Speech Enhancement Using Phase-Dependent A Priori SNR Estimator in Log-Mel Spectral Domain

Yun-Kyung Lee, Jeon Gue Park, Yun Keun Lee, and Oh-Wook Kwon

We propose a novel phase-based method for single-channel speech enhancement to extract and enhance the desired signals in noisy environments by utilizing the phase information. In the method, a phase-dependent a priori signal-to-noise ratio (SNR) is estimated in the log-mel spectral domain to utilize both the magnitude and phase information of input speech signals. The phase-dependent estimator is incorporated into the conventional magnitude-based decision-directed approach that recursively computes the a priori SNR from noisy speech. Additionally, we reduce the performance degradation owing to the one-frame delay of the estimated phase-dependent a priori SNR by using a minimum mean square error (MMSE)-based and maximum a posteriori (MAP)-based estimator. In our speech enhancement experiments, the proposed phase-dependent a priori SNR estimator is shown to improve the output SNR by 2.6 dB for both the MMSE-based and MAP-based estimator cases as compared to a conventional magnitude-based estimator.

Keywords: Phase modeling, speech enhancement, speech separation, decision-directed approach, minimum mean square error estimator.

I. Introduction

Along with the recent development in digital signal processing and multimedia communication technologies, a variety of speech communication services based on speech recognition systems have become popular. In general, although the speech recognition systems show high accuracy in quiet environments, they suffer rapid performance degradation in noisy environments. However, in a realistic speech recognition scenario, speech signals are frequently contaminated by background noisy sources. These noise sources have prevented the widespread use of automatic speech recognition systems in real environments. Automatic speech processing techniques still yield an inferior performance to the human ear in separating target speech signals from other mixed audio signals. A reduction of acoustical background noise or an enhancement of the speech signals is important to enhance the speech quality, reduce the degree of fatigue for speech communication terminals, and improve the speech recognition accuracy of smartphones [1]–[4].

Single-channel speech enhancement technologies enhance the speech signals or reduce noise from the noisy signals captured by a single microphone. In a unified view toward single-microphone speech enhancement systems, the enhancement process depends on the estimation of the spectral gain, which is a function of the *a priori* signal-to-noise ratio (SNR) or *a posteriori* SNR, to enhance the desired signal [1]. A decision-directed (DD) approach is widely used to determine *a priori* SNR from noisy speech signals because it effectively reduces musical noise, which is the residual noise of estimated

Manuscript received Jan. 29, 2014; revised June 3, 2014; accepted June 16, 2014.

This work was supported by the ICT R&D program of MSIP/IITP (10035252, Development of dialog-based spontaneous speech interface technology on mobile platform) and the research grant of the Chungbuk National University in 2011.

Yun-Kyung Lee (yunklee@etri.re.kr), Jeon Gue Park (jgp@etri.re.kr), and Yun Keun Lee (yklee@etri.re.kr) are with the SW·Content Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Oh-Wook Kwon (corresponding author, owkwon@cbnu.ac.kr) is with the School of Electronics Engineering, Chungbuk National University, Cheongju, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.14.2214.0039

frames and is annoying to listeners [3]–[5]. However, this method has a serious drawback in that the estimated *a priori* SNR follows the shape of the *a posteriori* SNR with the delay of a single short time frame [5]. This delay is due to the use of the speech spectrum estimated in the previous frame to compute the current *a priori* SNR. In addition, in a conventional DD approach, only spectral magnitude components are used to compute the *a priori* SNR, and the phase components were disregarded based on the assumption that the phase difference has zero mean. However, the phase components are known to have some speech information and to be useful in human speech perception and automatic speech recognition [6]–[10].

In this paper, we estimate the phase-dependent *a priori* SNR in the log-mel spectral domain by applying a nonlinear transform. In the proposed *a priori* SNR estimator, we do not assume that the phase components have zero mean. After translating the power spectral vector from a noisy speech signal to the log-mel spectral vector, we estimate the phase-dependent *a priori* SNR by utilizing both the magnitude and phase information. The conventional DD approach is also combined with the estimator to recursively obtain the *a priori* SNR. Detailed descriptions of the phase-dependent *a priori* SNR estimator can be found in [11], which was previously published as an article conference proceeding. In addition, we refine the estimated phase-dependent *a priori* SNR using the MMSE-based and MAP-based *a priori* SNR estimator to solve the delay problem while maintaining the advantages of the DD approach. Experimental results show that the proposed estimator improves the output SNR.

The remainder of this paper is organized as follows. Section II describes the signal modeling. Section III describes a conventional DD approach, the proposed phase-dependent *a priori* SNR estimator in the log-mel domain, the phase-based DD approach, and the minimum mean square error (MMSE)-based and the maximum *a posteriori* (MAP)-based two-step *a priori* SNR estimator. Section IV describes the experimental results, and finally, Section V offers some concluding remarks.

II. Signal Modeling

Let $x(t)$ and $n(t)$ represent the original speech signal and a noise from a single microphone, respectively. The mixed speech signal $y(t)$ is simply the sum of these two signals.

$$y(t) = x(t) + n(t). \tag{1}$$

We assume that $x(t)$ and $n(t)$ are uncorrelated with each other. Let X and N represent the spectral magnitude of speech signal and noise, respectively. Denoting the spectral magnitude of the noisy speech signal by Y, the relationship between the noisy speech, clean speech, and noise in the power spectral domain can be shown as follows [7]:

$$Y^2 = X^2 + N^2 + 2\cos(\theta)XN, \tag{2}$$

where θ is the phase vector with a phase difference between X and N.

