are below the threshold, then the current frame is a *noise* frame; otherwise, the current frame is a *speech* frame. W_s is a smoothing length; in our experiments we set $W_s = 5$. The rule is justified because in practice low-energy unvoiced frames usually happen immediately before or after voiced frames. Figure 2 shows an example where we see that this smoothing process helps to reduce the error of misclassifying *speech* frames into *noise* frames.

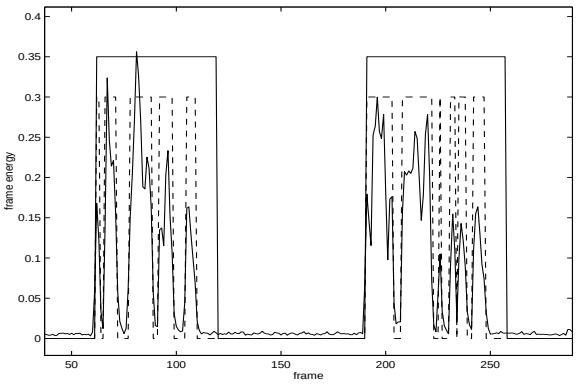

Figure 2: Comparison of energy-based classification results before (hard-decision, dashed lines) and after smoothing (soft-decision, solid lines) ($W_s = 5, \delta = 0.2, W_e = 20$).

4 Two-Stage Wiener Filtering

After classification, we use each *noise* frame to adapt the noise spectrum estimate $|\hat{N}(i,k)|$ by

$$|\hat{N}(i,k)| = \beta|\hat{N}(i-1,k)| + (1-\beta)|X(i,k)| \tag{6}$$

where the parameter β controls the adaptation speed. In our experiments, we use $\beta = 0.9$.

A Wiener filter [14] is the optimal Bayesian linear filter that minimizes the expected mean-squared error $E[|\hat{s} - s|^2]$ for the noise corruption model in Eqn. (1). In the frequency domain, the Wiener filter gain can be written as

$$G(k) = \frac{|S(k)|^2}{|S(k)|^2 + |N(k)|^2} = \frac{P(k)}{1 + P(k)} \tag{7}$$

where $S(k), N(k)$ are respectively the frequency spectrum of the signal and noise. $P(k) \equiv |S(k)|^2/|N(k)|^2$ is the *a priori* SNR. The output spectrum $\hat{S}(k)$ is computed by $\hat{S}(k) = G(k)X(k)$.

The Wiener filter is essentially an adaptive gain that gets smaller as the SNR $P(k)$ gets smaller. Its efficiency is tied to the assumptions that both signal and noise are wide-sense stationary random processes and the a priori SNR is known. In practice, many noise sources such as computers and fans are reasonably stationary, but speech certainly isn't. Therefore, we have to replace the a priori statistics by spectral estimates.

When frame-adaptive spectral estimates are used to compute the Wiener filter gains in Eqn. (7), low-level speech frames can make $G(k)$ fluctuate rapidly, generating annoying *musical noise* in the filtered signal [6].

To improve the spectrum estimation of speech signals, we propose to use a two-step Wiener filtering algorithm. In the first stage, the input signal is Wiener filtered

using an adjusted SNR estimate:

$$P^0(i,k) = \alpha \hat{P}(i-1,k) + (1-\alpha)P(i,k) \tag{8}$$

where

$$P(i,k) = (|X(i,k)|^2 - |\hat{N}(i,k)|^2)/|\hat{N}(i,k)|^2 \tag{9}$$

and $\hat{P}(i-1,k)$ is calculated, using the filtered signal from the previous frame, as

$$\hat{P}(i-1,k) = |\hat{S}(i-1,k)|^2/|\hat{N}(i-1,k)|^2 \tag{10}$$

We see that $P(i,k)$ is equivalent to that resulted from a spectral subtraction system [5, 11]. However, direct spectral subtraction leads to musical noise while oversubtraction increases speech distortion.

