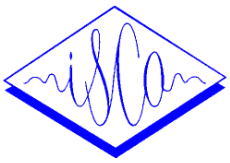

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Computation-Efficient Parameter Adaptation Algorithm for the Generalized Spectral Subtraction Method

Jin-Jie Zhang, Zhi-Gang Cao, Zheng-Xin Ma

State Key Lab on Microwave & Digital Communications, Electronic Engineering Department,
Tsinghua University, Beijing, 100084, P.R.China

ABSTRACT

The background noise in speech is not only objectionable to listeners with proper hearing but also very harmful to those hearing impaired. Hence, there is a strong need to employ speech enhancement to suppress the background noise in speech. In this paper, a computation-efficient parameter-adaptation algorithm is proposed for the generalized spectral subtraction method, which can reduce noise level more significantly.

1. INTRODUCTION

The presence of background noise in speech is not only objectionable to listeners but also causes the degradation of speech quality and/or intelligibility, which is especially harmful to those hearing impaired. In addition, noise degrades many automatic speech processing systems, such as speech coding, speech recognition and so on. Hence, there is a strong need to employ speech enhancement to suppress the background noise in speech.

This paper focuses on the single channel speech enhancement based on generalized spectral subtraction (GSS) method [1-2]. The GSS method enhances the short-time-spectral-amplitude (STSA) of speech by subtracting the averaged noise spectra from the noisy speech. Due to the fact that the speech phase is not very important in speech enhancement [3], the GSS method makes no efforts to estimate the speech phase and simply reuses that of the noisy speech. The original GSS method is easy to be implemented, but its enhancement performance is poor since it does not give a way to choose parameter values properly. In most cases its parameter values are chosen empirically beforehand and then left unchanged in the following speech enhancement. As a result, its residual noise level is often kept high.

In this paper, a computation-efficient algorithm is proposed to choose parameter values adaptively. Objective measures and informal listening test show that the proposed algorithm can suppress noise more effectively.

2. ALGORITHM DESCRIPTION

The noisy speech is modeled in time domain as

$$y(n) = x(n) + d(n),$$

where $x(n)$ is the noise-free speech and $d(n)$ is the background noise assumed to be stationary and uncorrelated with $x(n)$. Since the enhancement is done frame by frame in frequency domain, the above model can be rewritten as

$$Y(k,i) = X(k,i) + D(k,i),\ k = 0,1,\cdots,N-1,\ i = 1,2,\cdots,$$

where N is the size of fast Fourier transform (FFT), i is frame index, $Y(k,i)$, $X(k,i)$ and $D(k,i)$ are the FFT of windowed $y(n,i)$, $x(n,i)$ and $d(n,i)$. Let $\hat{x}(n,i)$ and $\hat{X}(k,i)$ denote the enhanced speech and its spectrum.

The GSS is originally formulated in [1-2] as

$$\left\|\hat{X}(k,i)\right\|^{\alpha} = \left\|Y(k,i)\right\|^{\alpha} - \beta E\left[\ \left\|D(k,i)\right\|^{\alpha}\right], \qquad (1)$$

where α and β are parameters introduced to optimize the enhancement performance. As experiments show the value choice of α is not so critical as that of β [4], we fix it to 1 here for simplicity and only make the value of β be updated in time-frequency domain. Then, GSS can be re-formulated as

$$\left\|\hat{X}(k,i)\right\| = \left\|Y(k,i)\right\| - \beta(k,i)\left\|D(k,i)\right\|. \qquad (2)$$

Here $\left\|D(k,i)\right\|$ is the noise spectral amplitude estimated by moving average in speech pauses as,

$$\left\|D(k,i)\right\|^2 = \varepsilon\left\|D(k,i-1)\right\|^2 + (1-\varepsilon)\left\|Y(k,i)\right\|^2, \qquad (3)$$

where $0 < \varepsilon < 1$, is a forgetting factor. An algorithm based on [7] is employed to detect speech pause.