Typically, the phase term $2\cos(\theta)XN$ is disregarded based on the assumption that it is zero on average.

$$Y^2 = X^2 + N^2\,. \tag{3}$$

However, when (1) is nonlinearly transformed into the log-mel domain by taking the logarithm, the phase term might not be zero on average [7] because the mean of a nonlinearly transformed pdf is not necessarily equal to the transformed mean of the original pdf.

The mel-scale filter is a filter bank whose center frequency is located in the mel-frequency scale, and its bandwidth increases as the center frequency increases. It resembles the human auditory system in that it is more sensitive in the low-frequency bands. Let *mel* represent the index at the mel-scale at f Hz, which is given as

$$mel = 1127\ln\left(f / 700 + 1\right). \tag{4}$$

Figure 1 shows the 23 normalized mel-scale filters used in this work. For each mel-scale filter, a single coefficient is obtained by weighting the power spectrum coefficients within the mel-scale filter with a filter bank matrix.

Let Y_p, X_p, and N_p denote the products of Y^2, X^2, and N^2 with the mel–filter bank matrix W, respectively [12]–[13]. Their relationship in the mel spectral domain becomes

$$Y_p = X_p + N_p + 2\sqrt{X_p N_p}\cos\left(\theta_{X_p} - \theta_{N_p}\right), \tag{5}$$

where θ_{X_p} and θ_{N_p} are the phase spectrum of X_p and N_p, respectively. Since (5) is a quadratic function of $\sqrt{X_p}$, we can obtain the following two solutions:

$$X_p = \left(-c_{X_pN_p}\sqrt{N_p} \pm \sqrt{(c^2_{X_pN_p} - 1)N_p + Y_p}\right)^2, \tag{6}$$

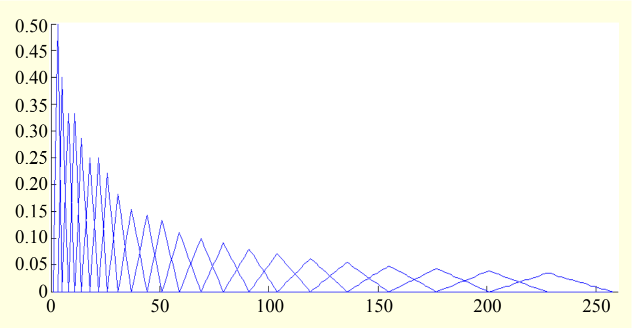

Fig. 1. Normalized mel-scale filters used in this work.

where $c_{X_pN_p}$ is defined as follows:

$$c_{X_pN_p} = \cos(\theta_{X_p} - \theta_{N_p}). \tag{7}$$

III. Phase-Based Speech Enhancement Algorithm

1. DD Approach

The DD approach is a widely used method to determine *a priori* SNR from a noisy signal [4], where the *a priori* SNR is recursively estimated based on the definition of *a priori* SNR and its relationship with the *a posteriori* SNR. The *a posteriori* SNR, which is the parameter for noise suppression, is defined as the ratio of power spectra of a noisy signal and noise. The *a posteriori* SNR at the *m*th frame and *k*th frequency bin, $\gamma(m,k)$, is given by

$$\gamma(m,k) = |Y(m,k)|^2 \Big/ E\left(|N(m,k)|^2\right). \tag{8}$$

The noise power spectrum is estimated during speech pauses by using the weighted noise estimation method [14]. The *a priori* SNR can be defined as

$$\xi(m,k) = E\left(|X(m,k)|^2\right) \Big/ E\left(|N(m,k)|^2\right). \tag{9}$$

The instantaneous SNR can be defined as

$$\begin{aligned} \upsilon(m,k) &= |X(m,k)|^2 \Big/ E\left(|N(m,k)|^2\right) \\ &= \left(|Y(m,k)|^2 - |N(m,k)|^2\right) \Big/ E\left(|N(m,k)|^2\right) \\ &= \left[|Y(m,k)|^2 \Big/ E\left(|N(m,k)|^2\right)\right] - 1. \end{aligned} \tag{10}$$

From the linear combination of the two expressions in (9) and (10), we obtain the new *a priori* SNR as

$$\xi(m,k) = E\left\{\alpha \frac{|X(m,k)|^2}{E\left(|N(m,k)|^2\right)} + [(1-\alpha)\times\upsilon(m,k)]\right\}, \tag{11}$$

with a weighting factor that is constrained to be $0 < \alpha < 1$.

However, as the above expression is hard to implement in practice, approximations were made to determine the new *a priori* SNR recursively.

$$\hat{\xi}(m,k) = \alpha \frac{\left|\hat{X}(m-1,k)\right|^2}{E\left(|N(m-1,k)|^2\right)} + (1-\alpha)\max\left[\gamma(m,k)-1, 0\right], \tag{12}$$

where $\left|\hat{X}(m-1,k)\right|^2$ and $E\left(|N(m-1,k)|^2\right)$ are the speech and noise power spectra estimated in the previous analysis frame, respectively.

