PSYCHOACOUSTICALLY CONSTRAINED AND DISTORTION MINIMIZED SPEECH ENHANCEMENT ALGORITHM

Seokhwan Jo and Chang D. Yoo

Div. of EE, School of EECS, KAIST,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
antiland00@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

A psychoacoustically constrained and distortion minimized speech enhancement algorithm is considered. In general, noise reduction leads to speech distortion, and thus, the goal of an enhancement algorithm should reduce noise and speech distortion so that both are inaudible. In this paper, a constrained optimization problem is formulated so that speech distortion is minimized while distortion that includes residual noise and speech distortion is kept below the masking threshold of the clean speech. Experimental results show that the algorithm considered in this paper outperforms some of the more popular algorithms in terms of improvement in segmental signal-to-noise ratio (SegSNR) and spectral distance (SD).

Index Terms— speech enhancement, speech distortion, residual noise, constrained optimization, masking threshold

1. INTRODUCTION

Noisy speech degrades the performance of speech communication and recognition systems. Most important of all, noise induces listener fatigue. Up till now, many speech enhancement algorithms have removed noise based on the minimum mean square criterion [1, 2]. This may be an over simplification of what needs to be considered. Noise reduction inevitably leads to distortion in speech and should be kept to a minimum. There should be a delicate balance between speech distortion and noise reduction. In this paper, a psychoacoustically constrained and distortion minimized speech enhancement algorithm is considered. The estimator aims to minimize speech distortion with the distortion and residual noise below the masking threshold.

In the past, numerous speech enhancement algorithms using psychoacoustic masking threshold have been proposed [3]-[8]. In [3] and [4], the masking threshold was used to guide the derivation of oversubtraction factor of the clean speech estimator. In [5], the masking threshold was used as a constraint to keep the residual noise inaudible. In [6], clean speech was estimated iteratively so that the residual noise was kept below the masking threshold. Some algorithms used the masking threshold not in the frequency domain but in the signal subspace domain [7, 8] to constrain the residual noise to be below the masking threshold. Most of the methods mentioned had not considered constraining the speech distortion.

The algorithm considered in this paper derives the clean speech estimator by psychoacoustically constraining the distortion to be below the masking threshold. The estimator minimizes speech distortion while keeping the distortion and residual noise below the masking threshold. This paper provides a solution for the constrained optimization problem. A modification of the estimator to improve speech enhancement measurements is also provided.

This paper is organized as follows. The considered algorithm is presented in Section 2. Section 2.1 presents psychoacoustically constrained and distortion minimized speech enhancement algorithm. Section 2.2 presents the parameters for speech enhancement and the calculation of masking threshold. Section 3 provides experimental results, and finally Section 4 concludes the paper.

2. PSYCHOACOUSTICALLY CONSTRAINED AND DISTORTION MINIMIZED SPEECH ENHANCEMENT ALGORITHM

2.1. The constrained optimization

When a clean speech $s[n]$ is corrupted by uncorrelated additive noise $z[n]$, noisy speech $y[n]$ is given by

$$y[n] = s[n] + z[n]. \tag{1}$$

In the frequency domain, the following relationship holds:

$$Y(\omega, l) = S(\omega, l) + Z(\omega, l), \tag{2}$$

where $Y(\omega, l)$, $S(\omega, l)$ and $Z(\omega, l)$ are the short-time Fourier transforms (STFT) of $y[n]$, $s[n]$ and $z[n]$ at time frame l, respectively. Here, ω and l denote frequency and time frame index, respectively.

Let $\hat{S}(\omega, l) = G(\omega, l)Y(\omega, l)$ be an estimate of $S(\omega, l)$ where $G(\omega, l)$ is the gain function for enhancement. The estimation error can be expressed as

$$\begin{aligned} \varepsilon(\omega, l) &= \hat{S}(\omega, l) - S(\omega, l) \\ &= (G(\omega, l) - 1)S(\omega, l) + G(\omega, l)Z(\omega, l) \\ &= \varepsilon_s(\omega, l) + \varepsilon_z(\omega, l), \end{aligned} \tag{3}$$

where $\varepsilon_s(\omega, l)$ and $\varepsilon_z(\omega, l)$ denote lth frame STFT of speech distortion and residual noise.