Ideally, the enhanced speech should have the same STSA as the noise-free speech, that is,

$$\left\|X(k,i)\right\| - \left\|\hat{X}(k,i)\right\| = 0. \qquad (4)$$

Substituting (2) into (4) gets the solution as

$$\beta(k,i) = \frac{\left\|Y(k,i)\right\| - \left\|X(k,i)\right\|}{\left\|D(k,i)\right\|}. \qquad (5)$$

Let

$$\gamma(k,i) = \frac{\left\|Y(k,i)\right\|^2}{\left\|D(k,i)\right\|^2},\ \xi(k,i) = \frac{\left\|X(k,i)\right\|^2}{\left\|D(k,i)\right\|^2},$$

then (5) can be rewritten as

$$\beta(k,i) = \gamma(k,i) - \xi(k,i). \qquad (6)$$

Here $\gamma(k,i)$ is *a posteriori* SNR and $\xi(k,i)$ is *a priori* SNR [5]. Obviously the value of $\beta(k,i)$ is proportional to the difference between the *a posteriori* SNR and the *a priori* SNR, i.e., the larger the difference is, the greater the value of $\beta(k,i)$ is taken. This is reasonable since larger difference means higher noise level in noisy speech and thus needs a greater $\beta(k,i)$ to compress noise according to (2). However, such a $\beta(k,i)$ is not applicable in practical situation since the exact value of $\xi(k,i)$ is unavailable then. Thus, approximation must be performed.

First, it is supposed that $\|Y(k)\|^2 \approx \|X(k)\|^2 + \|D(k,i)\|^2$. Replace the $\|Y(k,i)\|$ term in (5) with this approximation and let $\beta^*(k,i)$ denote $\beta(k,i)$, we get,

$$\beta^*(k,i) = \left(\sqrt{1+\xi(k,i)} + \sqrt{\xi(k,i)}\right)^{-1}. \quad (7)$$

Then $\beta(k,i)$ is re-estimated by averaging with its value in previous frame as

$$\beta(k,i) = \mu\beta(k,i-1) + (1-\mu)\lambda\beta^*(k,i), \quad (8)$$

where $0 \le \mu \le 1$ and $\lambda > 0$ is a correctional factor introduced. In the proposed algorithm $\xi(k,i)$ is estimated as,

$$\xi(k,i) = \eta \frac{\left\|\hat{X}(k,i-1)\right\|^2}{\|D(k,i)\|^2} + (1-\eta)\times \max\{\gamma(k,i) - \beta_0,\ 0\}, \quad (9)$$

where $0 < \eta < 1$, $\beta_0 \ge 1$ and $\left\|\hat{X}(k,i-1)\right\|$ is the STSA of the enhanced speech in previous frame. Here (9) uses a β-fixed GSS to perform a coarse estimation of $\left\|\hat{X}(k,i)\right\|$. It is obvious that in (7)-(9) the components used to calculate $\beta(k,i)$ are same as that for GSS and the operations involved are very simple, so the computation load is very low.

With this adaptive $\beta(k,i)$, the spectrum of the enhanced speech can then be estimated as

$$\hat{X}(k,i) = \max\left\{\|Y(k,i)\| - \beta(k,i)\|D(k,i)\|, 0\right\} . \exp\left(j \arg\left(Y(k,i)\right)\right), \quad (10)$$

where half-wave rectification is performed to avoid producing negative spectral components. Finally, with inverse FFT and overlapped addition, the enhanced speech can be obtained.

3. SIMULATION AND PERFORMANCE EVALUATION

In the simulation and performance evaluation, 6 speech sentences of 6 native Chinese speakers (3 male and 3 female) selected from Chinese mandarin MOS test database are used. The noise data, which includes white noise, speech-like noise and aircraft cockpit noise, are selected from the NOISEX-92 database. They are both digitized in 8kHz/16bits and added digitally together to form the noisy speech with SNRs from 10dB to -5dB.