2. Estimation of A Priori SNR in the Log-Mel Domain

Most speech enhancement methods generally use only the spectral magnitude by totally disregarding the phase [8]. The spectral phase component holds speech information and is used for human speech perception. There have been a few research activities on utilizing the phase information for speech recognition systems. In this paper, we propose a phase-dependent *a priori* SNR estimator to remove the background noise effectively and improve the performance of speech enhancement algorithms. We transform the power spectral vectors of noisy speech signals into the log-mel spectral vectors to make the non-zero phase term. Then, by estimating the *a priori* SNR in the log-mel domain and enhancing the desired speech signal, we utilize both the magnitude and phase component in the speech enhancement.

The subtractive rule given in (6) is simple to use. However, it is ambiguous regarding the ± sign of (6) in that we have no simple way of knowing which sign to use. To avoid the sign ambiguity in (6), we algebraically derive an alternative subtractive rule by applying the cosine law to the vector diagram [4] shown in Fig. 2.

$$\begin{aligned} X_p &= Y_p + N_p - 2\sqrt{Y_pN_p}\,c_{Y_pN_p}, \\ c_{Y_pN_p} &= \cos(\theta_{Y_p} - \theta_{N_p}). \end{aligned} \tag{13}$$

We derive the phase term $c_{Y_pN_p}$ at the *m*th frame and *k*th frequency bin by making the recursive equation relative to $c_{X_pN_p}$ using the cosine law and spectrum $\hat{X}_p(m-1,k)$ estimated in the previous frame.

$$\begin{aligned} &\sqrt{X_p} = \left(\sqrt{Y_p}\,\hat{c}_{Y_pN_p} - \sqrt{N_p}\right) \Big/ c_{X_pN_p}, \\ &\hat{c}_{Y_pN_p}(m,k) = \left(\frac{\sqrt{X_p(m-1,k)}}{\sqrt{Y_p(m-1,k)}} c_{X_pN_p}(m-1,k) + \frac{\sqrt{N_p(m,k)}}{\sqrt{Y_p(m,k)}}\right). \end{aligned} \tag{14}$$

Let $\lambda_{\log(N_p)}(m,k)$ be the noise power spectrum in the log-mel domain estimated during the speech pauses. The phase-dependent *a priori* SNR is defined as

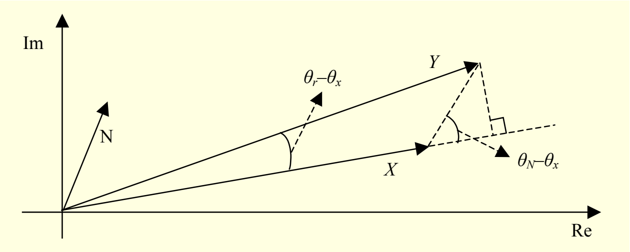

Fig. 2. Diagram illustrating the trigonometric relationship of the clean, noise, and noisy signals.

$$\begin{aligned}&\xi_p(m,k)=E\Big[\log\big(X_p(m,k)\big)\Big]\Big/\lambda_{\log(N_p)}(m,k),\\&\lambda_{\log(N_p)}(m,k)=E\Big[\log\big(N_p(m,k)\big)\Big].\end{aligned}\tag{15}$$

The phase-dependent *a posteriori* SNR is given by

$$\gamma_p(m,k)=E\Big[\log\big(Y_p(m,k)\big)\Big]\Big/\lambda_{\log(N_p)}(m,k).\tag{16}$$

In addition, the phase-dependent instantaneous SNR can be defined as

$$\upsilon_p(m,k)=\frac{E\Big\{\log\Big[Y_p(m,k)+N_p(m,k)-2\hat{c}_{Y_pN_p}\sqrt{Y_p(m,k)N_p(m,k)}\Big]\Big\}}{\lambda_{\log(N_p)}(m,k)}.\tag{17}$$

3. DD Approach in the Log-Mel Domain

To determine the *a priori* SNR from a noisy signal, we use the DD approach, where the *a priori* SNR is computed as a linear combination of (15) and (17) [4].

$$\begin{aligned}&\xi_p(m,k)=E\left[\frac{1}{2}\frac{\log\big(X_p(m,k)\big)}{\lambda_{\log(N_p)}(m,k)}\right]+\frac{1}{2}\bar{\xi}(m,k),\\&\bar{\xi}(m,k)=\frac{\log\Big[Y_p(m,k)+N_p(m,k)-2\hat{c}_{Y_pN_p}\sqrt{Y_p(m,k)N_p(m,k)}\Big]}{\lambda_{\log(N_p)}(m,k)}.\end{aligned}\tag{18}$$

The final estimator is derived by making the preceding equation recursive.

$$\hat{\xi}_p(m,k)=\alpha\frac{\log\Big[\hat{X}_p(m-1,k)\Big]}{\lambda_{\log(N_p)}(m-1,k)}+(1-\alpha)\max\Big[\bar{\xi}(m,k),0\Big],\tag{19}$$

where $0<\alpha<1$ is the weighting factor and $\hat{X}_p(m-1,k)$ is the magnitude estimator obtained in the previous analysis frame. In this paper we chose $\alpha=0.98$.

4. Two-Step A Priori SNR Estimation

In the conventional *a priori* SNR determination system with the DD approach, the estimated *a priori* SNR consequently follows the *a posteriori* SNR with a one-frame delay. This delay is due to the use of the speech spectrum estimated in the previous frame to compute the current *a priori* SNR; therefore, it degrades the speech enhancement performance.

