A SPEECH SPECTRAL ESTIMATOR USING ADAPTIVE SPEECH PROBABILITY DENSITY FUNCTION

Arata Kawamura, W. Thanhikam, and Youji Iiguni

Dept. of Systems Innovation, Graduate School of Engineering Science, Osaka University
560-8531, Toyonaka, Osaka, Japan
phone: + (81) 6-6850-6377, fax: + (81) 6-6850-6377, email: kawamura@sys.es.osaka-u.ac.jp

ABSTRACT

In this paper, we propose a speech spectral enhancer based on the MAP estimation using variable speech probability density function (PDF). The proposed speech enhancement algorithm adaptively changes the speech PDF used in the MAP estimation according to the observed spectral power. In speech segments, the speech spectral density approaches a Rayleigh distribution to keep the quality of the enhanced speech. In non-speech segments, it approaches a delta function to reduce noise effectively. The proposed technique is effective in suppressing residual noise well. Computer simulation results show that the proposed speech enhancer is superior to the conventional methods in the noise reduction capability.

1. INTRODUCTION

Speech enhancement technique is necessary in a wide range of applications including mobile communication and speech recognition systems. Single microphone speech enhancement has been a research topic for decades [1]-[5], and one of the famous methods in the spectral domain is the spectral subtraction algorithm proposed by Boll [1]. Unfortunately it provides annoying artifacts called "musical noise" in the enhanced speech. Ephraim and Malah have thus proposed an effective method for removing musical noise, called the MMSE-STSA (minimum mean square error short time spectral amplitude) method [2]. The MMSE-STSA becomes a strong tool of speech enhancement. The improved methods are also proposed in [3], and a noise suppressor employing the algorithm is implemented in a cellular phone [3].

The MMSE-STSA method minimizes the mean square error of the short time spectral amplitude. This method assumes that the discrete Fourier Transform (DFT) coefficient of speech obeys Gauss probability density function (PDF). The PDF of the speech spectral amplitude then results in Rayleigh distribution. However, Martin has pointed out that the DFT coefficient is more likely to fit a Gamma PDF and has shown that the estimator designed under Gamma model decreases the mean square error as compared with the one under Gaussian model [4]. However, neither of the speech models fits the actual DFT coefficient of the speech sufficiently.

Lotter and Vary have proposed an efficient speech enhancement method using the joint Maximum a Posteriori (MAP) estimation with a parametric PDF of the speech spectral amplitude [6]. This PDF is modeled by a single set of parameters estimated from a large amount of actual speech data. The enhanced speech is obtained by applying the MAP estimation rule with the derived PDF. The performance of the MAP estimator is superior to that of the MMSE-STSA method in terms of noise attenuation. However, the speech intelligibility in a speech segment is not sufficiently good, because the parameters of the PDF are determined, regardless whether the observed signal is in a speech or a non-speech segment.

To solve this problem, we have previously proposed an adaptive algorithm for speech enhancement, so that it adaptively changes the PDF parameters depending whether the observed signal is in a speech segment or in a non-speech segment. In a speech segment, we adjust the parameters so that the speech PDF approaches a Rayleigh distribution under the assumption that the speech PDF in speech segment approaches a Rayleigh distribution [7]. In a non-speech segment, since the speech signal does not exist, the speech PDF can be assumed a Delta function. In this case, we adjust the PDF parameters so that the speech PDF approaches the Delta function to strongly reduce the noise. Unfortunately, in [7], the one PDF parameter is fixed, while the another one is adaptively changed. Although the simulation results provided a good performance of this method, the approximation of the Delta function in non-speech segments was very rough. As a result, the noise suppression effect of this method was not sufficiently exercised.

To obtain more faithful approximation of the Delta function in non-speech segments, we propose an adaptive algorithm that adaptively changes the both of the two PDF parameters. Since the proposed adaptive speech PDF can considerably approaches the Delta function, the speech enhancer can suppress a large amount of noise signal from an observed signal especially in non-speech segments. Simulation results show that the noise reduction capability of the proposed method is superior to the other conventional methods.

