An adaptive MAP speech spectral amplitude estimator combined with a zero phase noise suppressor

Sayuri Kohmura*, Arata Kawamura*, and Youji Iiguni*
* Osaka University, Japan
E-mail: kohmura@sip.sys.es.osaka-u.ac.jp

***Abstract*—We previously proposed an efficient MAP (Maximum *a posteriori*) speech spectral amplitude estimator for stationary noise suppression. Although this method can strongly reduce stationary noise signals, it cannot reduce impulsive noise signals, such that a thunder, clap, and other impact noise signals. On the other hand, we also previously proposed a zero phase noise suppression method to achieve impulsive noise reduction, where its effectiveness was confirmed through some simulations. In this paper, we combine these two effective noise reduction methods and achieve a noise suppressor which can remove both of stationary and impulsive noise signals. We evaluate its noise reduction capability for some types of noise. The simulation results show the effectiveness of the proposed noise suppression method.**

I. Introduction

Continuous improvement of multimedia and communication systems has led to the widespread use of speech recording and processing devices, e.g., mobile phones, mobile video cameras, speech recognition tools, and so on. In practical situations, these devices are being used in environments where undesirable background noise exists. Speech with background noise can cause problems for both mobile communication and speech recognition systems. Noise suppression techniques are required to extract a speech signal from an observed signal which includes noise.

Single channel noise suppression is an important tool to improve the quality of speech communication systems. We focus on single microphone noise suppression systems. A number of single channel noise suppression methods have been proposed and extensively studied for decades [1] - [18]. These methods can be classified into two groups. One is for suppressing stationary noise and the other is for suppressing non-stationary noise. Although there exists a variety of non-stationary noise, we focus on its extreme case that is impulsive noise (clap, thunder, other impact noises, etc.). To suppress stationary and impulsive noise signals, noise suppressors have been individually researched.

An efficient stationary noise suppression method that employs a joint MAP (maximum *a posteriori*) estimator has been proposed by Lotter and Vary [3]. In the literature [3], the speech PDF (probability density function) is modeled 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 interval. However, the residual noise still be persistently perceived. Andrianakis and White [4] were aware that the speech PDF may change in some SNR intervals. They utilized three histograms made from speech signals in three different narrow SNR intervals and approximate them with Gamma density function. As reported in [4], changing these three speech PDFs according to SNR can improve the noise reduction capability. On the other hand, an adaptive PDF method has been proposed by Tsukamoto et al. [5], based on the assumption that speech PDF continuously changes its shape according to SNR. They employed the parametric super Gaussian function used in [3] and adaptively changes its shape parameters according to SNR. The adaptive speech PDF improved the noise suppression capability as reported in [5]. A sophisticated version of the adaptive PDF method has been proposed by Thanhikam et al., and its effectiveness has also been confirmed in [18].

The above mentioned stationary noise suppression methods require a priori estimated noise spectral amplitude. Thus, they cannot reduce non-stationary noise such as an impulsive noise since its pre-information is not available. The median filter [13] is often used to remove an impulsive noise, while it had been originally established in an image processing area. The median filter outputs the median value of an observed signal in a short interval. As a result, it achieves to remove impulsive noise signals and keep edge components of the observed signal, simultaneously. On the other hand, we have previously proposed a more efficient impulsive noise suppression method based on ZP (zero phase) signal, the ZP signal is defined as the IDFT (Inverse Discrete Fourier Transform) of the spectral amplitude [19]. The ZP signal has values only at around the origin when the spectral amplitude is almost flat, and the ZP signal has periodicity when the spectral amplitude has values only at equally spaced frequencies. We assume that an impulsive noise spectral amplitude is approximately flat, and a speech signal is periodic in a short observation. Then, we can reduce the impulsive noise by replacing the noisy ZP signal around the origin with the ZP signal in the second or latter period. After this replacement, taking the DFT of the ZP signal gives the estimated speech spectral amplitude. The IDFT of the estimated speech spectral amplitude with the observed spectral phase provides the estimated speech signal in time domain. Simulation results showed that the ZP signal

replacement method is effective to reduce impulsive noise signals [20].

In this paper, we combine the adaptive PDF noise suppressor proposed in [18] with the ZP noise suppressor proposed in [20] for suppressing real-world noise which includes both of stationary and impulsive noise signals. Since the stationary noise is comparatively easily reduced, we firstly reduce the stationary noise and then suppress the impulsive noise. To evaluate the capability of the combined noise suppressor, we performed noise suppression simulations for some real-world noise signals. Simulation results showed that the combined noise suppressor improved the noise suppression capability in comparison to conventional methods.

This paper is organized as follows. In Section 2, we describe the adaptive PDF method for stationary noise reduction. Section 3 presents the impulsive noise suppression method based on the ZP signal. After that, we combine the stationary and impulsive noise suppressors in Section 4. Then, we show experimental results to confirm the effectiveness of the proposed method. In Section 5, we conclude this research.

II. Stationary noise suppression using adaptive speech probability density function

A general single channel stationary noise suppression system is shown in Fig. 1, where $s(n)$ and $d(n)$ are a clean speech and an additional noise at time n, respectively. In a typical situation of mobile phone communication, the observed signal in time domain $x(n)$ is composed of $s(n)$ and $d(n)$ as

$$x(n) = s(n) + d(n). \tag{1}$$

The noisy signal $x(n)$ is transformed into frequency domain by segmentation and windowing with a window function $h(n)$, e.g., Hanning window. The DFT coefficient of the noisy signal at frame l and frequency bin k is calculated with

$$X_l(k) = \sum_{n=0}^{L-1} x(lQ+n)h(n)e^{-j2\pi nk/L}, \tag{2}$$

where L denotes the DFT frame size. For the computation of the next DFT, the window is shifted by Q samples. The DFT coefficient $X_l(k)$ also consists of speech and noise parts, as given by

$$X_l(k) = S_l(k) + D_l(k), \tag{3}$$

where $S_l(k)$ and $D_l(k)$ represent the DFT coefficients obtained from $s(n)$ and $d(n)$, respectively. The noise suppressor first estimates the noise variance $\lambda_l(k) = E[|D_l(k)|^2]$, where $E[\cdot]$ is an expectation operator. Next, the noise suppressor calculates the *a priori* SNR $\xi_l(k)$ and the *a posteriori SNR* $\gamma_l(k)$ for each DFT bin k, which are defined as

$$\xi_l(k) = \frac{E[|X_l(k)|^2]}{\lambda_l(k)}\ ,\ \gamma_l(k) = \frac{|X_l(k)|^2}{\lambda_l(k)}. \tag{4}$$