We propose a two-step phase-dependent *a priori* SNR estimator based on the MMSE and the MAP in the log-mel spectral domain to overcome the performance degradation caused by the one-frame delay.

A. MMSE-Based A Priori SNR Estimator

The MMSE estimator for the power spectral density X^2 can be given by the conditional expectation

$$\begin{aligned}\hat{X}^2&=E(X^2\mid Y)\\&=\frac{\int_{-\infty}^{\infty}X^2P(Y\mid X)P(X)dX}{\int_{-\infty}^{\infty}P(Y\mid X)P(X)dX}\end{aligned}\tag{20}$$

and redefined from (20) as [6]

$$\hat{X}^2=\left[\frac{E(|X|^2)}{E(|X|^2)+E(|N|^2)}\right]^2|Y|^2+\frac{E(|X|^2)E(|N|^2)}{E(|X|^2)+E(|N|^2)}.\tag{21}$$

Using (8) and (9) in (21), the MMSE-based *a priori* SNR estimation is given by

$$\begin{aligned}\xi_{\text{MMSE}}&=\hat{X}^2\big/E(|N|^2)\\&=\{\xi/(1+\xi)\}\{1+[\xi/(1+\xi)]\gamma\}.\end{aligned}\tag{22}$$

The first step of the phase-dependent *a priori* SNR estimation is the DD approach in the log-mel domain, whereas the second step, (22), is used to refine the estimated *a priori* SNR of the DD approach. Thus, the refined phase-dependent *a priori* SNR using the MMSE estimation is given by

$$\begin{aligned}&\tilde{\xi}_{\text{MMSE}}=\Big\{\hat{\xi}_p\big/(1+\hat{\xi}_p)\Big\}\Big\{1+\Big[\hat{\xi}_p\big/(1+\hat{\xi}_p)\Big]\gamma_p\Big\},\\&\gamma_p=|Y_p|^2\big/E\big(|N_p|^2\big).\end{aligned}\tag{23}$$

B. MAP-Based A Priori SNR Estimator

The MAP-based estimator for the speech amplitude *X* can be given by the conditional expectation with the noisy speech amplitude *Y* as follows:

$$\begin{aligned}\hat{X}&=\arg\max_x p(X\mid Y)\\&=\arg\max_x\frac{p(Y\mid X)p(X)}{p(Y)}.\end{aligned}\tag{24}$$

The Rician pdf $p(Y\mid X)$ is given by

$$p\big(Y\mid X\big)=\frac{2Y}{E(|N|^2)}\mathrm{e}^{\left[-\frac{X^2+Y^2}{E(|N|^2)}\right]}\mathrm{I}_0\left[\frac{2YX}{E(|N|^2)}\right],\tag{25}$$

where I_0 denotes the modified Bessel function with order zero. The pdf of the noisy spectrum *Y* conditioned on the speech amplitude and phase can be written as

$$p(Y\mid X,\theta)=\frac{1}{\pi E(|N|^2)}\mathrm{e}^{\left[-\frac{|Y-X\mathrm{e}^{j\theta}|^2}{E(|N|^2)}\right]}.\tag{26}$$

The MAP-based *a priori* SNR estimation is obtained by maximizing $p(Y\mid X)p(X)$ and is given by

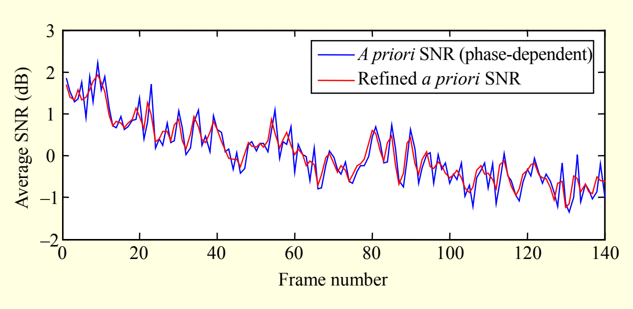

Fig. 3. Variations of the *a priori* SNR before and after refinement of the phase-dependent *a priori* SNR.

$$\begin{aligned}\xi_{\text{MAP}} &= \hat{X}^2 \big/ E(|N|^2) \\ &= [\xi/(1+\xi)][(1/4)+H\gamma].\end{aligned} \tag{27}$$

The refined phase-dependent *a priori* SNR using the MAP estimation is given by

$$\tilde{\xi}_{\text{MAP}} = \left[\hat{\xi}_p \big/ (1+\hat{\xi}_p)\right]\left[(1/4)+H\gamma_p\right], \tag{28}$$

where *H* is the noise suppression gain (de-noising filter).

5. Speech Reconstruction

The multiplicative gain function (de-noising filter) in the DD approach is a function of the *a priori* SNR given in [4].

$$H(m,k) = \hat{\xi}(m,k) \big/ \left[1+\hat{\xi}(m,k)\right]. \tag{29}$$

The enhanced speech spectrum is then obtained as follows:

$$\hat{X}(m,k) = H(m,k)\tilde{Y}(m,k). \tag{30}$$

After translating the enhanced log-mel spectrum into the power spectral domain again, we finally obtain the entire enhanced speech signal by taking the inverse discrete Fourier transform (DFT) and applying the overlap-add method [15]. Figure 3 shows an example of the variations of the *a priori* SNR before and after refinement of the phase-dependent *a priori* SNR estimated using the MMSE-based approach. We can confirm that the proposed two-step phase-dependent *a priori* SNR estimation approach solves the delay problem.