Let $E_s(\omega, l) = E\{\varepsilon_s^H(\omega, l)\varepsilon_s(\omega, l)\}$ and $E_z(\omega, l) = E\{\varepsilon_z^H(\omega, l)\varepsilon_z(\omega, l)\}$. Finding the optimal gain function $G(\omega, l)$ can be formulated as solving the following constrained optimization problem:

$$\min_G E_s(\omega, l)$$

$$\text{subject to } E_s(\omega, l) + E_z(\omega, l) \leq T(\omega, l), \tag{4}$$

where $T(\omega, l)$ is the masking threshold of the lth frame. The con-

This work was supported by grant No. R01-2007-000-20949-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

strained optimization problem will lead to a solution that minimizing speech distortion while constraining the total distortion below the masking thresholds so that distortion and residual noise are inaudible.

The problem can be solved by using the method of Lagrangian multipliers as shown below:

$$\min_{G} \max_{\alpha \geq 0} L(G, \alpha), \tag{5}$$

where $L(G,\alpha) = E_s(\omega,l) - \alpha(\omega,l)(E_s(\omega,l) + E_z(\omega,l) - T(\omega,l))$ and $\alpha(\omega,l)$ is the Lagrangian multiplier. Substituting (3) into (5),

$$\begin{aligned} L(G,\alpha) &= (G(\omega,l)-1)^2 P_s(\omega,l) \\ &+ \alpha(\omega,l)((G(\omega,l)-1)^2 P_s(\omega,l) \\ &+ G^2(\omega,l) P_z(\omega,l) - T(\omega,l)), \end{aligned} \tag{6}$$

where $P_s(\omega,l) = E\{S^H(\omega,l)S(\omega,l)\}$ and $P_z(\omega,l) = E\{Z^H(\omega,l)Z(\omega,l)\}$ are the power spectrum of clean speech and noise, respectively. Then, by partially differentiating L with respect to the gain function G, an optimal gain function can be derived as

$$G(\omega,l) = \frac{\xi(\omega,l)}{\xi(\omega,l) + \frac{\alpha(\omega,l)}{1+\alpha(\omega,l)}}, \tag{7}$$

where $\xi(\omega,l) = \frac{P_s(\omega,l)}{P_z(\omega,l)}$ which is defined as the *a priori* signal-to-noise ratio (SNR). Substituting (7) into (5) leads to L as follows:

$$\begin{aligned} L &= \frac{\alpha^2\xi}{((1+\alpha)\xi+\alpha)^2} + \frac{\alpha^3\xi}{((1+\alpha)\xi+\alpha)^2} \\ &+ \frac{\alpha(1+\alpha)^2\xi^2}{((1+\alpha)\xi+\alpha)^2} - C\alpha \\ &= \frac{-(\xi+1)(\xi(C-1)+C)\alpha^3}{((1+\alpha)\xi+\alpha)^2} \\ &+ \frac{(\xi+2\xi^2-2C\xi^2-2C\xi)\alpha^2 + (\xi^2-C\xi^2)\alpha}{((1+\alpha)\xi+\alpha)^2}, \end{aligned} \tag{8}$$

where $C = \frac{T(\omega,l)}{P_z(\omega,l)}$. If $-(\xi+1)(\xi(C-1)+C) > 0 \rightarrow \frac{\xi}{\xi+1} > C$, then α does not exist, because the value of α that maximize L is ∞. Therefore, $\frac{\xi}{\xi+1} \leq C$ must be satisfied for a value of α to be found such that

$$\frac{\partial L}{\partial \alpha} = 0 \text{ if } \frac{\xi}{\xi+1} \leq C. \tag{9}$$

By solving the equation (9), two possible values of α are obtained. By imposing the condition $\alpha \geq 0$, α is given by

$$\alpha = \frac{\xi(\xi(C-1)+C) - \xi\sqrt{(\xi(C-1)+C)}}{-(\xi+1)(\xi(C-1)+C)}. \tag{10}$$

But, if $\xi(C-1)+C > 1$ which means $C > 1$ is satisfied, then both values of α are negative. Thus, if $C > 1$, then α is set to 0. The optimal gain function is derived as follows:

$$G = \begin{cases} 1 & C > 1 \\ \frac{\xi}{\xi+\beta} \quad \left(\beta = \frac{\xi - \xi\sqrt{(\xi(C-1)+C)}}{\xi+\sqrt{(\xi(C-1)+C)}}\right) & C \leq 1, \frac{\xi}{\xi+1} \leq C \\ \text{not exist} & C \leq 1, \frac{\xi}{\xi+1} > C \end{cases}. \tag{11}$$