Many noise suppressors utilize these two SNRs to calculate the speech spectral gain $G_l(k)$, where various definitions of

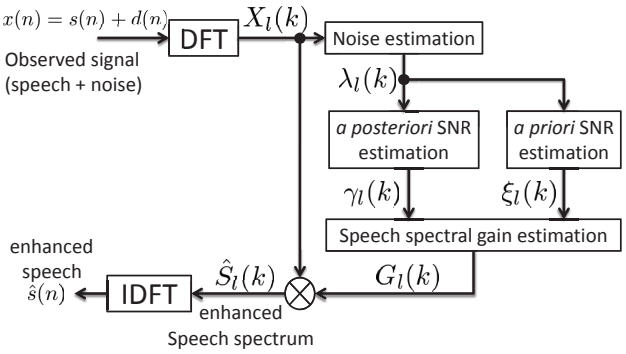


Fig. 1. Overview of single-channel noise suppression system.

$G_l(k)$ have been proposed in previous studies e.g., [1]– [18]. The enhanced speech spectrum $\hat{S}_l(k)$ is given by

$$\hat{S}_l(k) = G_l(k)X_l(k). \tag{5}$$

Finally, we obtain the enhanced speech $\hat{s}(n)$ after the IDFT of $\hat{S}_l(k)$ and overlap-add. The speech spectral gain $G_l(k)$ is the key to effective suppress the noise.

Here, we show an efficient spectral gain $G_l(k)$ used in [3], [5]. This gain function is also utilized in our method. For simplicity, the frame index l and frequency index k are omitted. Let $p(|S|)$ and $p(\angle S)$ represent the PDFs of the speech spectral amplitude and phase, respectively. Lotter and Vary introduced [3]

$$p(|S|) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{|S|^\nu}{\sigma_S^{\nu+1}}\exp\left\{-\mu\frac{|S|}{\sigma_S}\right\}, \tag{6}$$

$$p(\angle S) = \frac{1}{2\pi}, \tag{7}$$

respectively. Here, $\Gamma(\cdot)$ and σ_S^2 denote the Gamma function and the variance of the speech spectrum, respectively. The parameters μ and ν are positive scalars to determine the shape of the speech PDF. Assuming that $p(|S|)$ and $p(\angle S)$ are independent, the MAP solution provides

$$G = u + \sqrt{u^2 + \frac{\nu}{2\gamma}}, \tag{8}$$

$$u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\hat{\xi}}}. \tag{9}$$

In the literature [3], fixed shape parameters μ and ν had been derived from a large amount of speech data in a single narrow SNR interval. On the other hand, Tsukamoto [5] changes two shape parameters according to the SNR to improve the capability of noise reduction. However, these variable shape parameters are determined from the speech data may be incorrectly estimated in many frames, because the shape parameters are determined from only two speech histograms in extremely high and low SNR intervals respectively.

We have previously sophisticated the noise suppressor proposed by Tsukamoto et al. We derived appropriate shape parameters for the speech PDF in (6) [18]. We have used the real

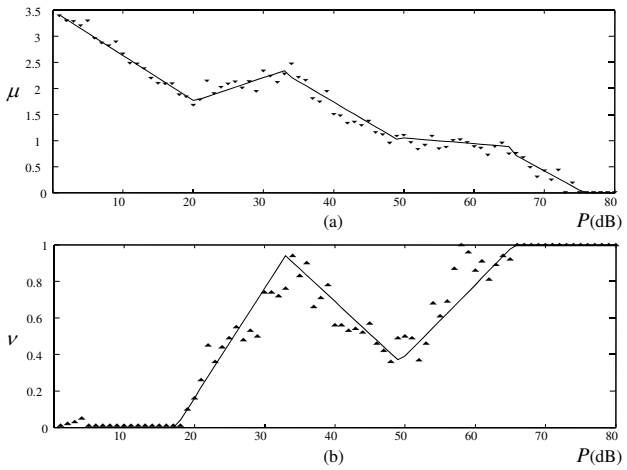


Fig. 2. Relation between shape parameters and SNR intervals (a) μ for SNR (b) ν for SNR.

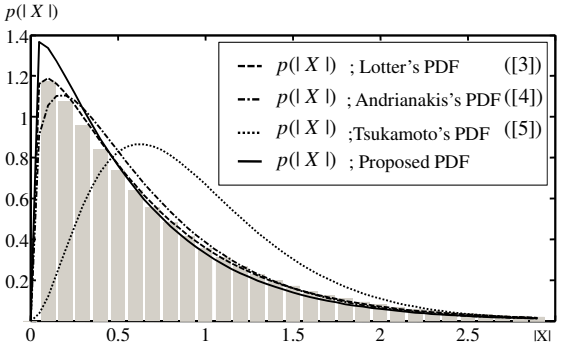


Fig. 3. Speech histogram in 19-20 dB interval and speech PDFs which are Lotter's PDF [3] (dashed line), Andrianakis's PDF [4] (dotted-dash line), Tsukamoto's PDF [5] (dotted line), and proposed PDF (solid line).

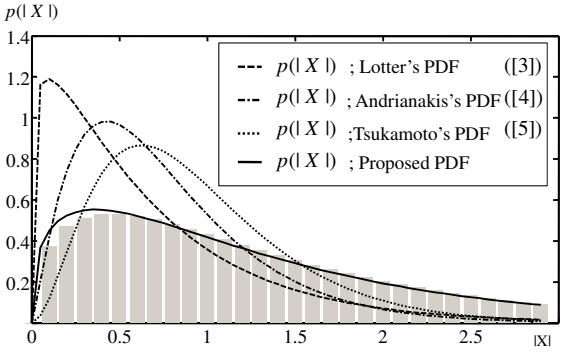


Fig. 4. Speech histogram in 49-50 dB interval and speech PDFs which are Lotter's PDF [3] (dashed line), Andrianakis's PDF [4] (dotted-dash line), Tsukamoto's PDF [5] (dotted line), and proposed PDF (solid line).

speech histograms to derive the appropriate shape parameters. In our previous work, various histograms were created from speech data from low to high SNR intervals. We derived noise suppression algorithm based on data matching between histograms and the speech PDF for each SNR intervals.