IV. Experimental Results and Discussion

1. Speech Database

A speech database was selected from the Interspeech 2006 Speech Separation Challenge [16], which was drawn from the GRID speech database consisting of six-word sentences with a vocabulary size of 52. The database was recorded by 34 different speakers and was sampled at 25 kHz. For speech enhancement experiments, three types of noise were used: car (N1), babble (N2), and white Gaussian (N3). The speech signal and noise sources were mixed by adding and scaling digitally to create three sets of noisy speech signal recordings. The mixed speech database was created at multiple SNRs: –10 dB, –5 dB, 0 dB, 5 dB, and 10 dB.

For all experiments reported in this paper, the sampling rate of the speech database was reduced from 25 kHz to 16 kHz. The sampling rate of the noise was originally 16 kHz. All speech signals were normalized to have zero mean and unit variance, and were divided into frames of 32 ms in length with an overlap of 16 ms between adjacent frames. For each Hamming-windowed frame, a power spectral (magnitude spectrum) vector of 257 components was derived from a 512-point DFT analysis, and the power spectral vector was then transformed into a log-mel spectral vector. The number of dimensions of the log-mel spectral vector was chosen to be 128.

2. Results of Speech Enhancement

To validate the performance of the speech enhancement from a noisy signal, we compared the waveforms and spectrograms of the original speech signal, the noisy signal, and the enhanced speech signal using the proposed method. The output SNR was computed to evaluate the performance quantitatively. The conventional magnitude-based estimator (baseline) [3] was used as a reference for a performance comparison.

A. Waveform and Spectrogram

Figures 4 and 5 show examples of the waveforms and spectrograms of the original signal, the noisy speech signal, and the enhanced speech signals, where the N3 noise was added to make the noisy signal at a 0 dB SNR. The MMSE-based and MAP-based *a priori* SNR estimators were used to refine the phase-dependent *a priori* SNR.

The DD approach was used in both the proposed method and the conventional DD magnitude-based method (baseline method). Comparing the waveforms and spectrograms, we confirmed that the noise was suppressed remarkably to yield an enhanced speech signal. In addition, in the listening tests using re-synthesized speech, we could hardly hear any background noise.

B. Output SNR

We also calculated the output SNR of the enhanced speech signals, which is defined as the ratio of the power of the original clean speech signal and the power of the error signal between the original signal *x*(*t*) and the enhanced signal. It can be computed as follows:

$$SNR = 10\log_{10}\left[\frac{|X|^2}{\left(|X|-|\hat{X}|\right)^2}\right], \tag{31}$$

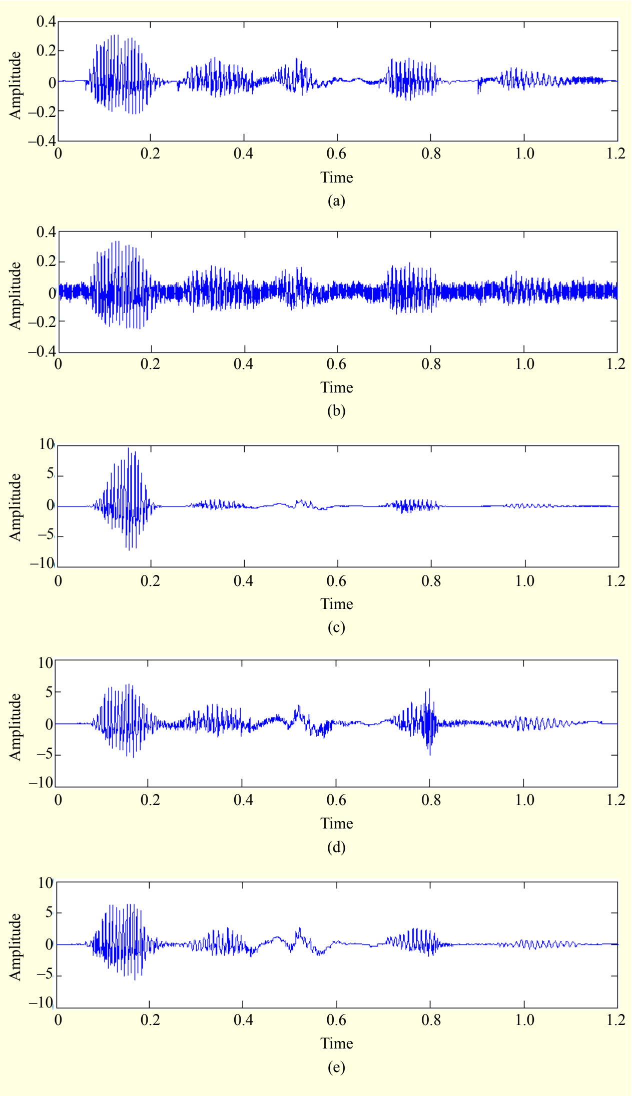

Fig. 4. Speech waveforms when N3 noise is added at 0 dB SNR: (a) original speech signal, (b) noisy signal, (c) enhanced signal using the conventional DD method, (d) enhanced signal using the proposed MMSE-based method, and (e) enhanced signal using the proposed MAP-based method.

where $|X|$ represents the magnitude spectrum of the clean speech signal and $|\hat{X}|$ represents the magnitude spectrum of the reconstructed signal.

