International Scholarly Research Network
ISRN Mechanical Engineering
Volume 2012, Article ID 919234, 9 pages
doi:10.5402/2012/919234

Review Article

Single Channel Speech Enhancement Techniques in Spectral Domain

Arata Kawamura, Weerawut Thanhikam, and Youji Iiguni

Department of Systems Innovations, Graduate School of Engineering Science, Osaka University, 1–3 Machikaneyama, Osaka, Toyonaka 560-8531, Japan

Correspondence should be addressed to Weerawut Thanhikam, weerawut@sip.sys.es.osaka-u.ac.jp

Received 13 February 2012; Accepted 30 April 2012

Academic Editor: D. Aggelis

This paper presents single-channel speech enhancement techniques in spectral domain. One of the most famous single channel speech enhancement techniques is the spectral subtraction method proposed by S.F. Boll in 1979. In this method, an estimated speech spectrum is obtained by simply subtracting a preestimated noise spectrum from an observed one. Hence, the spectral subtraction method is not concerned with speech spectral properties. It is well known that the spectral subtraction method produces an annoying artificial noise in the extracted speech signal. On the other hand, recent successful speech enhancement methods positively utilize the speech property and achieve an efficient speech enhancement capability. This paper presents a historical review about some speech estimation techniques and explicitly states the difference between their theoretical background. Moreover, to evaluate their speech enhancement capabilities, we perform computer simulations. The results show that an adaptive speech enhancement method based on MAP estimation gives the best noise reduction capability in comparison to other speech enhancement methods presented in this paper.

1. Introduction

In recent years, speech enhancement is required in a wide area of applications including mobile communication and speech recognition systems, where the major example is a cell-phone as shown in Figure 1. Many speech enhancement methods have been established in decades [1–15]. These speech enhancement techniques can be classified to time domain methods and spectral domain methods. Recent major speech enhancement techniques are of the spectral domain method which is preferably used in a cell phone. In this paper, we focus on the spectral domain speech enhancement techniques that employ a single microphone.

The spectral subtraction method [3] is one of the most popular methods among numerous noise reduction techniques in spectral domain. This method achieves noise reduction by simply subtracting a pre-estimated noise spectral amplitude from an observed spectral amplitude, where the spectral phase is not processed. The spectral subtraction method is easy for implementation and effectively reduces stationary noises. However, it incurs an artificial noise, called musical noise, which is caused from speech estimation errors. Because the spectral subtraction method is not concerned with speech spectral information, it often gives estimation errors. Ephraim and Malah have proposed the MMSE-STSA (Minimum Mean Square Error-Short-Time Spectral Amplitude) method [4] which utilizes a speech PDF (Probability Density Function) and a noise PDF. In the literature in [4], the speech and noise PDFs were modeled by Rayleigh and Gauss density functions, respectively. This method gives an optimal solution of the estimated speech signal in the sense of MMSE-STSA (the solution may change to Wiener filter [5] if we assume Gauss distributions for both of the speech and noise PDFs). Although the MMSE-STSA method gives an estimated speech signal with less musical noise, it requires more complicated computations, for example, the solution required to calculate the modified Bessel function. Moreover, as pointed out by some researchers, real speech histograms do not fit to Rayleigh function employed in [4].

A more efficient method that is based on a maximum *a posteriori* (MAP) estimation has been established by Lotter

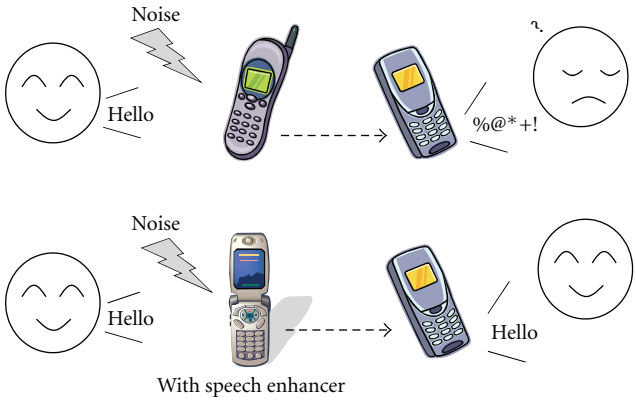


Figure 1: Application of speech enhancement.

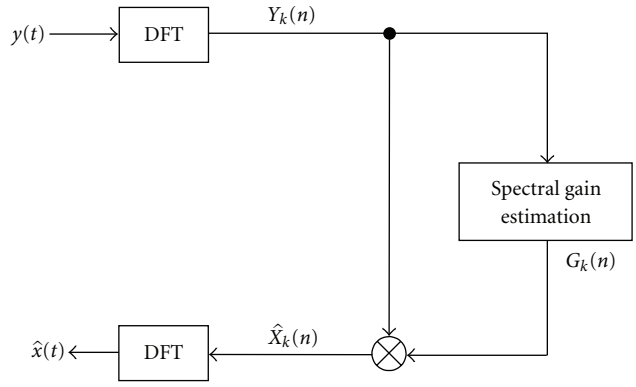


Figure 2: General speech enhancement system.

and Vary [11]. Lotter and Vary modeled the speech PDF by a parametric super-Gaussian function, controlled by two shape parameters. The parametric super-Gaussian function has been developed from a histogram made from a large amount of real speech data in a single narrow SNR (Signal to Noise Ratio) interval. The noise suppression capability of this method is superior to the Wiener filter. However, the residual noise is still persistently perceived. Andrianakis and White were aware that the speech PDF may change in some SNR intervals [12]. They utilized three histograms made from speech signals in three different narrow SNR intervals and approximate them with Gamma density function. As reported in [12], changing these three speech PDFs according to the SNR can improve the noise reduction capability. While Andrianakis discretely changes the speech PDF, Tsukamoto et al. continuously change the speech PDF according to the SNR [13]. They employed the parametric super-Gaussian function proposed in [11] and adaptively changed its shape parameters according to the SNR. Recently, Thanhikam et al. [16] sophisticated this approach by making and evaluating many real speech histograms made from various narrow SNR intervals. As shown in [16], this method has a very strong noise reduction capability in comparison to other traditional speech enhancement methods, and hence it is effective especially in low SNR environments.