Figure 2 (a) and (b) show the obtained optimal value of shape parameters. Here, we can find that an interesting property of the two parameters that include some increases and decreases in value. The fitting results may include fluctuations due to a limited number of the speech data. To reduce the fluctuation of the results, we use averaged values. From Fig. 2, we see a higher linearity by dividing the region into several parts, e.g., 19-33 dB and 33-50 dB. We used the several linear line to represent the results of Fig. 2 (a) and (b). The results are also shown in Fig. 2 (a) and (b) as the solid lines. Table I shows $R_l^{\mu}(k)$ and $R_l^{\nu}(k)$ that represent the derived linear curves, we called them as the shape parameter functions.

Here, we show examples of the speech histogram and the speech PDFs. Figure 3 depicts the histogram of speech amplitude, which is obtained from the 19-20 dB SNR intervals. Figure 3 also shows the conventional speech PDFs [3]– [5] and the proposed speech PDF with the derived shape parameter functions, respectively. The conventional [3], [4], and the proposed speech PDFs give good fitting results, while the speech PDF from [5] is different from other methods in this SNR interval. To observe fitting result in another range, we show the speech histogram and the speech PDFs in 49-50 dB interval in Fig. 4. Here, it appears that the proposed speech PDF provides the best fit for speech histogram. These results support the assumption that the speech histogram has various shapes, and the fixed values of shape parameters from the other conventional methods are no longer appropriate.

The following algorithm is the previously proposed stationary noise suppression algorithm based on the adaptive speech PDF with shape parameter functions shown in Table I.

$$G_l(k) = u_l(k) + \sqrt{u_n^2(k) + \frac{\nu_l(k)}{2\gamma_l(k)}}, \quad (10)$$

$$u_l(k) = \frac{1}{2} - \frac{\mu_l(k)}{4\sqrt{\gamma_l(k)\hat{\xi}_l(k)}}, \quad (11)$$

$$\mu_l(k) = \alpha\mu_{n-1}(k) + (1-\alpha)R_l^{\mu}(k), \quad (12)$$

$$\nu_l(k) = \alpha\nu_{n-1}(k) + (1-\alpha)R_l^{\nu}(k), \quad (13)$$

where α is a forgetting factor, and $\mu_l(k)$ and $\nu_l(k)$ are the adaptive shape parameters. The forgetting factor was introduced to average $R_l^{\mu}(k)$ and $R_l^{\nu}(k)$. In our previous work [18], the highest noise reduction capability was provided when α=0.98.

We show the property of the proposed spectral gain by observing the theoretical gain curve. Fig. 5 (a), (b), (c) and (d) show the gain curves of Lotter's method [3], Andrianakis's method [4], Tsukamoto's method [5] and the proposed method [18], respectively. Here, the spectral gains are depicted for *a priori* SNR ξ and the instantaneous value of ξ, $\gamma - 1$. We first focus on the effect of the spectral gains in high *a posteriori* SNR ξ situation. As we can see from Fig. 5 (a) and (b) when SNR ξ is higher than zero and $\gamma - 1$ is lower than

TABLE I
SHAPE PARAMETER FUNCTIONS $R_l^{\mu}(k)$ AND $R_l^{\nu}(k)$.

SNR range [dB]	$R_l^{\mu}(k) = a_0 P_l(k) + b_0$		$R_l^{\nu}(k) = c_0 P_l(k) + d_0$	
	a_0	b_0	c_0	d_0
$P_l(k) \leq 20$	-0.087	3.50	0.060	-1.04
$20 < P_l(k) \leq 33$	0.045	0.84	0.060	-1.04
$33 < P_l(k) \leq 49$	-0.079	4.90	-0.035	2.11
$49 < P_l(k) \leq 65$	-0.011	1.60	0.039	-1.56
$65 < P_l(k)$	-0.074	5.60	0	1.00

zero, the value of gain reaches to 1 steadily. It means that the gain curves of [3] and [4] have less capability to remove existed background noise in high SNR ξ. While in case of [5] and the proposed gain curves, in high *a priori* SNR ξ situation they show a good capability of noise removal when low value of $\gamma - 1$ persisted. Then, we move on to the next observation. When the *a posteriori* SNR ξ is low and $\gamma - 1$ is high, Fig. 5 (c) and (d) becomes relatively small. It means that Tsukamoto's method [5] and the proposed spectral gain perform better reducing of the noise in low SNR situation and non-speech segments.

In this section, we reviewed an efficient stationary noise suppressor with an adaptive speech PDF. An impulsive noise suppressor is reviewed in the next section.

III. IMPULSIVE NOISE SUPPRESSOR BASED ON ZERO PHASE SIGNAL

The impulsive noise suppressor can be derived by using ZP signal. We define the ZP signal of $x(n)$, $x_0(n)$, as

$$x_0(n) = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^{\beta} e^{j\frac{2\pi n}{N}k}, \tag{14}$$

where β is a certain constant, and we omitted the frame index l for simplicity. Obviously, $|X(k)|^{\beta}$ can be reproduced from the DFT of the ZP signal $x_0(n)$ as

$$|X(k)|^{\beta} = \sum_{n=0}^{N-1} x_0(n) e^{-j\frac{2\pi k}{N}n}. \tag{15}$$

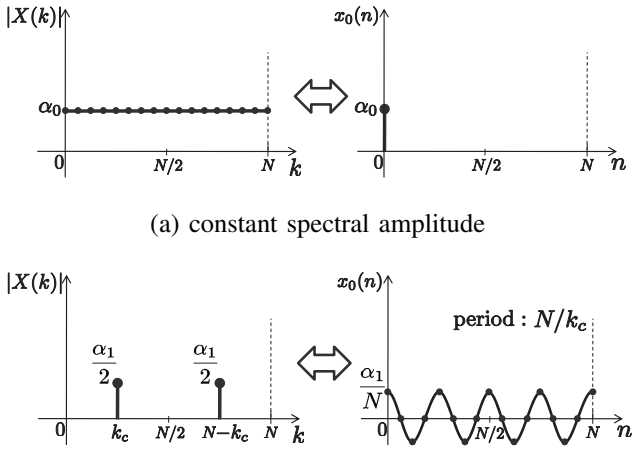
Through this paper, we put $\beta = 1$. In addition, we assume that $x(n)$ is a real valued signal. In this case, the ZP signal $x_0(n)$ come to real even signals [19].