2. CONVENTIONAL SPEECH ENHANCEMENT SYSTEM

2.1 Structure of the Speech Enhancer

Fig.1 shows the structure of the conventional speech enhancement system based on the MAP estimation [6], where $x(t)$ denotes the input signal at time t. After $x(t)$ is segmented and windowed, the spectral amplitude $|X_n(k)|$ and the phase $\angle X_n(k)$ are calculated by using the fast Fourier transform (FFT), where n and k denote the analysis frame number and the frequency index, respectively. Using the noise power spectrum $\lambda_n(k)$, the *a priori* SNR $\xi_n(k)$ and the *a posteriori* SNR $\gamma_n(k)$ are calculated as

$$\xi_n(k) = \frac{E[|S_n(k)|^2]}{\lambda_n(k)}, \quad (1)$$

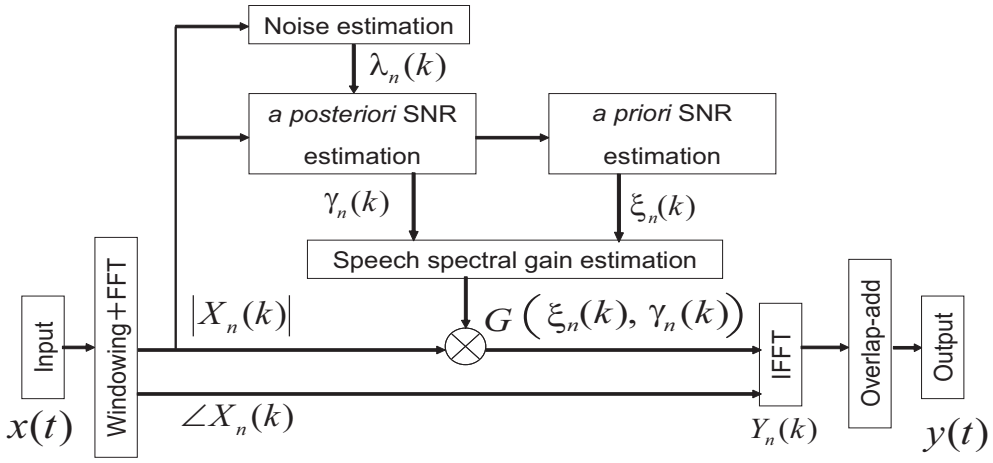


Figure 1: Structure of the speech enhancement system.

$$\gamma_n(k) = \frac{|X_n(k)|^2}{\lambda_n(k)}, \quad (2)$$

where $S_n(k)$ denotes the speech spectrum and $E[\cdot]$ is an expectation operator. Since $E[|S_n(k)|^2]$ is not directly available, $\xi_n(k)$ is calculated by using the following decision-directed method [2]:

$$\xi_n(k) = \alpha\xi_{n-1}(k) + (1-\alpha)\cdot\max[\gamma_n(k)-1, 0], \quad (3)$$

where α is a forgetting factor satisfying $0 < \alpha < 1$. The spectral gain function $G_n(k) = G(\xi_n(k), \gamma_n(k))$, which is characterized by $\xi_n(k)$ and $\gamma_n(k)$, magnifies the speech spectral amplitude. The enhanced speech spectrum $Y_n(k)$ is then expressed as

$$\begin{aligned} Y_n(k) &= G_n(k)X_n(k) \\ &= G_n(k)|X_n(k)|\exp(j\angle X_n(k)), \end{aligned} \quad (4)$$

where $j = \sqrt{-1}$. The enhanced speech $y(t)$ is obtained from $Y_n(k)$ by using the inverse FFT with the overlap-add method.

The speech enhancement system includes two important parts, namely the noise estimation and the spectral gain estimation. If either does not work well, serious distortion or insufficient residual noise occurs in the enhanced speech. Accurate noise and speech estimators are indispensable to maintain good quality of the enhanced speech.

2.2 Noise Estimation

Noise estimation is also an important issue in speech enhancement systems. A weighted noise estimator is proposed [3], and it exhibits a better performance than the methods based on minimum statistics [8]. We shall briefly describe the weighted noise estimation method proposed in [3]. This method recursively updates the noise power spectrum by

$$\lambda_n(k) = \begin{cases} \beta\lambda_{n-1}(k) + (1-\beta)H_n(k)|X_n(k)|^2, & H_n(k) > 0 \\ \lambda_{n-1}(k), & H_n(k) = 0 \end{cases},$$

where β is a forgetting factor satisfying $0 < \beta < 1$ and $H_n(k)$ is the weight on the power spectrum $|X_n(k)|^2$. The weight coefficient is designed so that it is almost inversely proportional to the estimated SNR:

$$\tilde{\gamma}_n(k) = 10\log_{10}\left(\frac{|X_n(k)|^2}{\lambda_{n-1}(k)}\right). \quad (5)$$