In the following sections, we present a historical review of useful speech enhancement methods mentioned above and compare their speech enhancement capabilities by computer simulations.

2. Speech Enhancement in Spectral Domain

This section presents several speech enhancement techniques including both traditional methods and recent methods. Particularly, we will carefully explain the difference between them.

2.1. General Speech Enhancement System. Firstly, we explain about a general single-channel speech enhancement system in spectral domain.

We assume that an observed signal is a sum of a speech signal and a noise signal given as

$$y(t) = x(t) + d(t), \tag{1}$$

where $y(t)$ is the observed signal at time t. $x(t)$ and $d(t)$ denote the speech signal and the noise signal, respectively. We assume that $x(t)$ is uncorrelated with $d(t)$ through the paper. Taking the DFT of (1), we have

$$Y_k(n) = \sum_{t=nQ}^{nQ+N-1} y(nQ+t)h(t)e^{-j2\pi nk/N} \quad (k = 0, 1, \ldots, N-1), \tag{2}$$

where N, n, and k denote the frame length, the frame index, and the frequency bin index, respectively. The analysis frame is shifted by Q samples, where $Q = N/2$ is used through the paper. The function $h(t)$ denotes an analysis window function, where the Hanning window of size N is used as $h(t)$. The DFT spectrum $Y_k(n)$ can be rewritten as

$$Y_k(n) = X_k(n) + D_k(n), \tag{3}$$

where $X_k(n)$ and $D_k(n)$ are the kth spectra of $x(t)$ and $d(t)$, respectively. The enhanced speech spectrum $\hat{X}_k(n)$ is given as

$$\hat{X}_k(n) = G_k(n)Y_k(n), \tag{4}$$

where $G_k(n)$ is a spectral gain. The enhanced speech is obtained as the observed signal $Y_k(n)$ multiplied by the spectral gain $G_k(n)$. Hence, speech enhancement capability depends only on the spectral gain.

A general speech enhancement system can be illustrated in Figure 2, where the value of the spectral gain $G_k(n)$ depends on an employed speech enhancement algorithm. We see from (3) and (4) that the ideal spectral gain is given as

$$G_{k,\mathrm{opt}}(n) = 1 - \frac{D_k(n)}{Y_k(n)}. \tag{5}$$

This spectral gain perfectly provides the original speech signal as the enhanced speech. Since the ideal spectral gain above cannot be directly obtained from $Y_k(n)$, we have to approximate the ideal spectral gain by introducing additional assumptions for the speech or the noise signals.

In the following sections, we give some typical spectral gains which have been derived from respective assumptions for the speech or the noise. For avoiding redundant expressions, we omit the indices n and k if they do not play an important role.

2.2. Spectral Subtraction. The most simple and famous speech enhancement technique is the spectral subtraction proposed by Boll in 1979 [3]. This method just subtracts a pre-estimated noise spectral amplitude from an observed one to obtain the estimated speech spectral amplitude. In the spectral subtraction method, the spectral phase is not modified; that is, the estimated speech spectral phase is identical to the observed one. This is based on the fact that the spectral phase is unimportant in comparison to the spectral amplitude in human speech perception [17]. The spectral subtraction method is achieved by using the following spectral gain.

$$G_{\mathrm{SS}} = 1 - \frac{\left|\hat{D}\right|}{|Y|}, \tag{6}$$

where $|\hat{D}|$ is the pre-estimated noise spectral amplitude. Usually, we choose $|\hat{D}| = E[|D|]$. We note that formula (6) is an absolute version of (5).

The spectral subtraction is not concerned with speech spectral property. As a result, the estimated speech signal includes many estimation errors. The estimation error produces an isolated spectrum in the estimated speech signal. This noise is called "musical noise" and it is perceived as an annoying sound for human. To obtain an estimated speech signal with less musical noise, we should introduce a speech property into speech enhancement scheme. In the following sections, we present some speech enhancement methods taking into account speech probabilistic properties.

2.3. Wiener Filter. In this section, we explain the Wiener filter [5] which utilizes both of the speech and the noise spectral probabilistic properties. It is well known that the Wiener filter provides an estimated speech signal with less musical noise in comparison to the spectral subtraction method.

To derive the Wiener filter, we assume that the speech spectrum X is uncorrelated with the noise spectrum D and $E[X] = 0$, $E[|X|^2] = \sigma_x^2$, $E[D] = 0$, $E[|D|^2] = \sigma_d^2$. The Wiener filter is obtained by minimizing the following cost function:

$$J = E\left[\left|X - \hat{X}\right|^2\right] = E\left[|X - GY|^2\right], \tag{7}$$

where $E[\cdot]$ denotes the expected value. We can rewrite J as

$$\begin{aligned} J &= E\left[|X|^2\right] + |G|^2 E\left[|Y|^2\right] - GE[XY^*] - G^* E[X^* Y] \\ &= \sigma_x^2 + |G|^2\left(\sigma_x^2 + \sigma_d^2\right) - G\sigma_x^2 - G^*\sigma_x^2. \end{aligned} \tag{8}$$

Differentiating J with respect to G^* gives

$$\frac{\partial J}{\partial G^*} = G\left(\sigma_x^2 + \sigma_d^2\right) - \sigma_x^2. \tag{9}$$

Putting (9) to zero and solving it with respect to G, we have the spectral gain of the Wiener filter given as

$$G_{\mathrm{Wiener}} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_d^2} = \frac{\xi}{1+\xi}, \tag{10}$$

where $\xi = \sigma_x^2/\sigma_d^2$ is the *a priori* SNR. The Wiener filter requires one parameter ξ or two variances σ_x^2 and σ_d^2.