Here, we show a few examples of the ZP signal. Let the spectral amplitude $|X(k)|$ be a constant α_0 (≥ 0). Substituting $|X(k)| = \alpha_0$ into (14) with $\beta = 1$, we have

$$x_0(n) = \alpha_0 \delta(n), \tag{16}$$

where $\delta(n)$ denotes the Kronecker's delta function. Equation (16) implies that the ZP signal of any flat spectral amplitude is expressed as the delta function. Next, let $|X(k)|$ be equally-spaced line-spectral pairs (i.e., $x(n)$ is periodic), where each frequency interval is k_c ($0 < k_c < N/2$). That is

$$|X(k)| = \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{2} \{\delta(k - mk_c) + \delta(k + mk_c - N)\}, \tag{17}$$



(a) constant spectral amplitude

(b) equally spaced line spectra

Fig. 6. Examples of zero phase signal: (a) constant, (b) equally spaced line spectra.

where $\lfloor \cdot \rfloor$ denotes a floor function, and α_m is an amplitude of the m^{th} frequency. Substituting (17) into (14) with $\beta = 1$, we have

$$x_0(n) = \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{N} \cos \frac{2\pi m k_c}{N} n. \tag{18}$$

Hence, the ZP signal of a periodic signal becomes also a periodic signal whose period is N/k_c. These properties are shown in Fig. 6.

As the same manner of conventional noise reduction methods [2]–[17], we also assume that the spectral phase of the estimated speech signal is equal to that of the observed signal, i.e., $\angle \hat{S}(k) = \angle X(k)$. It means that

$$x_0(n) = s_0(n) + d_0(n), \tag{19}$$

where $s_0(n)$ and $d_0(n)$ are the ZP signals of $s(n)$ and $d(n)$, respectively. Firstly, we models a speech signal $s(n)$ in a short observation as a HNM (Harmonic plus Noise Model) [21], [22] given as

$$s(n) = \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \alpha_m \cos(2\pi \frac{k_c}{N} mn + \theta_m) + \varepsilon(n), \tag{20}$$

where k_c/N is the normalized fundamental frequency, and α_m and θ_m are the amplitude and the phase of the m^{th}

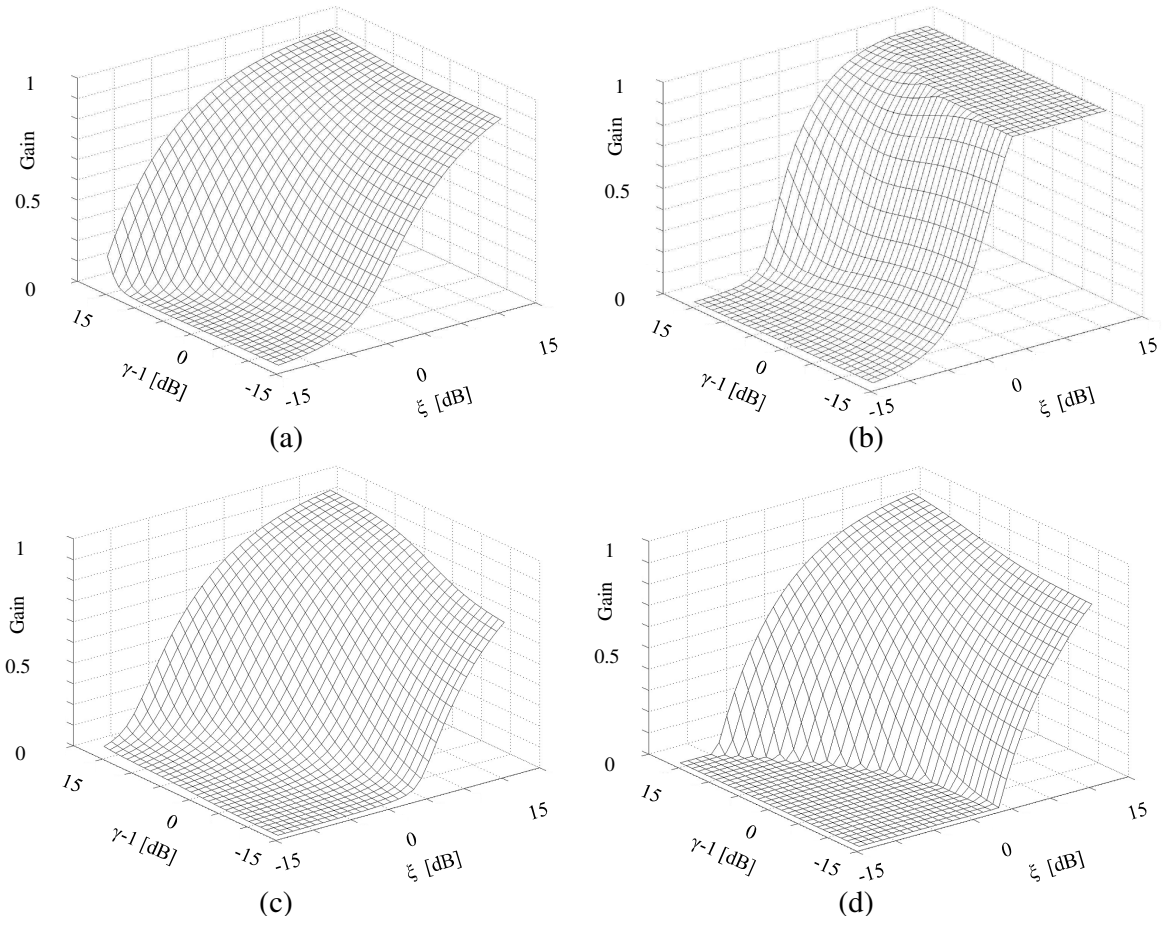


Fig. 5. Gain curves as a function of the *a priori* SNR ξ and instantaneous SNR γ -1. (a) The Lotter's method [3], (b) Andrianakis's method [4], (c) Tsukamoto's method [5] and (d) proposed method.

harmonic frequency, respectively. The signal $\varepsilon(n)$ is a noise signal generated by passing a white noise through an all-pole filter [22]. Here, we assume that the energy of $\varepsilon(n)$ in an observation frame is sufficiently small in comparison to one of the harmonic part. This assumption is appropriate for a voiced speech, but it is not appropriate for an unvoiced speech. Although this assumption may give a degradation to an enhanced speech, the degradation is not fatal. Because, voiced speech energy is usually much greater than unvoiced one.