Then, $H_n(k)$ is empirically chosen as

$$H_n(k) = \begin{cases} 1, & \tilde{\gamma}_n(k) \le 0 \\ -\frac{1}{\gamma_z}\tilde{\gamma}_n(k) + 1, & 0 < \tilde{\gamma}_n(k) \le \theta_z \\ 0, & \theta_z < \tilde{\gamma}_n(k) \end{cases}, \quad (6)$$

where γ_z is a constant to decide a slope of graph and θ_z is a threshold to eliminate an unreliable $\tilde{\gamma}_n(k)$.

2.3 Gain Estimation

We shall explain the gain estimation method based on the joint MAP method [6]. We here omit the subscripts, the frame number n and the frequency number k for simplicity. Let $p(S)$ and $p(\angle S)$ denote the PDFs of the speech spectral amplitude and the phase, respectively. $p(X)$ denotes the PDF of the input DFT coefficient and $p(S, \angle S|X)$ is the conditional joint PDF. The joint MAP estimator gives the speech spectral amplitude $\hat{S}$ that maximizes $p(S, \angle S|X)$ as follows:

$$\begin{aligned} Y &= \arg\max_S \; p(S, \angle S|X) \\ &= \arg\max_S \; \frac{p(X|S, \angle S)p(S, \angle S)}{p(X)}. \end{aligned} \quad (7)$$

We assume that $p(X|S, \angle S)$ is Gaussian and that $p(S)$ and $p(\angle S)$ are statistically independent. Moreover, $p(S)$ and $p(\angle S)$ are assumed to be

$$p(S) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{S^\nu}{\sigma_S^{\nu+1}}\exp\left(-\mu\frac{S}{\sigma_s}\right), \quad (8)$$

$$p(\angle S) = \frac{1}{2\pi}, \quad (9)$$

where $\Gamma(\cdot)$ denotes Gamma function, σ_S^2 is the variance of the speech spectrum. The PDF $p(S)$ is characterized by positive parameters μ and ν. Substituting Eqs.(8) and (9) into Eq.(7), and solving it for S, we have

$$Y = GX \quad (10)$$

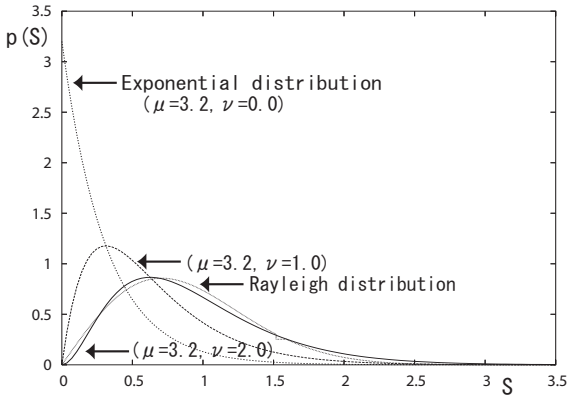


Figure 2: Parametric speech PDF.

with

$$G = u + \sqrt{u^2 + \frac{\nu}{2\gamma}}, \qquad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}. \tag{11}$$

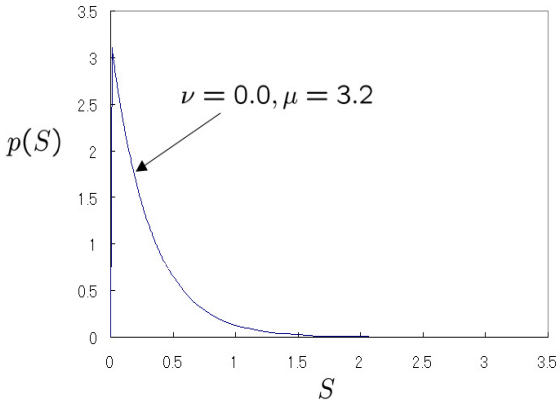
This G is the Lotter's spectral gain [6].