2.4. MMSE-STSA Method. In this section, we explain a historically important speech enhancement method, that is, the MMSE-STSA method [4] proposed by Ephraim and Malah in 1984. Ephraim and Malah have proposed not only an efficient spectral gain, but also an efficient estimation technique to get the *a priori* SNR.

The MMSE-STSA method is derived by minimizing a conditional mean square value of the short time spectral amplitude. The cost function to be minimized is given by

$$\begin{aligned} J_{\mathrm{MMSE}} &= E\left[\left|X - \hat{X}\right|^2 \mid Y\right] \\ &= \int_{-\infty}^{\infty} |X|^2 p(X \mid Y)dx + \left|\hat{X}\right|^2 - \hat{X}\int_{-\infty}^{\infty} X^* p(X \mid Y)dx \\ &\quad - \hat{X}^*\int_{-\infty}^{\infty} X p(X \mid Y)dx, \end{aligned} \tag{11}$$

where $p(X \mid Y)$ denotes the conditional PDF of X. The estimated speech spectrum which minimizes J_{MMSE} is given as

$$\hat{X}_{\mathrm{MMSE}} = \int_{-\infty}^{\infty} X p(X \mid Y)dx = E[X \mid Y]. \tag{12}$$

As shown in [6], when we assume $p(X)$ and $p(D)$ as Gauss functions, (12) produces the Wiener filter again. On the other hand, Ephraim and Malah considered the PDFs of the speech spectral amplitude and phase, that is, $p(|X|)$ and $p(\angle X)$. They assumed that $p(|X|)$ and $p(\angle X)$ as the Rayleigh distribution and the uniform distribution, respectively [18]. They assumed $p(D)$ as the Gauss function, where the noise variance σ_d^2 is assumed to split equally into real and imaginary parts. These PDFs are expressed as

$$p(|X|) = \frac{2|X|}{\sigma_x^2}\exp\left\{-\frac{|X|^2}{\sigma_x^2}\right\}, \tag{13}$$

$$p(\angle X) = \frac{1}{2\pi}, \tag{14}$$

$$p(Y \mid X) = \frac{1}{\pi\sigma_d^2}\exp\left\{-\frac{|Y - X|^2}{\sigma_d^2}\right\}, \tag{15}$$

where $P(Y \mid X)$ is corresponding to $p(D)$. Assuming $p(X) = p(|X|)p(\angle X)$, we can calculate (12) by using the relation $p(X \mid Y) = p(Y \mid X)p(X)/p(Y)$. After tedious and complex computations, the spectral gain is given as [4]

$$\begin{aligned} G_{\mathrm{MMSE}} &= \frac{(\pi\nu)^{1/2}}{2\gamma}\exp\left(\frac{-\nu}{2}\right) \\ &\quad \times \left[(1+\nu)I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right], \end{aligned} \tag{16}$$

where $I_i(\cdot)$ is the modified Bessel function of order i and

$$\nu = \frac{\xi}{1+\xi}\gamma, \qquad \gamma = \frac{|Y|^2}{\sigma_d^2}. \tag{17}$$

Here, γ is called as the *a posteriori* SNR. As shown in [4], the optimal spectral phase in the sense of MMSE-STSA

is identical to the observed one. Hence, G_{MMSE} is also a real value. The MMSE-STSA solution, G_{MMSE}, is completely characterized by σ_d^2, ξ, and γ. When the noise variance σ_d^2 is known or can be estimated, γ is simply obtained by the observed spectrum. On the other hand, estimating the *a priori* SNR ξ is difficult, although it needs to be required for many other spectral speech enhancers. One of the valuable contributions in [4] is to present a useful estimation method of ξ, called the decision-directed method. We will show and use it to estimate ξ in Section 3.

2.5. MAP Estimation Method. As confirmed in many literatures, the spectral gain G_{MMSE} derived in the previous section is superior to the spectral subtraction method. But G_{MMSE} is not easy to implement due to a large amount of computational complexity. Indeed, we can obtain a more theoretically relevant and reasonable spectral gain from the same cost function shown in (11). The MMSE-STSA method has chosen $\hat{X} = E[X \mid Y]$ to minimize (11). Here, we can note that $E[X \mid Y]$ is the best choice when the PDF is an even function like a Gauss function. Because the Rayleigh distribution is asymmetric function, $\hat{X} = E[X \mid Y]$ is not appropriate. The MAP estimation method [6] denotes that the best choice for minimizing (11) is to employ the speech spectrum maximizing $p(X \mid Y)$.

To illustrate the difference between the MMSE-STSA solution and the MAP solution, we show an example of the specific PDF. Figures 3(a) and 3(b) show the Gauss and Rayleigh distributions, respectively. Here, the horizontal axis denotes the value of an argument x and the vertical axis is a PDF $p(x)$. The vertical dotted lines denote the argument values giving the mean value and maximum value of $p(x)$, respectively. The former value is corresponding to the MMSE-STSA solution and the latter value is corresponding to the MAP solution. As shown in Figure 3(a), the MMSE-STSA solution is identical to the MAP solution for the Gauss distribution which is an even function. On the other hand, the solutions of them are different for the asymmetric Rayleigh distribution as shown in Figure 3(b). Obviously, we should choose the solution of the MAP estimation rather than the MMSE-STSA solution to minimize the cost function (11).