Next, we show some examples of practical noise and speech signals in the ZP domain. We plotted the ZP signals of some practical wide-band noises and a female speech signal in Fig. 7, where (a) shows a tunnel noise, (b) shows a motor noise, (c) shows a babble noise, (d) shows a clap noise, (e) and (f) show voiced and unvoiced speech signals, respectively. Here, all signals were sampled at 8kHz and $N = 256$. We see from Figs. 7(a)–(d) that the energy of all wide-band noises is concentrated around the origin in the ZP domain. Hence, when we remove the ZP signal around the origin, then the noise is greatly reduced. On the other hand, from Fig. 7(e), we see that the voiced speech becomes a periodic signal with amplitude attenuation in the ZP domain. This attenuation arises due to the window function. Since the window function is known, we can compensate the attenuation. We also see from Fig. 7(e) that the effect of $\varepsilon(n)$ is extremely low for the voiced speech. On the other hand, the ZP signal of the unvoiced speech shown in Fig. 7(f) is similar to that of the noises. As shown in Fig. 7(e) and (f), the energy of the unvoiced speech is less than the voiced one. In this paper, we concentrate on extracting the voiced speech rather than the unvoiced one.

The noise ZP signal has nonzero values mainly around origin. Hence, we assume that the noise ZP signal $d_0(n)$ at $(n > L)$ is sufficiently small for $x_0(n)$. Then we have

$$x_0(n) \approx \begin{cases} s_0(n) + d_0(n), & 0 \le n \le L \\ s_0(n), & L < n \le \frac{N}{2}, \end{cases} \quad (21)$$

$$x_0(n) = x_0(N-n), \quad \frac{N}{2} < n < N. \quad (22)$$

When the pitch period of the speech ZP signal, $T = N/k_c$, is greater than L, we can estimate T as the time index of the second peak of $x_0(n)$ as shown in Fig. 8. Since the observed ZP signal $x_0(n)$ in $T \le n < N+L$ does not include the noise components, we obtain the estimated speech ZP signal $\hat{s}_0(n)$

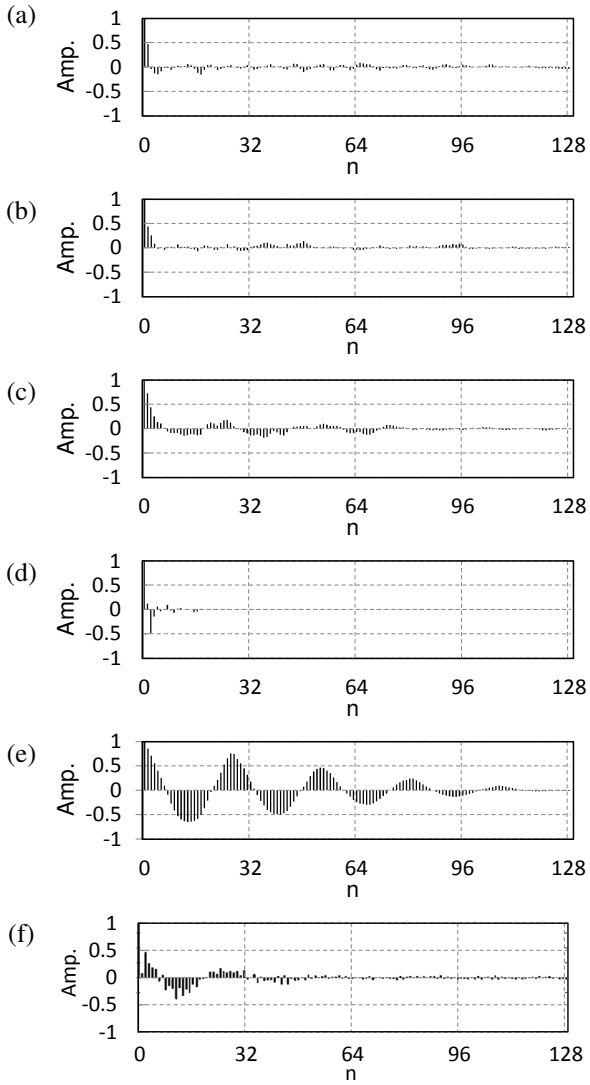


Fig. 7. Zero phase signals. (a) tunnel noise, (b) motor noise, (c) babble noise, (d) clap noise, (e) voiced speech signal, (f) unvoiced speech signal.

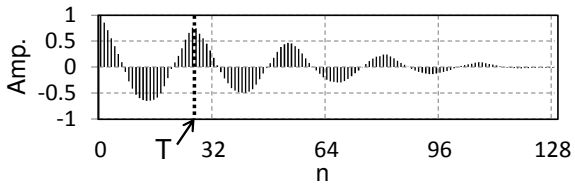


Fig. 8. T obtained from second peak of ZP signal.

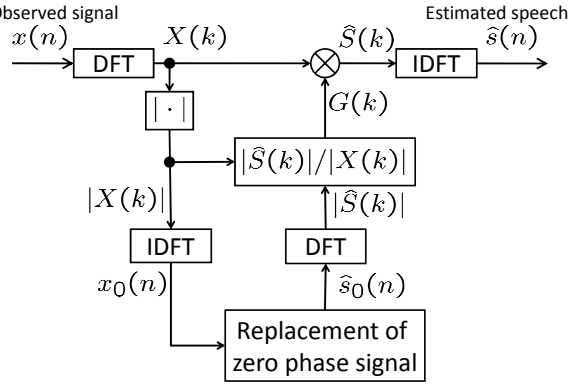


Fig. 9. Proposed wide-band noise reduction system using zero phase signal.