3. SPEECH SPECTRAL ESTIMATOR USING ADAPTIVE PDF MODEL

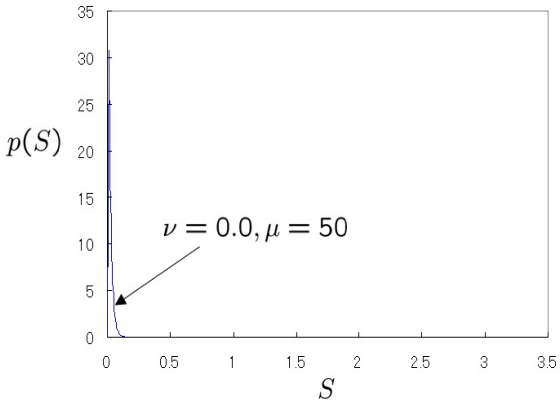
3.1 Derivation of the proposed algorithm

In the parametric speech PDF derived by Lotter and Vary, the two parameters μ and ν are fixed [6]. However, an actual speech signal involves the both of speech segment (speech exists) and non-speech segment (speech does not exists). Clearly, in the non-speech segments, the speech PDF becomes a Delta function because all speech data are zero. On the other hand, in the speech segments, the speech PDF can be approximated as a Rayleigh PDF [7]. Hence, we have previously proposed the adaptive speech PDF model based on Eq.(8). Our approach is to adaptively change the PDF with the parameter ν. Fig. 2 shows some parametric PDFs made by different ν, where the other parameter μ is fixed to 3.2. This result supports that the PDF with variable ν can provide the Rayleigh PDF in speech segments and an exponential PDF in non-speech segments. Unfortunately, in the non-speech segments, the actual speech PDF is not identical to the exponential PDF. So, the more appropriate PDF is the Delta function. Since the speech enhancement algorithm proposed in [7] cannot use the appropriate speech PDF in non-speech segments, it produced residual noises in the enhanced speech.

To solve this problem, we adaptively change the both of two parameters ν and μ to approximate the Delta function in non-speech segments. μ is the parameter to adjust the descent of the PDF, while ν controls its ascent. Fig. 3.1 (a) shows the most steep ascent with $\nu = 0$. It is impossible to get more steep ascent. To approximate the Delta function, we have to get more steep decent. A large value of μ gives such effect. Fig. 3.1 (b) shows the parametric speech PDF with $\mu = 50$ and $\nu = 0$. We see from this result that the obtained PDF considerably approaches to the Delta function in comparison to the conventional one. Hence, decreasing ν and increasing μ in non-speech segments may give a good performance for noise reduction.



(a) Conventional speech PDF model in noise segment [7] ($\mu = 3.2$, $\nu = 0.0$).



(b) Proposed speech PDF model in noise segment ($\mu = 50$, $\nu = 0.0$).

Figure 3: Speech PDF models.

We propose the following adaptive parameters ν_n and μ_n.

$$\nu_n = \begin{cases} 0, & \tilde{\nu}_n \le 0 \\ \tilde{\nu}_n, & \text{otherwise} \end{cases}, \tag{12}$$

$$\tilde{\nu}_n = A \cdot \log_{10} R_n, \tag{13}$$

$$\mu_n = \begin{cases} 0, & \tilde{\mu}_n \le 0 \\ \tilde{\mu}_n, & \text{otherwise} \end{cases}, \tag{14}$$

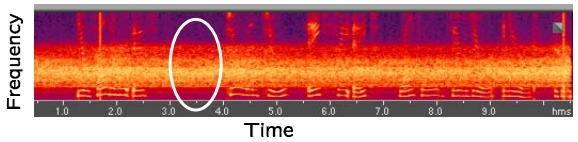
$$\tilde{\mu}_n = B/R_n \tag{15}$$

$$R_n = \frac{\sum_{k=0}^{M-1} |X_n(k)|^2}{\sum_{k=0}^{M-1} \lambda_n(k)}, \tag{16}$$

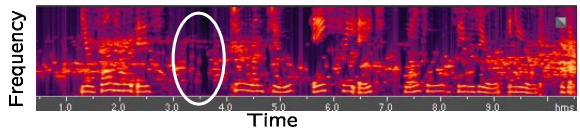
where M is the FFT size, and A and B are constants for adaptation. We see from Eq. (14) and (15) that μ_n changes reciprocal to the SNR, R_n. Substituting ν_n and μ_n into Eq.(11), we have the speech spectral gain.