To obtain the MAP solution, we have to maximize the conditional PDF $p(X \mid Y)$. Based on the Bayes's rule, we have [6]

$$\begin{aligned} p(X \mid Y) &= \frac{p(Y \mid X)p(X)}{p(Y)} \\ &\propto p(Y \mid X)p(X). \end{aligned} \tag{18}$$

The MAP estimation is to find the arguments X which maximize $p(X|Y)$, that is,

$$\begin{aligned} \hat{X} &= \arg\max_X \; p(X \mid Y) \\ &= \arg\max_X \; p(Y \mid X)p(X) \\ &= \arg\max_X \; \ln\{p(Y \mid X)p(X)\}. \end{aligned} \tag{19}$$

We assume the same PDFs from (13) to (15), and $p(X) = p(|X|)p(\angle X)$. After calculating $\ln\{p(Y \mid X)p(X)\}$ and differentiating it with respect to $|X|$ (or $\angle X$), we put the obtained derivative to zero and solve it with respect to $|X|$ (or $\angle X$). Then, we have [6]

$$G_{\mathrm{MAP}} = \frac{\xi + \sqrt{\xi^2 + 2(1+\xi)(\xi/\gamma)}}{2(1+\xi)}. \tag{20}$$

Since the MAP solution of $\angle X$ is identical to the observed spectral phase, G_{MAP} is also a real value. We see that G_{MAP} consists of ξ and γ only; thus its computational complexity is extremely low in comparison to (16).

2.6. Lotter's Spectral Gain. In the previous section, we obtained a MAP solution for speech enhancement under the assumption that the PDF of the speech spectral amplitude can be modeled as the Rayleigh distribution. However, some researchers pointed out that there exists other appropriate speech PDF [8–11]. In 2005, Lotter and Vary have proposed an original speech spectral amplitude PDF. This PDF was derived from a real speech histogram made from a large amount of real speech data. In the same manner as in the previous section, the speech spectral amplitude and phase were separately modeled in [11]. The PDF of the spectral phase was also modeled as the uniform distribution defined in (14). Lotter et al. modeled the PDF of the speech spectral amplitude as a super-Gaussian function represented by

$$p(|X|) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)} \frac{|X|^{\nu}}{\sigma_x^{\nu+1}} \exp\left(-\mu \frac{|X|}{\sigma_x}\right), \tag{21}$$

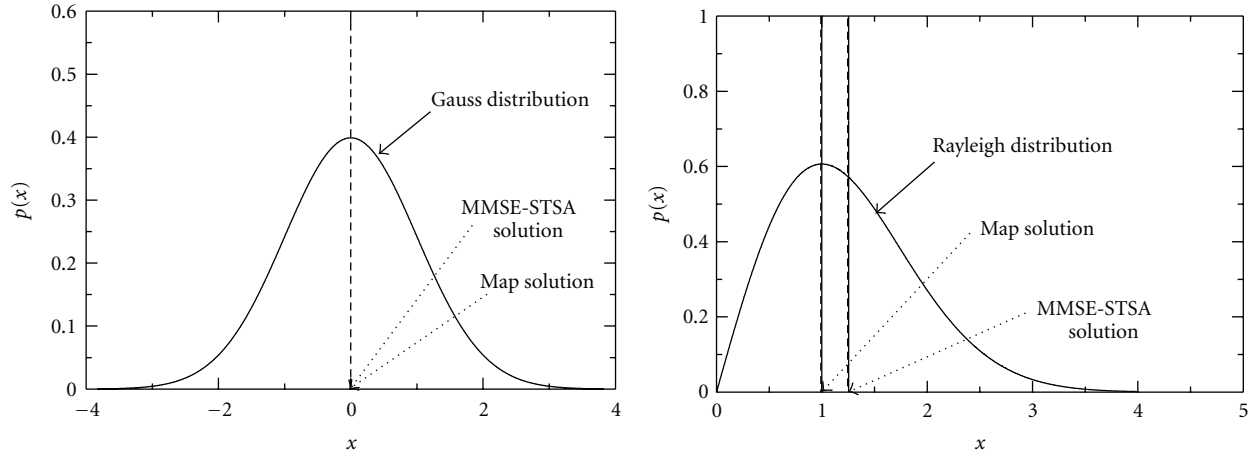
where $\Gamma(\cdot)$ is a Gamma function and μ and ν are the shape parameters which determine the shape of the above PDF. Using (21), (14) and (15), the same procedure in the previous section gives the MAP solution expressed as

$$G_{\mathrm{L{\cdot}MAP}} = u + \sqrt{u^2 + \frac{\nu}{2\gamma}}, \tag{22}$$

$$u = \frac{1}{2} - \frac{\mu}{4}\sqrt{\frac{1}{\gamma\xi}}. \tag{23}$$

The MAP solution of the speech spectral phase is also identical to the observed one, and thus $G_{\mathrm{L{\cdot}MAP}}$ is a real value. Lotter and Vary reported that the most appropriate shape parameters are $\mu = 1.74$ and $\nu = 0.126$ in [11]. The spectral gain $G_{\mathrm{L{\cdot}MAP}}$ also consists of ξ and γ only, hence it is easy to implement.