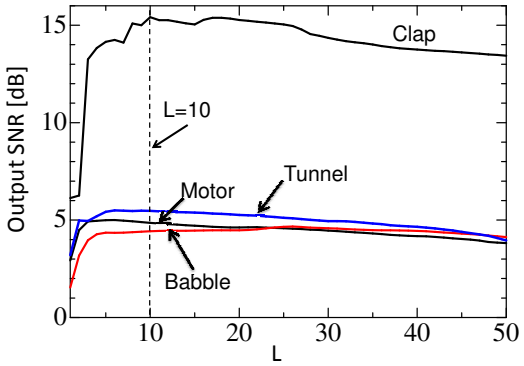


Fig. 10. Practical wide-band noise reduction results for various L with Input SNR= 0dB.

by the following replacement.

$$\hat{s}_0(n) = \begin{cases} \mathrm{sc}(n) \cdot x_0(T+n), & 0 \le n \le L \\ x_0(n), & L < n \le \frac{N}{2} \end{cases}, \qquad (23)$$

where $\mathrm{sc}(n)$ is a scaling function to compensate the envelope attenuation of the speech ZP signal. It is obtained as the reciprocal function of the window for signal segmentation. When we use the hanning window, the scaling function $\mathrm{sc}(n)$ is given as (see Appendix)

$$\mathrm{sc}(n) = \frac{1+\cos\frac{2\pi}{N}n}{1+\cos\frac{2\pi}{N}(n+T)}. \qquad (24)$$

After the replacement (23), the DFT of $\hat{s}_0(n)$ gives the estimated speech spectral amplitude $|\hat{S}(k)|$. Finally, taking the IDFT of $|\hat{S}(k)|e^{j\angle X(k)}$, we have the estimated speech signal $\hat{s}(n)$ in time domain.

Figure 9 shows the block diagram of the impulsive noise reduction system, where the spectral gain is given as $G(k) = |\hat{S}(k)|/|X(k)|$. Here, this system requires the additional DFT and IDFT to achieve impulsive noise reduction without a priori estimation of noise spectral amplitudes. The most important parameters in the impulsive noise suppression method are the pitch period T and the replacement size L shown in (23).

We first describe how to estimate the pitch period T, and then derive an appropriate replacement size L in an empirical manner. From the definition (14), we see that any ZP signal takes the maximum value at the origin. On the other hand, as shown in Fig. 8, a voiced speech provides a periodic ZP signal with amplitude attenuation. Hence, as we stated in the previous section, the index of the second peak in the speech ZP signal gives T. As reported in [24], an averaged pitch period of male

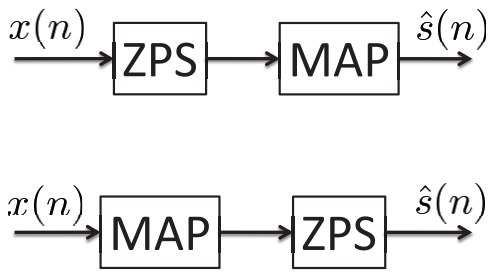


Fig. 11. Combined noise suppressor. There exists possible two combinations.

speakers is about 8ms, and that of female speakers is about 4ms. Hence, an computationally efficient peak search method can be established by restricting the search range. The pitch period T is given as

$$T = \arg\max_{t_L \le n \le t_H} \{x_0(n)\}, \tag{25}$$

where, t_L is the lowest index number of the search range, and t_H is the highest one.

Next, we choose the replacement size L in an empirical manner. For various L, we performed wide-band noise reduction simulations, and evaluated its capability by using

$$\text{InputSNR} = 10\log_{10}\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1} d^2(n)}, \tag{26}$$

$$\text{OutputSNR} = 10\log_{10}\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1} \{\hat{s}(n) - s(n)\}^2}, \tag{27}$$

where M is the number of samples. The results for the four practical noises with Input SNR of 0dB are shown in Fig. 10. We see from this figure that the proposed method is effective for reducing the non-stationary clap noise, and also other wide-band noise signals whose spectral amplitude is approximately flat. Although the respective maximum Output SNRs gave different values of L, all they were less than 10. Hence, we employ $L = 10$ as an appropriate value.

IV. Combined noise suppression system

To simultaneously suppress stationary and impulsive noise signals, we combine the above mentioned two noise suppression methods. We simply cascade these two noise suppressors. There exists two possible combinations for cascading as shown in Fig. 11. Here, "MAP" denotes the stationary noise suppressor explained in Section 2, and "ZPS" denotes the impulsive noise suppressor mentioned in Section 3. We individually evaluate these combinations.

The speech signals used in the simulations were taken from ATR-promotion database [23]. All signals used in simulations were sampled at 8kHz. We put $N = 256$ and $L = 10$, and used the Hanning window for signal segmentation. We put $t_L = 16$ and $t_H = 64$ that implies the pitch search range from 2ms to 8ms. The proposed method was compared with some conventional methods. We carried out noise reduction simulations for 8 kinds of noises which includes practical noise. As the stationary noises, we used a white noise, tunnel noise, motor noise, and babble noise. On the other hand, artificially generated impulse, clap noise, white mixed with impulse, and train noise were used as impulsive noise. Here, the motor and babble noises were obtained from a SPIB database [25], train noise was obtained from a noise database distributed from Sunrise Music inc. [26], and clap and tunnel noises were practically recorded by the authors. The speech signals are spoken by 10 male and 10 female from ATR-promotion database [23]. For evaluating noise reduction capability, we used the Output SNR as a time domain criterion and Itakura-Saito Distance (ISD) [24] as a frequency domain criterion. The ISD is defined as

$$\text{ISD} = \frac{1}{J}\sum_{j=0}^{J-1}\frac{1}{N}\sum_{k=0}^{N-1}(\log\frac{f(k,j)}{g(k,j)} + \frac{f(k,j)}{g(k,j)} - 1), \tag{28}$$

where J is the number of frames, and $f(k,j)$ and $g(k,j)$ are k^{th} bin of spectral envelopes in the j^{th} frame obtained by the maximum likelihood estimation. The spectral envelope $f(k,j)$ is given as [24]

$$f(k,j) = \frac{1}{N}\frac{\sigma_f^2}{1 + 2\sum_{i=1}^{P} A_i \cos(2\pi ki/N)}, \tag{29}$$

$$A_i = \sum_{m=0}^{P-|i|} a_m a_{m+|i|}, \tag{30}$$

where a_m ($m = 1, 2, \ldots, P$) is the m^{th} linear predictor coefficient for the speech signal $s(n)$ in the j^{th} frame. P denotes the order of the linear predictor, and σ_f^2 is the variance of the residual error. The same procedure for the estimated speech $\hat{s}(n)$ gives the other spectral envelope $g(k,j)$. For all of the following simulation results, we compared the proposed method with the spectral subtraction (SS) [1], a variable Maximum a Posteriori estimation method (MAP) [18], and the conventional ZP signal method (ZPS) [20].