Table 1 gives the average SNR with respect to the speech signals under the N1 noise condition using both the conventional DD magnitude-based method and the proposed phase-dependent *a priori* SNR estimator. Tables 2 and 3 show the averages SNR for the N2 and N3 noise conditions, respectively.

When the MMSE-based *a priori* estimator is used to refine the estimated *a priori* SNR, the proposed phase-dependent speech enhancement algorithm effectively improves the output SNR by 3.1 dB, 2.0 dB, and 2.8 dB for the N1, N2, and N3 noise conditions, on average, compared to the conventional DD magnitude-based estimator (baseline method) [3], as shown in Tables 1 through 3. Overall, the proposed method improves the SNR by 2.6 dB on average for the N1, N2, and N3 noise conditions compared with the baseline.

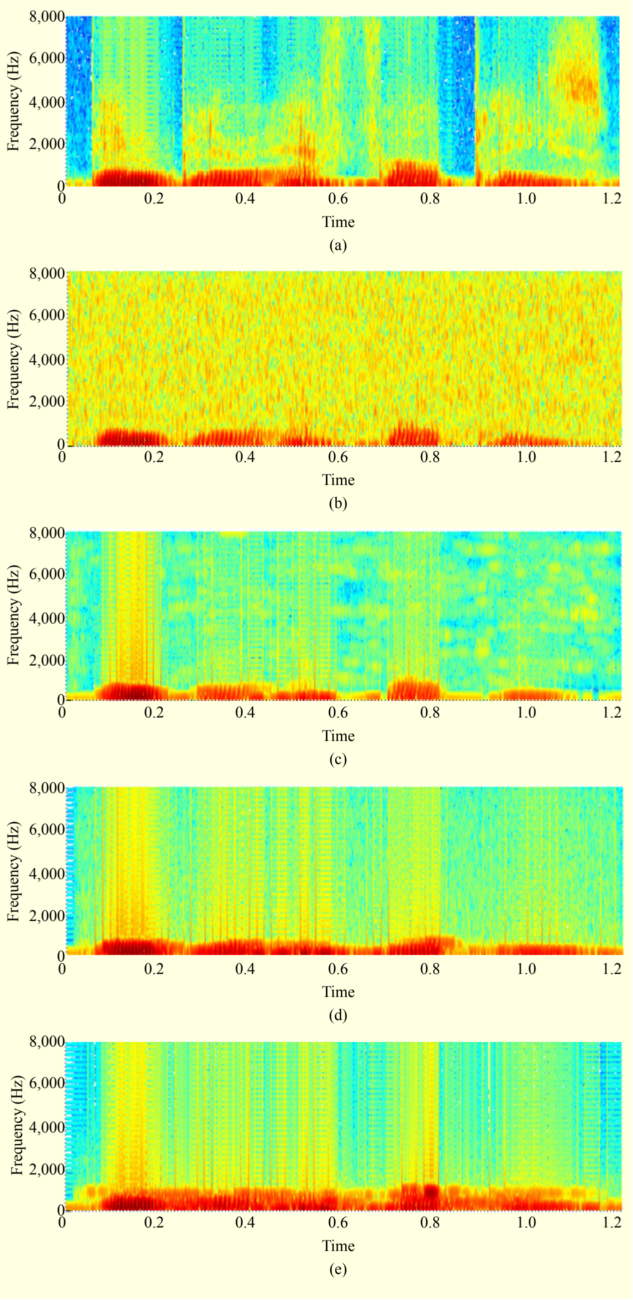

Fig. 5. Spectrograms when N3 noise is added at 0 dB SNR: (a) original speech signal, (b) noisy signal, (c) enhanced signal using the conventional DD method, (d) enhanced signal using the proposed MMSE-based method, and (e) enhanced signal using the proposed MAP-based method.

In the case of using the MAP-based *a priori* SNR estimator, the proposed algorithm improves the output SNR by 3.1 dB,

Table 1. Comparison of output SNRs (dB) for the N1 (car) noise.

Input SNR (dB)	Baseline method	Proposed method		
		Step 1 (before refine)	Step 2 (MMSE)	Step 2 (MAP)
–10	1.5	4.2	4.4	4.3
–5	2.4	5.4	5.7	5.7
0	3.3	7.2	7.5	7.4
5	5.0	7.8	8.1	8.0
10	6.5	8.3	8.5	8.5
Average	3.7	6.6	6.8	6.8

Table 2. Comparison of output SNRs (dB) for the N2 (babble) noise.

Input SNR(dB)	Baseline method	Proposed method		
		Step 1 (before refine)	Step 2 (MMSE)	Step 2 (MAP)
–10	2.2	5.0	5.2	5.3
–5	4.0	6.1	6.3	6.3
0	6.0	7.5	7.9	7.8
5	6.7	7.8	8.1	8.0
10	7.1	8.1	8.4	8.3
Average	5.2	6.9	7.2	7.1

Table 3. Comparison of output SNRs (dB) for the N3 (white Gaussian) noise.