2.7. Adaptive Speech PDF Method. In [11], the shape parameters of the speech spectral amplitude PDF, μ and ν, had been derived from a large amount of speech data in a single narrow SNR interval. However, in a practical situation, a speech signal includes both activity segments and pause segments. Since the value of the speech spectral amplitude is always zero in the pause segments, we expect that its PDF can be modeled as a delta function. On the other hand, in the activity speech segments, the PDF of the speech spectral amplitude obeys



(a) Gauss distribution (b) Rayleigh distribution

Figure 3: Maximum and mean values for the specific PDFs.

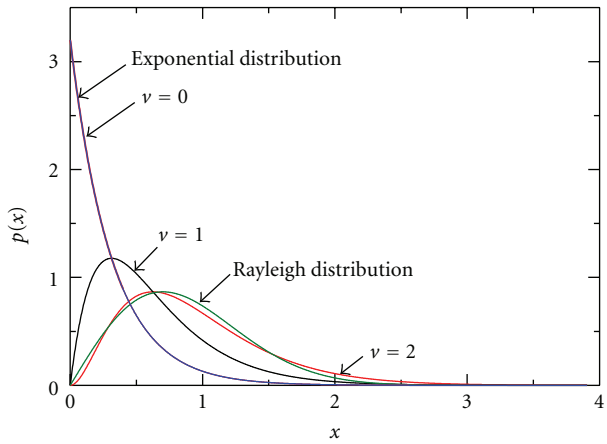


Figure 4: Shape examples of the PDF in (21) with different parameters.

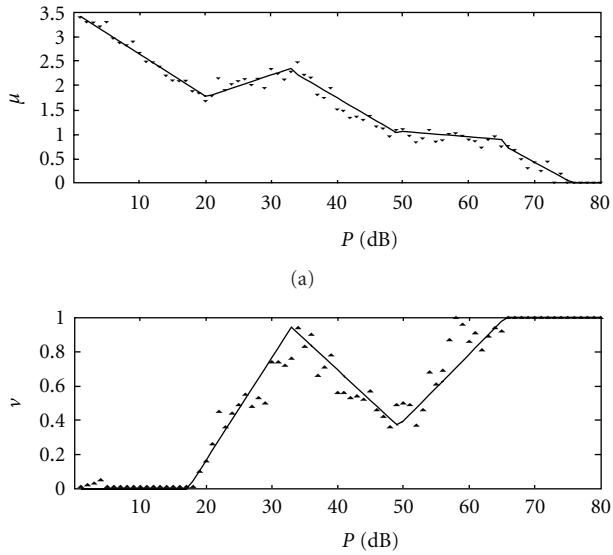


Figure 5: Shape parameter fitting result for the SNR.

other functions. Tsukamoto et al. have noticed the fact and investigated an adaptive method to change the PDF of the speech spectral amplitude, according to the SNR [13]. They have chosen Lotter's PDF defined in (21) as the adaptive PDF, because its shape is easily controlled by ν and μ. Here, we show examples of Lotter's PDF with different shape parameters in Figure 4. We see from this figure that the PDF can fit the exponential distribution and the Rayleigh distribution by adjusting the shape parameters. Utilizing real speech histograms, Tsukamoto et al. derived adaptive shape parameters and showed its effectiveness through the computer simulations [13]. This basic idea is useful for speech enhancement in a practical situation. Unfortunately, a reliability of the derived adaptive shape parameter is comparatively low, because it is derived from only two speech histograms.

To sophisticate Tsukamoto's adaptive shape parameter, Thanhikam et al. have made and evaluated many real speech histograms in various narrow SNR intervals [16]. They tried to fit the speech histograms with (21) and revealed an interesting curve of the shape parameters for narrow SNR intervals. The obtained shape parameters as the fitting results and the derived curve are shown in Figures 5(a) and 5(b), where the narrow SNR was calculated as $P = 10\log_{10}\xi$ [dB]. The lines in the figures denote the curves obtained by the least mean square method. Thes curves denote the relation between the shape parameters and P. Table 1 shows the formulations of the derived shape parameter function for P, where we denote the derived shape parameters by $R_k^{\mu}(n)$ and $R_k^{\nu}(n)$, and

$$F[x] = \begin{cases} x, & x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

Thanhikam et al. used an averaged value of $R_k^{\mu}(n)$ and $R_k^{\nu}(n)$ to determine the present PDF shape of the speech

Table 1: Instantaneous shape parameter functions $R_k^{\mu}(n)$ and $R_k^{\nu}(n)$.

SNR range [dB]	$R_k^{\mu}(n) = F[a_0 P_k(n) + b_0]$		$R_k^{\nu}(n) = F[c_0 P_k(n) + d_0]$	
	a_0	b_0	c_0	d_0
$P_k(n) \le 20$	−0.087	3.50	0.060	−1.04
$20 < P_k(n) \le 33$	0.045	0.84	0.060	−1.04
$33 < P_k(n) \le 49$	−0.079	4.90	−0.035	2.11
$49 < P_k(n) \le 65$	−0.011	1.60	0.039	−1.56
$65 < P_k(n)$	−0.074	5.60	0	1.00

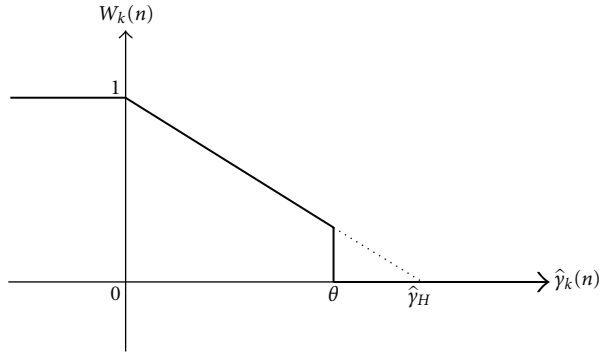


Figure 6: Weighting function.