Table II shows the output SNR of the stationary noise reduction results. We see from the results for the stationary noise that both of the proposed combined systems can improve the noise reduction capability in comparison to the conventional methods. On the other hand, for non-stationary impulsive noise signals, the noise reduction capability of the proposed method are superior to MAP and ZPS. When the input SNR was 0dB in clap noise situation, the proposed method achieved the SNR of 13.4dB which is higher than the result of the MAP and is equivalent to ZPS. Table III shows the ISD of the simulation results, where it expresses speech spectral envelope distortion. Note that the lower value of ISD is better than the higher one. We see from the results that the proposed method almost improved the ISD in comparison to SS, MAP, ZPS. The difference of the cascade order provided the different results. Although the difference of the results is not so large, the MAP–ZPS order gave slightly better results than the ZPS–MAP.

V. Conclusion

In this paper, we have combined the efficient MAP speech spectral amplitude estimator with the impulsive noise sup-

TABLE II
OUTPUT SNR OF WIDE BAND NOISE REDUCTION RESULTS [DB]

System \ Noise	-10.0 [dB]				0.0 [dB]				10.0 [dB]			
	W	Tn	M	B	W	Tn	M	B	W	Tn	M	B
SS [1]	−2.0	−1.8	−3.2	−4.2	6.9	6.5	5.6	4.7	15.3	14.5	14.1	13.3
MAP [18]	0.5	−0.2	−1.3	−0.4	6.8	3.8	4.5	4.3	13.8	11.7	12.0	11.6
ZPS	−2.4	−3.9	−4.5	−4.9	6.3	5.1	4.5	4.1	12.4	11.8	11.4	11.1
MAP-ZPS	0.9	0.1	−0.8	0.1	6.5	3.6	4.5	4.2	11.8	10.3	10.6	10.1
ZPS-MAP	2.2	0.9	−0.5	−0.3	8.3	6.3	5.9	4.6	12.4	11.3	11.4	10.4

W : white noise Tn : tunnel noise M : motor noise B : babble noise

System \ Noise	-10.0 [dB]				0.0 [dB]				10.0 [dB]			
	I	C	WI	Tr	I	C	WI	Tr	I	C	WI	Tr
SS [1]	−9.9	−9.9	−5.1	−5.2	0.1	0.1	4.5	4.0	10.2	10.1	13.7	13.0
MAP [18]	−9.9	−9.9	0.0	−1.5	0.1	0.1	7.8	4.2	10.1	10.1	15.0	12.6
ZPS	9.0	9.4	−0.5	−4.2	11.7	13.5	7.6	4.7	14.0	14.7	12.9	11.5
MAP-ZPS	9.0	9.4	1.3	−0.7	11.6	13.4	7.9	4.1	13.9	14.5	12.5	10.9
ZPS-MAP	9.0	9.4	3.4	−1.4	11.6	13.4	9.0	5.6	13.9	14.5	12.8	11.5

I : impulsive noise C : clap noise WI : white and impulsive noise Tr : train noise

TABLE III
ISD OF WIDE BAND NOISE REDUCTION RESULTS ($\times 10^4$)

System \ Noise	-10.0 [dB]				0.0 [dB]				10.0 [dB]			
	W	Tn	M	B	W	Tn	M	B	W	Tn	M	B
SS [1]	40.1	36.8	32.8	50.2	4.0	3.7	3.3	5.0	0.4	0.4	0.3	0.5
MAP [18]	37.7	37.0	45.5	45.5	3.8	3.7	4.6	4.6	0.4	0.4	0.5	0.5
ZPS	47.7	66.6	58.5	81.0	4.8	6.6	5.8	8.1	0.5	0.7	0.6	0.8
MAP-ZPS	12.2	14.3	18.1	23.2	1.2	1.4	1.8	2.3	0.1	0.1	0.2	0.2
ZPS-MAP	10.6	12.9	15.7	22.9	1.1	1.3	1.6	2.3	0.1	0.1	0.2	0.2

W : white noise Tn : tunnel noise M : motor noise B : babble noise

System \ Noise	-10.0 [dB]				0.0 [dB]				10.0 [dB]			
	I	C	WI	Tr	I	C	WI	Tr	I	C	WI	Tr
SS [1]	247.2	297.2	81.7	59.4	24.7	29.7	8.2	5.9	2.5	2.9	0.8	0.6
MAP [18]	247.3	297.4	48.8	47.9	24.7	29.7	4.8	4.8	2.4	3.0	0.5	0.5
ZPS [20]	0.0	68.7	26.8	104.7	0.0	6.9	2.7	10.5	0.0	0.7	0.3	1.1
MAP-ZPS	0.0	1.7	10.6	21.2	0.0	0.2	0.1	2.1	0.0	0.2	0.1	0.2
ZPS-MAP	0.0	1.7	6.8	29.4	0.0	0.2	0.7	3.0	0.0	0.0	0.1	0.3

I : impulsive noise C : clap noise WI : white and impulsive noise Tr : train noise

pressor using ZP signal. The MAP estimator and the ZP signal method have been previously established by us for stationary noise suppression and impulsive noise suppression, respectively. The proposed method is very simple. We directly cascaded two noise suppression methods. The simulation results show the effectiveness of the proposed noise suppression method. The results showed that the suppression of stationary noise should be done before impulsive noise suppression.

REFERENCES

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

[2] Y. Ephraim, and D. Malah, "Noise suppression using a minimum mean square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6, pp. 1109–1121, Dec. 1984.