With the smoothed estimate $P^0(i,k)$, we reduce variations in the Wiener gain $G(i,k)$ over time. This helps to suppress the residual musical noise. The larger the α, the lower the level of the residual musical noise. In Figure 3 we show different estimations of the SNR. It can be seen that isolated small magnitude pulses (corresponding directly to the *musical noise*) are suppressed after the smoothing operation.

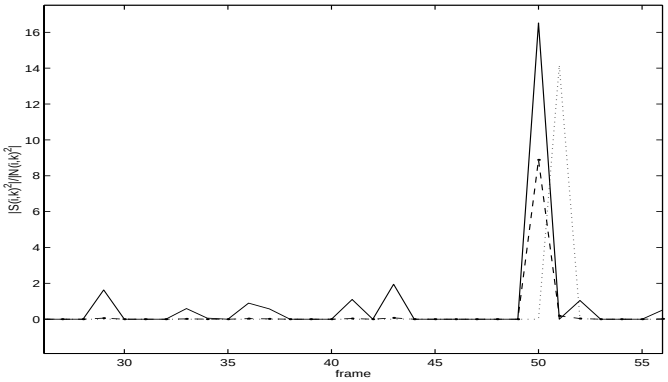

Figure 3: Different SNR estimates. Solid line: $P(i,k)$ before smoothing; dotted line: $P^0(i,k)$ (after smoothing) with $\alpha = 0.97$; dashed line: $P^1(i,k)$ final estimate.

In Figure 3 we note that the smoothed SNR estimate $P^0(i,k)$ is delayed with respect to $P(i,k)$ for large α (e.g.$\alpha = 0.97$). This time delay may lead to reverberation effects at the end of speech utterances. To avoid that kind of distortion, we propose the use of a second Wiener filter, which recomputes the SNR estimation by

$$P^1(i,k) = \alpha \hat{P}(i-1,k) + (1-\alpha)P^u(i,k) \tag{11}$$

where $P^u(i,k) = |\hat{S}(i,k)|^2/|\hat{N}(i,k)|^2$ with $\hat{S}(i,k)$ the filtered signal from the first Wiener filter. A typical plot of $P^1(i,k)$ is also shown in Figure 3. We note that the newly estimated $P^1(i,k)$ is shifted back and synchronized with that of $P_{old}(i,k)$ from spectrum subtraction, while suppressing the small magnitude pulses to avoid musical noise.

5 Experimental Results

To measure the performance of the proposed algorithm, we compute the sample SNR and the noise-to-masking ratio (NMR) for the filtered speech signals. The sample SNR is defined as

$$SNR = 10 \log 10 \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [y(n) - s(n)]^2} \tag{12}$$

where N is the length of the original signal $s(n)$ and $y(n)$ is the signal for which we want to compute the SNR (either the input speech $x(n)$ or the filtered output from our system). The NMR is an objective measure based on the human auditory system and it indicates the ratio of audible noise components to the hearing threshold. Therefore, an NMR of 0 dB indicates a noise at the threshold of audibility, whereas higher NMRs mean more noticeable noise. The NMR has been found to have a high degree of correlation with subjective tests. The NMR is defined as [5]

$$NMR = \frac{10}{M} \sum_{i=0}^{M-1} \log 10 \frac{1}{B} \sum_{b=0}^{B-1} \frac{1}{C_b} \frac{\sum_{k=k_l}^{k=k_h} |D(i,k)|^2}{T_b^2(i)} \tag{13}$$

where M is the total number of frames, B is the number of Critical Bands (CB) , C_b is the number of frequency components for the bth CB, and $|D(i,k)|^2$ is the power spectrum of the noise at frequency bin k and frame i. The k_l, k_h are respectively the low and high frequency bin indices corresponding to bth CB, and T_b is its masking threshold, which depends on the signal spectral magnitudes around the bth band [5].

To generate noisy speech signals, we used Eqn. (1) with six noise patterns. Besides white and pink noise, for more realistic results we also used four noise patterns recorded from office and conferencing rooms, with a mixture of air conditioning and computer noises. The speech material consisted of short sentences recorded by a male and a female speaker. All signals were sampled at 16 kHz (which is characteristic of "wideband" teleconferencing systems). We adjusted the noise level to an equivalent *a priori* SNR of 10 dB.