spectral amplitude. Their "adaptive" MAP solution is as follows:

$$G_k(n) = u_k(n) + \sqrt{u_k^2(n) + \frac{\nu_k(n)}{2\gamma_k(n)}}, \tag{25}$$

$$u_k(n) = \frac{1}{2} - \frac{\mu_k(n)}{4\sqrt{\gamma_k(n)\hat{\xi}_k(n)}}, \tag{26}$$

$$\mu_k(n) = \alpha\mu_k(n-1) + (1-\alpha)R_k^{\mu}(n), \tag{27}$$

$$\nu_k(n) = \&\alpha\nu_k(n-1) + (1-\alpha)R_k^{\nu}(n), \tag{28}$$

where α is the forgetting factor and $\mu_k(n)$ and $\nu_k(n)$ are the adaptive shape parameters. In [16], they put $\alpha = 0.98$, $\mu_k(0) = 20$, $\nu_k(0) = 0$. This paper also use these settings.

In the next section, we compare the speech enhancement capabilities of the spectral gains presented in this paper.

3. Speech Enhancement Simulation

To compare the speech enhancement capabilities of some spectral gains derived in this paper, we firstly explain about common conditions for speech enhancement simulation. After that, we show the simulation results and discuss them.

3.1. Common Conditions. The speech enhancement methods explained in this paper commonly require the noise variance $\sigma_{d,k}^2(n)$, *a priori* SNR $\xi_k(n)$, and *a posteriori* SNR $\gamma_k(n)$. To obtain these parameters, the following estimation methods were used.

Firstly, the noise variance was calculated by using the weighted noise estimator proposed in [19]. This method can update the estimated noise variance even if a speech signal exists. The weighted noise estimator calculates an instantaneous noise power by using the weight $W_k(n)$ as shown in Figure 6. Here, θ and $\hat{\gamma}_H$ are constant values. The literature in [19] recommends that $\theta = 7$ and $\hat{\gamma}_H = 10$. As shown in Figure 6, $W(n)$ is a function of $\hat{\gamma}(n)$ given as

$$\hat{\gamma}(n) = 10\log_{10}\frac{|Y(n)|^2}{\sigma_{d,k}^2(n-1)}. \tag{29}$$

The noise variance $\sigma_{d,k}^2(n)$ is updated as

$$\sigma_{d,k}^2(n) = \beta\sigma_{d,k}^2(n-1) + (1-\beta)W_k(n)|Y_k(n)|^2, \tag{30}$$

where β is a forgetting factor and $\beta = 0.92$ was used.

Next, the *a posteriori* SNR was directly calculated as

$$\gamma_k(n) = \frac{|Y_k(n)|^2}{\sigma_{d,k}^2(n)}. \tag{31}$$

Lastly, the *a priori* SNR was calculated by using the decision-directed method proposed in [4]. The decision-directed method is given by

$$\xi_k(n) = \alpha_{\text{snr}}\frac{\left|\hat{X}_k(n-1)\right|^2}{\sigma_{d,k}(n-1)} + (1-\alpha_{\text{snr}})F[\gamma_k(n)-1], \tag{32}$$

where α_{snr} is a forgetting factor and $\alpha_{\text{snr}} = 0.98$ was used according to [4].

The common speech enhancement system is shown in Figure 7, where the numbers denote the order of the estimation procedures. Of course, the spectral gain estimation is depending on the employed speech enhancement method. In simulations, the observed signal $y(t)$ was a female speech signal $x(t)$ corrupted with a practical tunnel noise $d(t)$ with SNR = 0 dB, where the noise was recorded in a tunnel in an expressway in Japan. All the signals used in the simulations were sampled at 8 kHz, and the DFT size was 256 (the FFT was used instead of the DFT). For objective evaluations, we utilized the SNR defined as

$$\text{SNR} = 10\log_{10}\frac{\sum_{t=0}^{L}|x(t)|^2}{\sum_{t=0}^{L}\left|x(t)-\hat{x}(t)\right|^2}, \tag{33}$$

where L denotes the number of the samples in time domain. It was also utilized the other evaluation function given as [17]

$$\text{LR} = \frac{1}{J}\sum_{j=0}^{J-1}\frac{1}{N}\sum_{k=0}^{N-1}\left(\log\frac{|X_k(n)|}{\left|\hat{X}_k(n)\right|} + \frac{\left|\hat{X}_k(n)\right|}{|X_k(n)|} - 1\right), \tag{34}$$

where J is the number of frames. The LR (Likelihood Ratio) denotes a spectral distance between the original speech and the estimated one, hence the perfect speech estimate gives LR = 0.

3.2. Simulation Results. Speech enhancement simulations were carried out to compare the presented speech enhancement methods. The chosen methods were the spectral

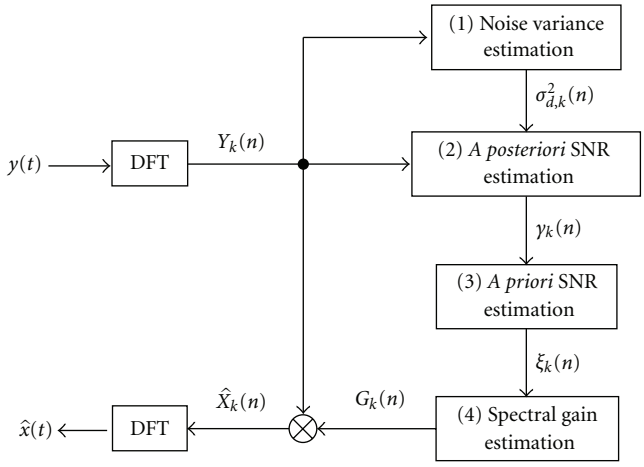


Figure 7: Speech enhancement system.