Input SNR(dB)	Baseline method	Proposed method		
		Step 1 (before refine)	Step 2 (MMSE)	Step 2 (MAP)
–10	1.2	3.5	3.9	3.8
–5	1.7	4.1	4.5	4.3
0	2.7	5.3	5.7	5.6
5	4.0	6.8	7.0	7.1
10	5.7	8.0	8.2	8.1
Average	3.1	5.5	5.9	5.8

1.9 dB, and 2.7 dB for the N1, N2, and N3 noise conditions, respectively, on average, compared to the baseline. The MMSE-based *a priori* SNR estimator has slightly better output SNR improvement than the MAP-based estimator, and it could be confirmed that both estimators solve the delay problem.

Overall, the proposed method improves the SNR by 2.6 dB on average for all noise conditions. In all cases, the proposed algorithm was better than the baseline in the output SNR measurements and outperformed stationary noise conditions such as N3. These results show that the proposed method significantly improves the objective quality measures by utilizing the phase information.

Table 4 provides the average SNR and the Perceptual Evaluation of Speech Quality (PESQ) for an enhanced speech using different mel spectrum dimensions of 23, 32, 64, and 128 at 0 dB SNR, where the MMSE-based method is used under the white Gaussian (N3) noise condition. Table 5 shows the average SNR and PESQ for an enhanced speech signal using the MAP-based method to refine the estimated *a priori* SNR. The enhanced speech signal is obtained by extracting the speech feature in the log-mel domain, enhancing the speech, transforming the enhanced log-mel spectrum into the power spectral domain again, and reconstructing the speech signal. When the relatively lower dimensions of the mel spectrum are applied, as in the 23 and 32 cases, the signals are lumped together while being resynthesized, which causes a performance degradation of the proposed speech enhancement algorithm. In this work, we chose 128 mel spectrum dimensions, which improves the performance significantly and outputs the most similar waveform with the original speech signal after speech enhancement.

Figure 6 shows the average PESQ with respect to the speech signals under the N3 noise condition using the conventional DD method and the proposed phase-dependent *a priori* SNR estimators. Overall, the proposed methods improve the PESQ on average. In addition, the proposed method with the MMSE-

Table 4. Average output SNR (dB) and PESQ of enhanced speech signal according to the mel-spectrum dimension (MMSE, N3, 0 dB).

Mel-spectrum dimension	Proposed method	
	Output SNR (dB)	PESQ
23	2.2	1.61
32	4.2	2.53
64	7.7	2.77
128	7.9	2.91

Table 5. Average output SNR (dB) and PESQ of enhanced speech signal according to the mel-spectrum dimension (MAP, N3, 0 dB).

Mel-spectrum dimension	Proposed method	
	Output SNR (dB)	PESQ
23	2.3	1.61
32	4.1	2.51
64	7.6	2.75
128	7.8	2.89

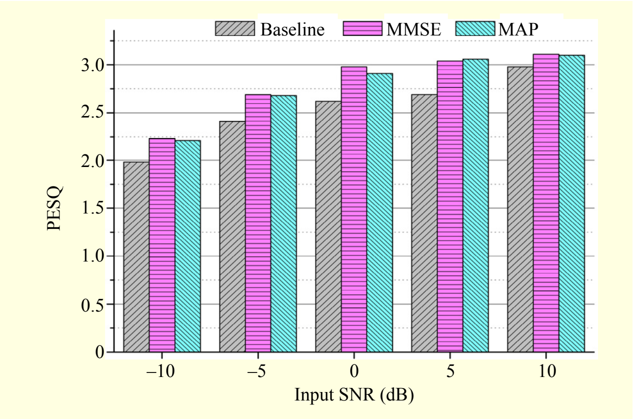

Fig. 6. Average PESQ of enhanced speech signal (N3).

based *a priori* SNR estimator improves the PESQ by 0.3 for –10 dB and –5 dB, 0.4 for 0 dB and 5 dB, and 0.1 for 10 dB, on average, compared with the conventional DD approach.

Using the MAP-based estimator, the proposed method improves the PESQ by 0.2 for –10 dB, 0.3 for –5 dB and 0 dB, 0.4 for 5 dB, and 0.1 for 10 dB, on average, compared with the baseline. As the phase component of the speech signal is relatively more insensitive to human ears than the magnitude component, the speech enhancement performance does not increase as much.

V. Conclusion

We proposed a new single-channel speech enhancement method that estimates the phase-dependent *a priori* SNR in the log-mel spectral domain by considering the magnitude components of the speech signal and the phase components. In the proposed method, the phase term is no longer assumed to be zero because the power spectral vectors of noisy signals are nonlinearly transformed into the log-mel spectral vectors. The new phase-dependent *a priori* SNR is recursively updated by adopting the DD approach. The estimated phase-dependent *a priori* SNR is then refined to solve the delay problem while maintaining the advantages of the decision-directed (DD) approach. In this paper, the MMSE-based and MAP-based *a priori* SNR estimator is used to refine the estimated *a priori* SNR of the DD approach.

By providing the waveforms and spectrograms of the enhanced speech signal, we showed that the enhanced signal was close to the original signal. In the listening tests, we could hardly hear any residual noise from the enhanced signal. In the quantitative evaluation tests, the proposed method with the MMSE-based *a priori* SNR estimator was shown to improve the output SNR by 3.1 dB, 2.0 dB, and 2.8 dB for car, babble, and white Gaussian noise, respectively. In addition, when the MAP-based *a priori* SNR estimator is used, the proposed method improves the output SNR by 3.1 dB, 1.9 dB, and 2.7 dB for the same noise conditions.