The results are given in Table 1. The rows indicate the SNR and NMR results before (suffix "in") and after (suffix "out") noise reduction, for male and female speech ("M:" and "F:" prefixes), and the columns indicate the noise patterns; the four recorded room noises (a)–(d) and pink and white noises ("PN" and "WN"). We see that the proposed algorithm significantly improves the SNR or equivalently reduces the NMR. The average SNR improvement is 5.8 dB or equivalently 12.9 dB NMR reduction. That level of SNR improvement is roughly the same as what is obtained with the best spectral subtraction systems [3], but our proposed algorithm leads to a significant reduction of the *musical noise* artifact, with low algorithmic complexity and low processing delay.

Table 1: SNR and NMR (in dB) before and after noise reduction.

	(a)	(b)	(c)	(d)	PN	WN
M: SNR_{in}	9.9	9.8	10.0	10.0	10.1	10.0
M: SNR_{out}	13.1	12.9	12.6	19.2	14.3	15.6
F: SNR_{in}	9.9	9.9	9.9	10.0	10.2	10.0
F: SNR_{out}	17.7	17.6	16.0	20.7	16.2	15.9
SNR Gain	5.5	5.4	4.4	9.9	4.1	5.7
M: NMR_{in}	11.7	15.0	16.3	11.9	21.7	28.5
M: NMR_{out}	2.7	4.0	4.9	-0.1	6.6	11.1
F: NMR_{in}	15.9	19.0	17.4	12.0	19.5	25.2
F: NMR_{out}	3.9	3.7	5.0	1.9	5.3	8.6
NMR Gain	10.5	12.2	11.9	11.1	14.7	17

6 Conclusion

We proposed an adaptive noise reduction algorithm based on Wiener filtering. It includes two main modifications compared to conventional approaches:(i) a smoothing rule for the energy-based speech/noise classification and (ii) a recursive two-stage Wiener filtering structure, to reduce the signal distortion from "musical noise." Preliminary experimental results have shown an average SNR improvement of about 6 dB and an NMR reduction of about 13 dB, for noisy speech at 10 dB input SNR.

With speech input, the performance of our system could be enhanced by adding speech production models (e.g. linear prediction – LP) as part of the *a priori* spectral information. However, such modification could hinder performance on handset-free telephony and similar applications, due to the mismatch of the LPC model to reverberant speech.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, pp. 1109–1121, 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on ASSP*, pp. 443–445, 1985.

[3] Y. Ephraim, "A signal subspace approach for speech enhancement," *IEEE Trans. on speech and audio processing*, pp. 251–266, 1995.

[4] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on speech and audio processing*, pp. 126–137, 1999.

[5] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. on speech and audio processing*, pp. 497–514, 1997.

[6] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on speech and audio processing*, pp. 345–349, 1994.

[7] P. Vary, "Noise suppression by spectral magnitude estimation: mechamism and theorectical limits," *Signal Processing*, pp. 387–400, 1985.

[8] P. Sovka, V. Davidek, P. Pollak, and J. Uhlir, "Speech/ pause detection for real-time implementation of spectral subtraction algorithm," in *The 6th Intl. Conf. on Signal Proc. Applications and Technology*, 1995, pp. 1955–1958.

[9] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings-I*, pp. 377–380, 1992.

[10] A. Benyassine, E. Shlomot, and H. Y. Su, "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simulations voice and data applications," *IEEE Communications Magazine*, pp. 64–73, 1997.

[11] G. S. Kang and L. J. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Trans. on ASSP*, pp. 939–942, 1989.

[12] C. Chrysafis and A. Ortega, "Efficient context-based entropy coding for lossy wavelet image compression," in *Proc. of DCC'97*, Snowbird, UT, Mar. 1997.

[13] H. Malvar, "A modulated complex lapped transform and its application to audio processing," in *Proc. ICASSP*, 1999, pp. 1421–1424.

[14] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, New York: Wiley, 1968.