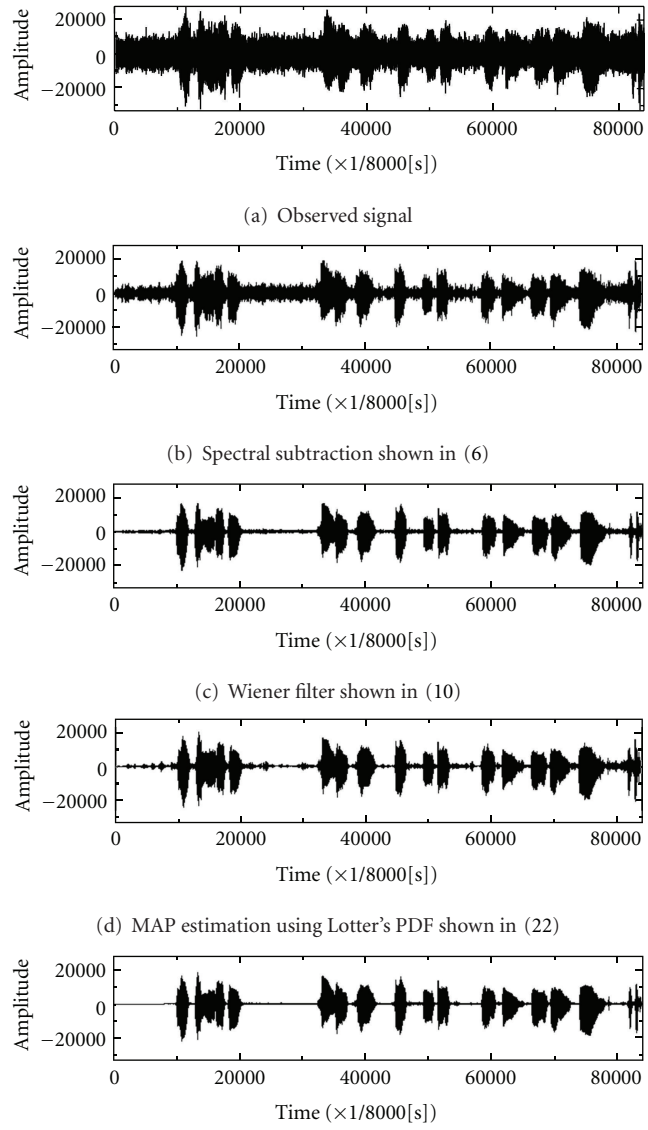
Table 2: Objective evaluation results for noisy signal with SNR = 0 dB.

	S	W	L	A
SNR [dB]	6.8	14.5	12.7	14.8
LR	141.8	29.0	27.9	7.0

S: spectral subtraction in (6), W: Wiener filter in (10), L: Lotter's spectral gain in (22), A: adaptive PDF in (25).

subtraction method [3] and Wiener filter [5] as traditional methods, Lotter's spectral gain [11] as a MAP method using a fixed speech PDF, and the adaptive speech PDF method [16] as the recent method.

Table 2 shows the results of the objective evaluation for each methods, where both of the best SNR and LR results were obtained from the adaptive speech PDF method proposed by Thanhikam et al. [16]. We see from this table that the Wiener filter and Lotter's method also gave comparatively good SNR and LR results in comparison to the spectral subtraction method. The waveforms of the simulation results are shown in Figures 8(a)–8(e), and the respective spectrograms are shown in Figures 9(a)–9(e). From Figures 8(b) and 9(b), we see that the spectral subtraction method provided many residual noises. The main reason of it may be that the spectral subtraction method does not use any speech spectral information. The residual noises are perceived as an annoying musical noise. From Figures 8 and 9(c), we see that the Wiener filter is superior to the spectral subtraction method for speech enhancement. The Wiener filter gave the estimated speech with less musical noise, although the amount of the residual noise was comparatively large. From the waveform shown in Figure 8(d), we can confirm that the Lotter's spectral gain method can effectively reduce the noise in some segments. But its spectrogram shown in Figure 9(d) showed that the Lotter's spectral gain method emphasized isolated spectra, that is, musical noises. As a result, it also causes a perception problem. In Figures 8 and 9(e), such estimation errors cannot be confirmed. It implies that the adaptive PDF method proposed by Thanhikam is appropriate to reduce the noise in speech pause segments.



(a) Observed signal

(b) Spectral subtraction shown in (6)

(c) Wiener filter shown in (10)

(d) MAP estimation using Lotter's PDF shown in (22)