However, the gain function G in (11) needs to be modified for the following two reasons. First, when $\frac{\xi}{\xi+1} > C$ (constraint can not be satisfied), and therefore solution for G does not exist. We need to set G to a reasonable value. Second, when $C > 1$ ($T(\omega,l) > P_z(\omega,l)$ and noise is inaudible), there is no improvement in SNR since $G = 1$. We need to set G to some value such that SNR is improved. Thus, the proposed gain function is as follows:

$$G(\omega,l) = \begin{cases} \frac{\xi(\omega,l)}{\xi(\omega,l)+\beta(\omega,l)} & C \leq 1, \frac{\xi(\omega,l)}{\xi(\omega,l)+1} \leq C \\ \frac{\xi(\omega,l)}{\xi(\omega,l)+\gamma(\omega,l)} & \text{otherwise} \end{cases}, \tag{12}$$

where $\beta(\omega,l) = \frac{\xi(\omega,l) - \xi(\omega,l)\sqrt{(\xi(\omega,l)(C-1)+C)}}{\xi(\omega,l)+\sqrt{(\xi(\omega,l)(C-1)+C)}}$ and $\gamma(\omega,l)$ is an over-subtraction factor controlled by a local averaged value of *a priori* SNR. The local averaged value of *a priori* SNR, $\bar{\xi}(\omega,l)$, is defined as follows:

$$\bar{\xi}(\omega,l) = \frac{1}{0.02\pi}\int_{\omega-0.01\pi}^{\omega+0.01\pi} \xi(\Omega,l)d\Omega. \tag{13}$$

When $\bar{\xi}(\omega,l)$ is low, noise reduction needs to be increased, thus, $\gamma(\omega,l)$ is a large value and the noise reduction is strengthened. The converse happens when $\bar{\xi}(\omega,l)$ is high.

2.2. Parameters for enhancement and calculation of masking threshold

2.2.1. Power spectrum of clean speech

In this paper, speech is modelled as an output of an autoregressive (AR) process of order p and is mathematically expressed as

$$\sum_{k=0}^{p} b_k s_l[n-k] + g \cdot d[n] = 0, \quad b_0 = -1, \tag{14}$$

where g and $d[n]$ represent the gain and white Gaussian noise with zero mean and unit variance. The subscript l signifies that $s_l[n]$ is a short-time segment which is obtained by applying a window at the region of interest. The power spectrum of $s_l[n]$ is expressed as

$$P_s(\omega,l) = \frac{g^2}{\left|1 - \sum_{i=1}^{p} b_i e^{-ji\omega}\right|^2}. \tag{15}$$

2.2.2. Masking threshold calculation

To constraint the residual noise and the speech distortion to be below the masking threshold, we need to calculate the masking threshold. The calculation of the masking threshold is summarized in a number of literatures [3, 9]. The steps involved in determining the masking threshold are as follows:

1. *Critical band analysis* : sum up the power spectrum in each critical band (Bark), where the power spectrum is obtained by magnitude squaring the Fourier coefficient.
2. *Spreading* : convolve with a spreading function to take into account the effect of adjacent critical bands.
3. *Offset* : subtract the offset by considering the tone-like or noise-like nature of the speech.
4. *Re-normalization* : convert the spread spectrum back to Bark domain.
5. *Absolute threshold* : compare with the absolute threshold and choose the maximum between them.

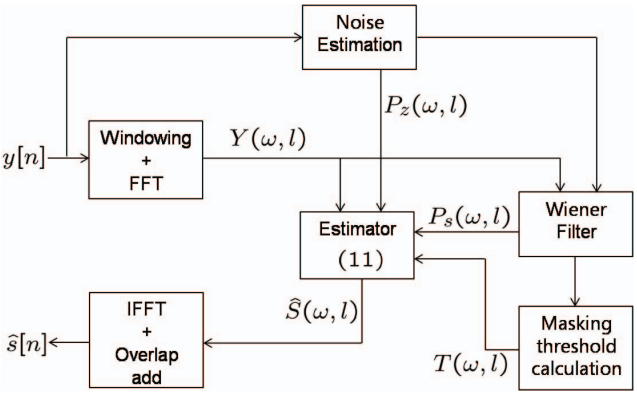


Fig. 1. Block diagram.

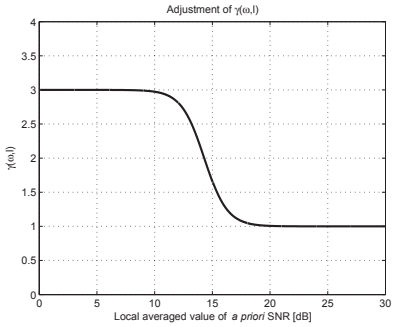


Fig. 2. Adjustment of $\gamma(\omega,l)$. The equation is $\gamma(\omega,l) = \frac{1}{0.5+e^{10\log_{10}\bar{\xi}(\omega,l)-15}} + 1$ which is decided based on the value of SegSNR.