Segmental SNR (SEGSNR) [8] improvement and Noise-to-Masking Ratio (NMR) [6] decrement are used as objective measures for performance evaluation. The SEGSNR of enhanced speech is defined as

$$\text{SEGSNR(dB)} = \frac{10}{M}\sum_{i=0}^{M-1} \log_{10} \frac{\frac{1}{L}\sum_{n=0}^{L-1} x^2(n,i)}{\frac{1}{L}\sum_{n=0}^{L-1}\left[x(n,i) - \hat{x}(n,i)\right]^2},$$

where L is the frame length and M is the total number of frames. The SEGSNR measures the waveform distortion of speech. The NMR is defined as

$$\text{NMR(dB)} = \frac{10}{M}\sum_{i=1}^{M} \log_{10} \frac{1}{B}\sum_{b=0}^{B-1} \frac{\frac{1}{C_b}\sum_{k=k_{bl}}^{k_{bh}} D_p(k,i)}{T_{p,b}(i)},$$

where $T_{p,b}(i)$ is the masking threshold of pure speech, $D_p(k,i)$ is the power spectra of noise in speech, k_{bl} and k_{bh} are the lower and upper limits of critical band b, C_b is the number of frequency components included in critical band b and B is the number of critical bands used. The NMR can measure the audible noise level of speech and was found to have high correlation with subjective test [6]. For the sake of convenience, the SEGSNR and NMR difference between the noisy speech and the enhanced speech are given in the following performance evaluation.

The value choices of ε, η, β_0, μ and λ have been investigated by a number of experiments with different noise types and different noise levels. By making tradeoff between residual noise level and speech distortion, the parameter values are chosen as $\varepsilon = 0.995$, $\eta = 0.98$, $\beta_0 = 5.0$, $\mu = 0.9$ and $\lambda = 1.9$.

The objective measures of the proposed algorithm have been compared with those of power spectral subtraction (PSS) and optimized generalized spectral subtraction (OGSS). The PSS is the special case of the GSS ($\alpha = 2, \beta = 1$) and was proposed first among the subtractive-type speech enhancement methods. The OGSS is another special case of GSS when $\alpha = 2$ and β is optimized to 5 under white noise with SEGSNR measure. Figure1-3 present the SEGSNR measures of 3 methods for the 3 different noise types respectively and show that the proposed algorithm gains more SEGSNR improvements, especially in lower input SNRs. The NMR measures are given in Figure4-6. It is obvious that the proposed algorithm outperforms the other two very significantly for all the noise types and all the noise levels, which means that it can suppress audible noise more effectively than the other two.

Informal subjective listening test shows that the speech enhanced by the proposed algorithm has less audible residual noise than that enhanced by the β-fixed ones and as a result, the enhanced speech sounds more clearly.

4. CONCLUSION

In this paper, a computation-efficient algorithm is proposed to select parameter value for the generalized spectral subtraction method adaptively. Objective measures and informal listening test show that the proposed algorithm can suppress noise more effectively than the parameter-fixed ones.