[3] T. Lotter, and P. Vary, "Noise suppression by MAP spectral amplitude estimation using a super-Gaussian speech model," EURASIP Journal on Applied Signal Processing, Vol. 7, pp. 1110–1126, Sept. 2005.

[4] I. Andrianakis, and P.R. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," Speech Communication, Vol. 51, Issue 1, Jan. 2009.

[5] Y. Tsukamoto, A. Kawamura, and Y. Iiguni, "noise suppression based on MAP estimation using a variable speech distribution," IEICE Trans. Fundamentals, Vol.E90-A, No.8, pp.1587-1593, Aug. 2007.

[6] 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.

[7] R. Martin, "Spectral subtraction based on minimum statistics," EURIPCO'94, pp.1182-1185, Sept. 1994.

[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.

[9] S. Kullback, Information Theory and Statistics, Dover Publication, 1968.

[10] P. Vary, and R. Martin, Digital Speech Transmission: Enhancement, Coding and Error Concealment, Wiley, 2005.

[11] Draper, N.R and H. Smith, Applied Regression Analysis, 3rd Ed., John Wiley & Son, New York, 1998.

[12] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., Vol.12, No.3, pp.504-512, Jul. 2001.

[13] M. Muneyasu and A. Taguchi, Nonlinear digital signal processing, Asakura Publishing Company, Tokyo, 1999.

[14] A. Kawamura, Y. Iiguni and Y. Itoh, "A noise reduction method based on linear prediction with variable step-size," IEICE Trans. Fundamentals, Vol.E88-A, No.4, pp.855–861, April 2005.

[15] 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.10, pp.1043–1051, Oct. 2003.

[16] A. Kawamura, W. Thanhikam, and Y. Iiguni, "A speech spectral estimator using adaptive speech probability density function," Proc. of EUSIPCO 2010, pp.1549–1552, Aug. 2010.
[17] W. Thanhikam, A. Kawamura and Y. Iiguni, "Noise suppression using speech model parameters refined by two-step technique" Proc. of the Second APSIPA Annual Summit and Conference, p.11, Dec. 2010.
[18] W. Thanhikam, A. Kawamura, Y. Iiguni, " Speech enhancement based on real-speech PDF in various narrow SNR intervals " IEICE Trans. Fundamentals, Vol.E95-A, No.3, pp.623-630, Mar. 2012.
[19] Y. Kamamori, A. Kawamura, and Y. Iiguni, "Zero phase signal analysis and its application to noise reduction," IEICE Trans. Fundamentals, Vol.J93-A, No.10, pp.658–666, Oct. 2010.
[20] W. Thanhikam, A. Kawamura, Y. Iiguni, "Stationary and non-stationary wide-band noise reduction using zero phase signal," IEICE Trans. Fundamentals, Vol.E95-A, No.5, pp.843-852, May. 2012.
[21] D.W. Griffin and J.S. Lim, "Multiband-excitation vocoder," IEEE trans. Acoust., Speech, Signal Processing, ASSP-36, pp. 1223–1235, Aug. 1988.
[22] J. Laroche, Y. Stylianou, and E. Moulines,"HNS: Speech modification based on a harmonic + noise model" in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing '93, Minneapolis, MN, pp. 550–553, Apr. 1993.
[23] http://www.atr-p.com
[24] S. Furui, Digital speech processing, Tokai University Press, Tokyo, 1985.
[25] http://spib.rice.edu/
[26] http://www.sunrisemusic.co.jp/database/fl/noisedata01_ fl. html
[27] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. 24, No. 5, pp.399–418, 1976.

APPENDIX

The segmented speech signal is given by

$$\tilde{s}(n) \quad = \quad s(n) \cdot h(n). \tag{31}$$

Under the assumption that the power of $|\varepsilon(n)|$ is small enough to be neglected in comparison to one of harmonic part in (20). Then, we can approximate a speech signal $s(n)$ as

$$s(n) \quad \approx \quad \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \alpha_m \cos(2\pi \frac{k_c}{N} mn + \theta_m). \tag{32}$$

We utilize the Hanning window function given as

$$h(n) \quad = \quad \frac{1}{2}\left\{1 - \cos\left(\frac{2\pi n}{N}\right)\right\}. \tag{33}$$

Then, the spectral amplitude of $\tilde{s}(n)$ is given by

$$\begin{aligned} |S(k)| \quad = \quad & \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{2} \left\{ \frac{1}{2}\delta(k - mk_c + 1) + \delta(k + mk_c) \right. \\ & + \frac{1}{2}\delta(k - mk_c - 1) + \frac{1}{2}\delta(k + mk_c - N + 1) \\ & \left. + \delta(k + mk_c - N) + \frac{1}{2}\delta(k + mk_c - N - 1) \right\}. \end{aligned} \tag{34}$$

By substituting (34) into (14) with $\beta = 1$, we get

$$\begin{aligned} s_0(n) \quad = \quad & \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{2} \left\{ \frac{1}{2} \cos \frac{2\pi(mk_c - 1)}{N} n \right. \\ & \left. + \cos \frac{2\pi(mk_c)}{N} n + \frac{1}{2} \cos \frac{2\pi(mk_c + 1)}{N} n \right\} \\ = \quad & \left(1 + \cos \frac{2\pi}{N} n\right) \cdot \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{N} \cos \frac{2\pi mk_c}{N} n. \end{aligned} \tag{35}$$

The scaling function for $s_0(n+T)$ is given as

$$\begin{aligned} sc(n) \quad = \quad & \frac{s_0(n)}{s_0(n+T)} \\ = \quad & \frac{\left(1 + \cos \frac{2\pi}{N} n\right) \cdot \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{N} \cos \frac{2\pi mk_c}{N} n}{\left\{1 + \cos \frac{2\pi}{N}(n+T)\right\} \cdot \sum_{m=1}^{\lfloor \frac{N}{2k_c} \rfloor} \frac{\alpha_m}{N} \cos \frac{2\pi mk_c}{N}(n+T)}. \end{aligned} \tag{36}$$

Using the following relation

$$\cos \frac{2\pi mk_c}{N}(n+T) \quad = \quad \cos \frac{2\pi mk_c}{N} n, \tag{37}$$

we have (24).