The scheme of the algorithm is shown in Fig. 1. The estimator in (12) is derived by constraining both the speech distortion and the residual noise to be inaudible. Because the masking threshold is difficult to calculate from the noisy speech, speech is roughly estimated, and this value is used to compute the masking threshold. And the power spectrum of the speech is also computed from this roughly estimated speech.

3. EXPERIMENTAL RESULTS

The algorithm considered was evaluated and compared to other speech enhancement algorithms. The test sentences were selected from the TIMIT database. Three kinds of background noises were used in the experiments: white Gaussian, car, and babble noises. Both speech and noise were sampled at 16 kHz. Noise was added to clean speech with various noise levels. To determine the variance of noise, a fast noise tracking estimator was employed [10]. The performance of the considered algorithm was evaluated in terms of segmental signal-to-noise ratio (SegSNR) and spectral distance (SD).

The amount of noise reduction is generally measured by the improvement of SegSNR, which is defined as

$$\text{SegSNR} = \frac{1}{T}\sum_{m=0}^{T-1} 10\log\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1} s^2[n+Nm]}{\frac{1}{N}\sum_{n=0}^{N-1}(s[n+Nm]-\hat{s}[n+Nm])^2}\right)$$

where $s[n]$ and $\hat{s}[n]$ are the original clean and the estimated speech samples, respectively. The upper and lower bound of the frame SNR were set to 35 dB and -5 dB, respectively. All the SegSNR results were averaged over 20 different speech signals.

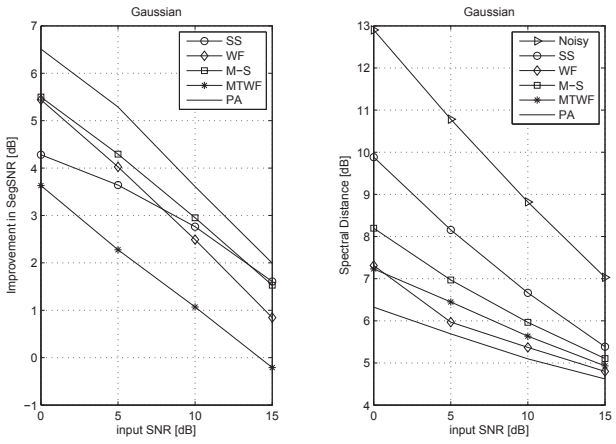
The SD measures the dissimilarity between the spectrum of frames of clean speech and enhanced speech. It is given by

$$\text{SD} = \frac{1}{T}\sum_{m=0}^{T-1}\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[20\log|S_m(\omega)| - 20\log|\hat{S}_m(\omega)|\right]^2 d\omega}$$

where $|S_m(\omega)|$ and $|\hat{S}_m(\omega)|$ are the magnitude spectra of the clean and the enhanced speech signals of the m^{th} signal segment, respectively. All SD results were averaged over 20 different speech signals. Large value of SD implies bad performance.

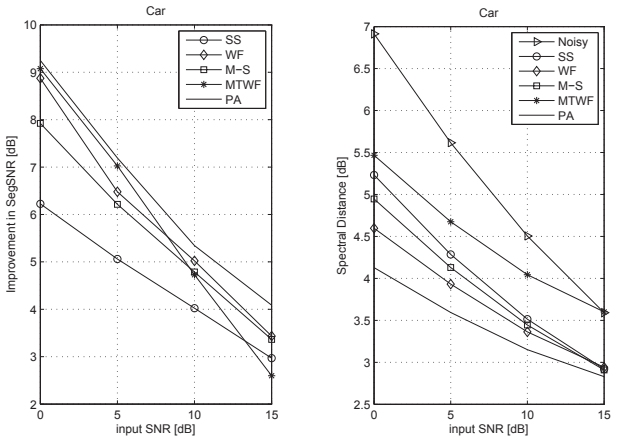
The AR order for the speech was varied depending on mean value of *a priori* SNR in frame: When the mean values of *a priori* SNR in frame is high, the order is high, and when it is low, the order is low. It varied from 0 to 30. $\gamma(\omega,l)$ in (12) was set to $\gamma(\omega,l) = \frac{1}{0.5+e^{10\log_{10}\bar{\xi}(\omega,l)-15}} + 1$ (shown on Fig. 2). A rough estimation for the parameters and the masking threshold was obtained using the Wiener filter (WF), and *a priori* SNR was estimated with the decision directed method [2] with weighing factor set to 0.97. The considered algorithm (PA) was compared to spectral subtraction (SS) [1], WF, MMSE-STSA estimator (M-S) [2], and WF with constrained optimization using the masking threshold (MTWF) proposed in [5]. For WF, M-S and MTWF, *a priori* SNR was estimated with the decision directed method with weighing factor equal to 0.98 as proposed in [2].