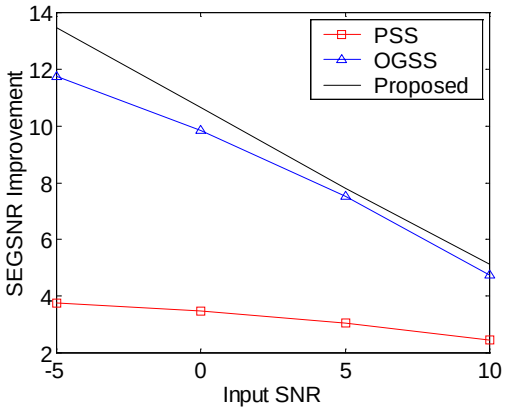

Figure 1. SEGSNR measures for white noise

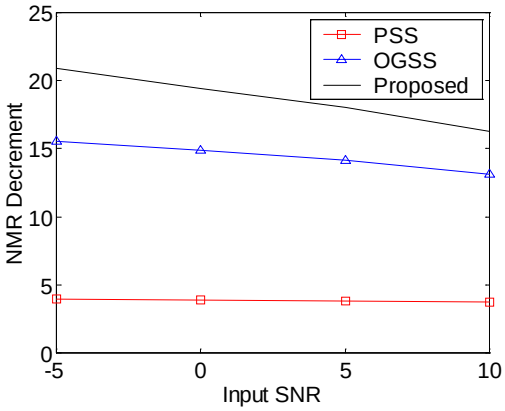

Figure 4. NMR measures for white noise

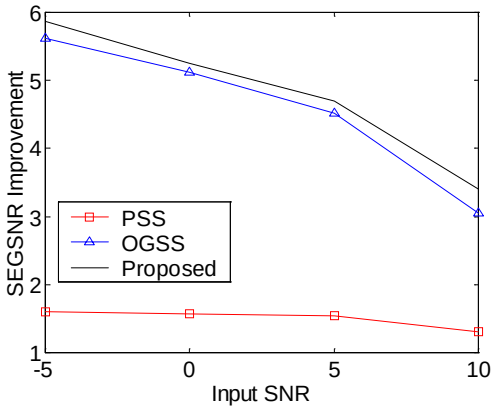

Figure 2. SEGSNR measures for speech-like noise

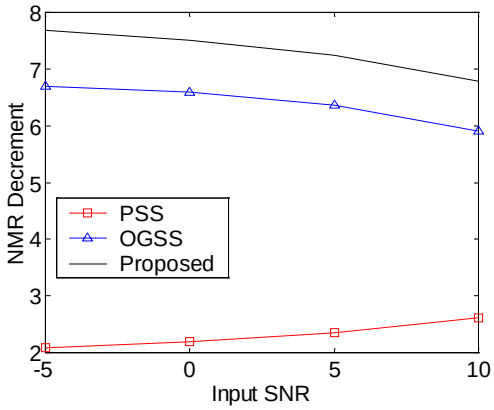

Figure 5. NMR measures for speech-like noise

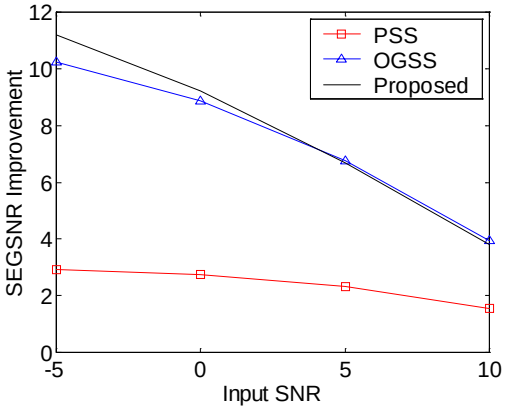

Figure 3. SEGSNR measures for aircraft cockpit noise

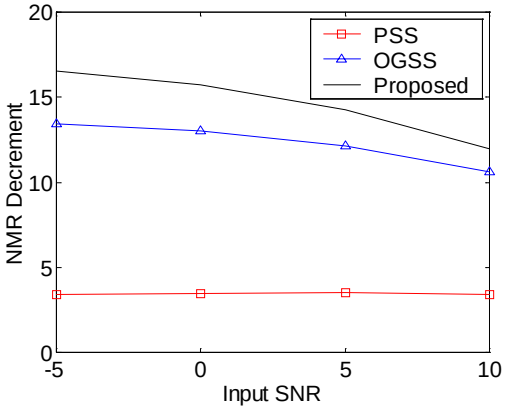

Figure 6. NMR measures for aircraft cockpit noise

5. RFERENCES

1. Berouti, M., Schwartz, R. and Makhoul, J. “Enhancement of speech corrupted by acoustic noise,” *ICASSP’79*, pp. 208-211, Apr. 1979.

2. Lim, J. S. and Oppenheim, A. V. “Enhancement and bandwidth compression of noisy speech, ” *Proc. IEEE ICASSP*, Washington, DC, Apr., 1979, pp. 208-211.

3. Wang , D.L. and Lim, J.S., “The unimportance of phase in speech enhancement,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-30, pp.679-681, Aug. 1982.

4. Virag, N., “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech and Audio Processing*, vol.7, pp.126-137, Mar. 1999.

5. Ephraim, Y. and Malah, D. “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-32, pp.1109-1121, Dec. 1984.

6. Tsoukalas, D.E., Mourjopoulos, J.N. and Kokkinakis, G. “Speech enhancement based on audible noise suppression,” *IEEE Trans. Speech Audio Processing*, vol. 5, pp.497-514, Nov. 1997.

7. Lynch J.F., Josenhans, J.G. and Crochiere R.E. “Speech/silence segmentation for real-time coding via rule based adaptive endpoint detection,” *ICASSP’87*, pp. 1348-1351, Apr. 1987.

8. Quakenbush, S., Barnwell, T. and Clements, M. *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.