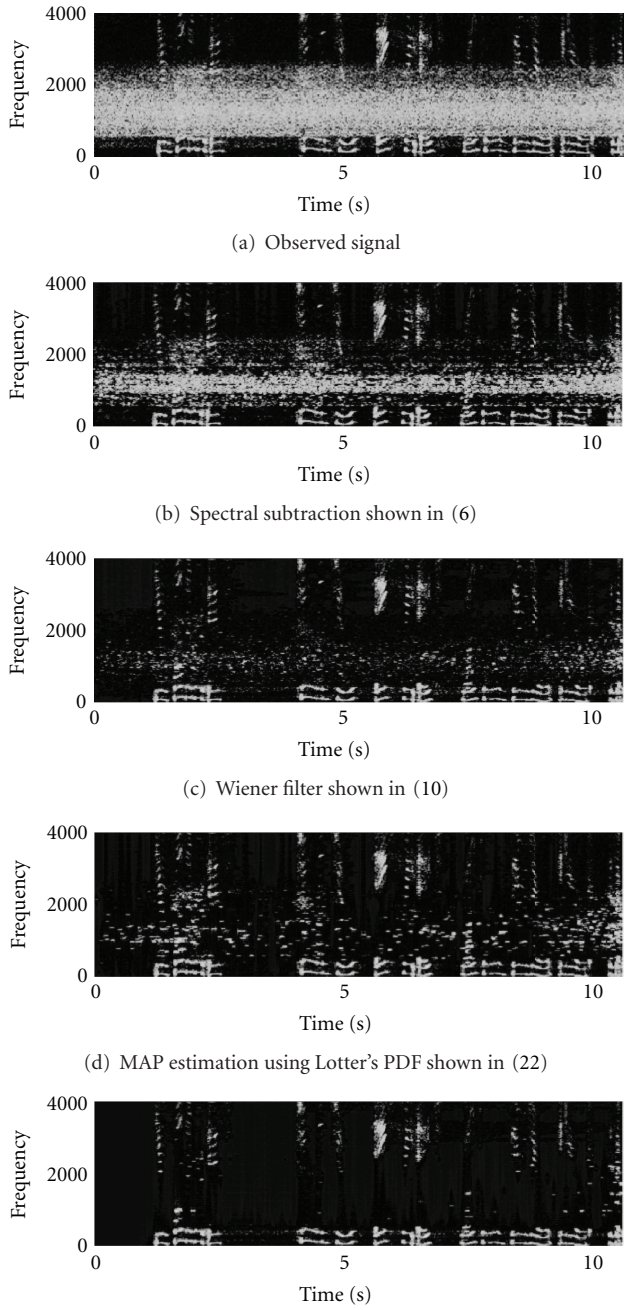
(e) MAP estimation using adaptive PDF shown in (25)

Figure 8: Waveforms of speech enhancement results.

However, in the speech activity segments, we can confirm that the speech spectral components were also vanished. The output speech quality of the adaptive speech PDF method may be improved by adjusting the forgetting factor in the adaptive shape parameters of the speech PDF.

4. Conclusion

Single channel speech enhancement methods have been extensively studied in decades. This paper have presented some spectral gain methods among numerous studies. Of course, there exists various noisy situations, and hence we cannot choose the best speech enhancement system among them. We just tried to explicitly denote theoretical backgrounds of the chosen speech enhancement methods. The noise reduction capability of the speech enhancement methods was roughly compared for an arbitrary noisy



(a) Observed signal

(b) Spectral subtraction shown in (6)

(c) Wiener filter shown in (10)

(d) MAP estimation using Lotter's PDF shown in (22)

(e) MAP estimation using adaptive PDF shown in (25)

Figure 9: Waveforms of speech enhancement results.

speech, although the simulation results may slightly change when different noise and speech signals are used. From the obtained simulation results, we confirmed that the MAP estimation methods gave a good noise reduction performance. Particularly, the recently proposed adaptive speech PDF method reduced the noise signal strongly and hence did not produce a musical noise in speech pause segments. In the speech activity segments, we however perceived a small-level musical noise and a degradation of the speech. Such degradation tends to become large as noise increases. Future works in speech enhancement include a development of an effective noise reduction method which can give a good performance for a noisy speech signal with SNR less than 0 dB.

References

[1] M. Muneyasu and A. Taguchi, *Nonlinear Digital Signal Processing*, Asakura Publishing, Tokyo, Japan, 1999.

[2] A. Kawamura, Y. Iiguni, and Y. Itoh, "A noise reduction method based on linear prediction with variable step-size," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 4, pp. 855–861, 2005.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] B. Widrow, J. G. R. Glover Jr., J. M. Mccool et al., "Adaptive noise cancelling: principles and applications," *Proceedings of The IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[6] 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*, vol. 2003, no. 10, pp. 1043–1051, 2003.

[7] 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. 28, no. 2, pp. 137–145, 1980.

[8] B. Chen and P. C. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with laplacian speech modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. I1097–I1100, March 2005.

[9] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[10] S. Gazor and W. Zhang, "Speech enhancement employing laplacian-gaussian mixture," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 896–904, 2005.

[11] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *Eurasip Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[12] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *Speech Communication*, vol. 51, no. 1, pp. 1–14, 2009.

[13] Y. Tsukamoto, A. Kawamura, and Y. Iiguni, "Speech enhancement based on MAP estimation using a variable speech distribution," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E90-A, no. 8, pp. 1587–1593, 2007.

[14] A. Kawamura, W. Thanhikam, and Y. Iiguni, "A speech spectral estimator using adaptive speech probability density function," in *Proceedings of the EUSIPCO 2010*, pp. 1549–1552, August 2010.

[15] W. Thanhikam, A. Kawamura, and Y. Iiguni, "Speech enhancement using speech model parameters refined by two-step technique," in *Proceedings of the 2nd APSIPA Annual Summit and Conference*, p. 11, December 2010.

[16] W. Thanhikam, A. Kawamura, and Y. Iiguni, "Speech enhancement based on real-speech PDF in various narrow SNR intervals," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E95-A, no. 3, pp. 623–630, 2012.
[17] S. Furui, *Digital Speech Processing*, Tokai University Press, Tokyo, Japan, 1985.
[18] S. L. Miller and D. G. Childers, *Probability and Random Processes*, Elsevier/Academic Press, 2004.
[19] M. Kato, A. Sugiyama, and M. Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E85-A, no. 7, pp. 1710–1718, 2002.