Fig. 3 (a) illustrates the average SegSNR improvement using the considered algorithms and the other algorithms for white Gaussian noise at various noise levels. In Fig. 3 (b), the average SD is shown in various white Gaussian noise level. Fig. 4 (a) illustrates the average SegSNR improvement using the considered algorithms and the other algorithms for car noise at various noise levels. In Fig. 4 (b), the average SD is shown in various car noise level. Fig. 5 (a) illustrates the average SegSNR improvement using the considered algorithms and the other algorithms for babble noise at various noise levels. In Fig. 5 (b), the average SD is shown in various babble noise level. In these results, PA outperformed the other algorithms in terms of SegSNR and SD.



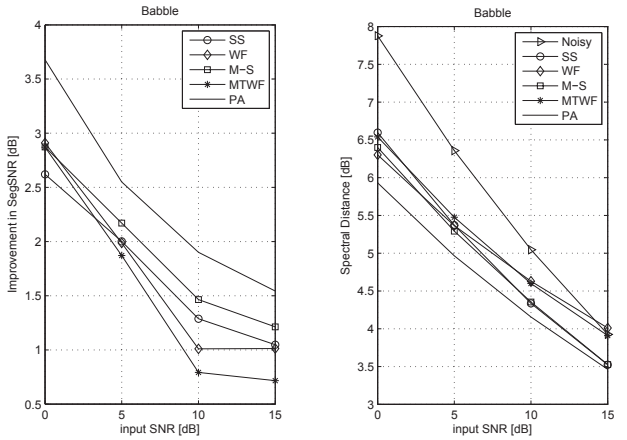
(a) (b)

Fig. 3. (a) SegSNR improvement of the considered algorithm and other algorithms in white Gaussian noise. (b) SD of the considered algorithm and other algorithms in white Gaussian noise.



(a) (b)

Fig. 4. (a) SegSNR improvement of the considered algorithm and other algorithms in car noise. (b) SD of the considered algorithm and other algorithms in car noise.



(a) (b)

Fig. 5. (a) SegSNR improvement of the considered algorithm and other algorithms in babble noise. (b) SD of the considered algorithm and other algorithms in babble noise.

4. CONCLUSION

This paper considers a psychoacoustically constrained and distortion minimized speech enhancement algorithm. Speech enhancement inevitably leads to distortion in speech. Thus, the goal of an enhancement algorithm should reduce noise and speech distortion. This paper provides the clean speech estimator for any distortion that includes residual noise and the speech distortion to be kept below the masking threshold of the speech. A modification of the estimator is also provided to improve SNR. Experimental results show that the algorithm considered in this paper outperforms some of the more popular algorithms.

5. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, April 1979.

[2] Y. Ephraim, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, September December 1984.

[3] N. Virag, "Single channel speech enhancement based on masking properties of human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 126–137, 1999.

[4] S.N. Koh C.H. You and S. Rahardja, "Masking-based beta-order mmse speech enhancement," *Speech Communication*, vol. 48, pp. 57–70, 2006.

[5] Y. Hu and P.C. Loizou, "Incorporation a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol. 11, February 2004.

[6] V. Radhakrishnan J.H.L. Hansen and K.H. Arehart, "Speech enhancement based on generalized minimum mean spuare error estimators and masking properties of the auditory system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, November 2006.

[7] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech Audio Processing*, vol. 11, November 2003.

[8] SangGyun Kim Jong Uk Kim and Chang D. Yoo, "The incorporation of masking threshold to subspace speech enhancement," *Proc. IEEE Int. Conf. Acoustic, Speechm and Signal Processing*, vol. 1, April 2003.

[9] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. of IEEE*, vol. 88, April 2000.

[10] Jan S. Erkelens and Richard Heusdens, "Fast noise tracking based on recursive smoothing of mmse noise power estimates," *Proc. IEEE Int. Conf. Acoustic, Speechm and Signal Processing*, vol. 1, April 2008.