3.2 Simulation

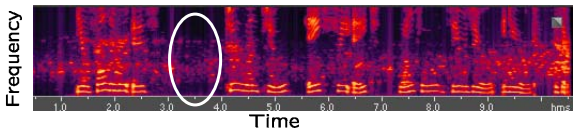
We carried out computer simulation for confirming the effectiveness of the proposed method. In the simulation, we added a tunnel noise to a female speech signal to make an observed signal. We put $\alpha = 0.98$, $\beta = 0.92$, $A = 0.5$, $B = 100$ for



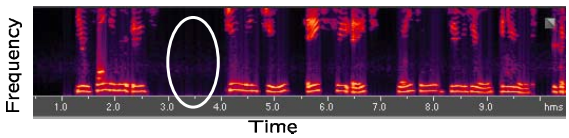
(a) Observed signal



(b) Lotter's method [6]



(c) Tsukamoto's method [7]



(d) Proposed method

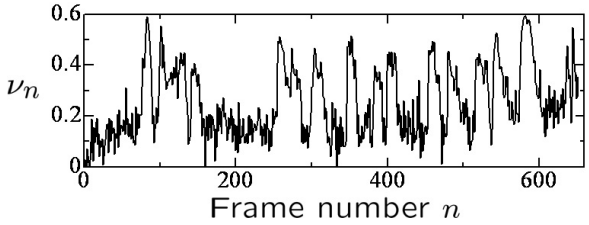
Figure 4: Results of speech enhancement

speech enhancer, and $\theta_z = 7$ and $\gamma_z = 10$ for noise estimator. The proposed method compared with Lotter's method [6] and Tsukamoto's methods [7] Fig. 4(a)–(d) show the spectrograms of simulation results, where horizontal axis shows the time and vertical axis shows the frequency. From Fig. 4(b), Lotter's method gives a good speech enhancement capability although residual noises arise especially in non-speech segments. In Fig. 4(c), the residual noise is more suppressed than the result of (b). But, the persistent residual noises are perceptible. We see from Fig. 4(d) that the proposed method reduces the noise sufficiently small, and it is almost not perceptible.

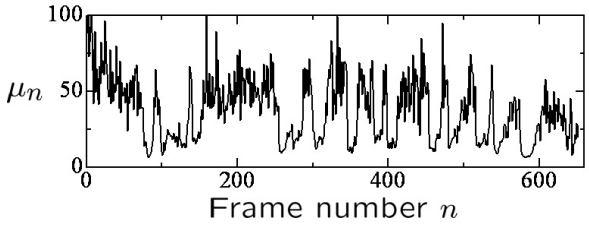
Next, we show the trajectories of the adaptive parameters ν_n and μ_n in Fig. 5(a) and (b). Here, (a) shows the trajectory of ν_n. Since ν_n is proportional to the SNR, ν_n becomes large when the observed signal includes speech signal, and becomes small in absence of the speech signal. Fig. 5(b) shows the trajectory of μ_n. We see from this results that μ_n increases when ν_n decreases and vice versa, because μ_n is reciprocal to the SNR. The pameter $\nu \approx 0.1$ and $\mu \approx 50$ around 200 frames that is in non-speech segment. In this case, the adaptive speech PDF approximates the Delta function like Fig. 3.1(b).

4. CONCLUSION

We have proposed a speech spectral enhancer using adaptive speech PDF that are controlled with two parameters ν and μ. In the proposed method, the speech PDF in non-speech segments approaches to Delta function by decreasing ν and increasing μ. Since the non-speech segment is unknown, we adaptively change the two parameters based on the SNR of the observed signal. Simulation results show that the proposed method effectively reduced the noise more than that



(a) Trajectory of ν_n



(b) Trajectory of μ_n

Figure 5: Trajectories of adaptive parameters

of the conventional methods, especially in non-speech segments.

REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol.ASSP-32, no.6, pp.1109-1121, Dec. 1984.

[3] M. Kato, A. Sugiyama and M. Serizawa, "Noise Suppression with high speech quality based on weighted noise estimation and MMSE STSA," *IEICE Trans. Fundamentals*, vol.E85-A, no.7, pp.1710-1718, Jul. 2002.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio process.*, vol.13, no.5, pp.845-856, Sep.2005.

[5] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, pp.1043-1051, Oct. 2003.

[6] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Applied Signal Processing*, pp.1110-1126, July 2005.

[7] Y. Tsukamoto, A. Kawamura, and Y. Iiguni, "Speech Enhancement Based on MAP Estimation Using a Variable Speech Distribution," *IEICE Trans. Fundamentals*, Vol.E90-A, No.8, pp.1587-1593, Aug. 2007.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio process.*, vol.9, no.5, pp.504-512, Jul. 2001.