The experimental results confirmed that the phase information is useful and can be used together with the magnitude information for improved speech enhancement systems.

References

[1] H.J. Song, Y.K. Lee, and H.S. Kim, “Probabilistic Bilinear Transformation Space-Based Joint Maximum A Posteriori Adaptation,” *ETRI J.*, vol. 34, no. 5, Oct. 2012, pp. 783–786.

[2] S.J. Lee et al., “Intra- and Inter-Frame Features for Automatic Speech Recognition,” *ETRI J.*, vol. 36, no. 3, June 2014, pp. 514–517.

[3] 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. 32, no. 6, Dec. 1984, pp. 1109–1121.

[4] P.C. Loizou, “Part II Algorithms,” in *Speech Enhancement*, CRC Press, 2007, pp. 97–289.

[5] M.J. Alam, D. O’Shaughnessy, and S.-A. Selouani, “Speech Enhancement Based on Novel Two-Step A Priori SNR Estimators,” *Proc. INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 565–568.

[6] D.L. Wang and J.S. Lim, “The Unimportance of Phase in Speech Enhancements,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, Aug. 1982, pp. 679–681.

[7] F. Faubel, J. Mcdonough, and D. Klakow, “A Phase-Averaged Model for the Relationship between Noisy Speech, Clean Speech, and Noise in the Log-Mel Domain,” *Proc. INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 553–556.

[8] L. Deng, J. Droppo, and A. Acero, “Enhancement of Log Mel Power Spectra of Speech Using a Phase–Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, Mar. 2004, pp. 133–143.

[9] K.K. Paliwal, “Usefulness of Phase in Speech Processing,” *Proc. IPSJ Spoken Language Process. Workshop*, Gifu, Japan, 2003, pp. 1–6.

[10] Y.-K. Lee, I.S. Lee, and O.-W. Kwon, “Single-Channel Speech Separation Using Phase-Based Methods,” *IEEE Trans. Consum. Electron.*, vol. 56, no. 4, Nov. 2010, pp. 2453–2459.

[11] Y.-K. Lee and O.-W. Kwon, “A Phase-Dependent A Priori SNR Estimator in the Log-Mel Spectral Domain for Speech Enhancement,” *IEEE Int. Conf. Consum. Electron.*, Las Vegas, NV, USA, Jan. 9–12, 2011, pp. 413–414.

[12] B. Andrassy, D. Vlaj, and C. Beaugeant, “Recognition Performance of the Siemens Front-End with and without Frame Dropping on the Aurora 2 Database,” *Proc. European Conf.*

Speech Commun. Technol., vol. 1, 2001, pp. 193–196.

[13] S. Sigurdsson, K.B. Petersen, and T. Lehn-Schiøle, “Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music,” *Proc. Int. Conf. Music Inf. Retrieval*, Victoria, Canada, Oct. 2006.

[14] M. Kato, A. Sugiyama, and M. Serizawa, “Noise Suppression with High Speech Quality Based on Weighted Noise Estimation and MMSE STSA,” *IEICE Trans. Fundam.*, vol. E85–A, no. 7, July 2002, pp. 1710–1718.

[15] A.V. Oppenheim and R.W. Schaefer, *Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[16] M.P. Cooke et al., “An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition,” *J. Acoust. Soc. America*, vol. 120, no. 5, Nov. 2006, pp. 2421–2424.

Yun-Kyung Lee received her BS degree in electronics engineering and MS degree in control and instrumentation engineering from Chungbuk National University (CBNU), Cheongju, Rep. of Korea, in 2007 and 2009, respectively. She received her PhD degree in control and robot engineering at CBNU, in 2013. Currently, she is in charge of the Spoken Language Processing Research Section, Electronic and Telecommunications Research Institute, Daejeon, Rep. of Korea. Her research interests are speech processing and automatic speech recognition technology.

Jeon Gue Park received his PhD degree in information and communication engineering from Pai Chai University, Daejeon, Rep. of Korea, in 2010. Currently, he is in charge of the Spoken Language Processing Research Section, Electronic and Telecommunications Research Institute, Daejeon, Rep. of Korea. His current research interests include automatic speech recognition technology and dialogue systems.

Yun Keun Lee received his BS and MS degrees in electronic engineering from Seoul National University, Seoul, Rep. of Korea and Korea Advanced Institute of Science and Technology (KAIST), Seoul, Rep. of Korea, in 1986 and 1988, respectively. He received his PhD degree in information and communication engineering from KAIST, in 1998. Currently, he is in charge of the Automatic Translation & Artificial Intelligence Research Center, ETRI, Daejeon, Rep. of Korea.

Oh-Wook Kwon received his BS degree in electronics engineering from Seoul National University, Seoul, Rep. of Korea, in 1986 and his MS and PhD degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 1988 and 1997, respectively. From 1988, he was with the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. In 2000, he joined the Brain Science Research Center, KAIST, as a research professor. From 2001 to 2003, he worked for the Institute for Neural Computation, University of California, San Diego, CA, USA, as a postgraduate researcher. Since 2003, he has been a professor at Chungbuk National University, Cheongju, Rep. of Korea. His research interests include speech recognition; speech and audio signal processing; and pattern recognition.

http://dx.doi.org/10.4218/etrij.14.2214.0039