PAPER

Reducing Speech Noise for Patients with Dysarthria in Noisy Environments

Woo Kyeong SEONG[†], **Ji Hun PARK**[††], ***Nonmembers***, ***and*** **Hong Kook KIM**[†a)], ***Member***

SUMMARY Dysarthric speech results from damage to the central nervous system involving the articulator, which can mainly be characterized by poor articulation due to irregular sub-glottal pressure, loudness bursts, phoneme elongation, and unexpected pauses during utterances. Since dysarthric speakers have physical disabilities due to the impairment of their nervous system, they cannot easily control electronic devices. For this reason, automatic speech recognition (ASR) can be a convenient interface for dysarthric speakers to control electronic devices. However, the performance of dysarthric ASR severely degrades when there is background noise. Thus, in this paper, we propose a noise reduction method that improves the performance of dysarthric ASR. The proposed method selectively applies either a Wiener filtering algorithm or a Kalman filtering algorithm according to the result of voiced or unvoiced classification. Then, the performance of the proposed method is compared to a conventional Wiener filtering method in terms of ASR accuracy.

key words: *dysarthric speech recognition, noise reduction, Wiener filter, Kalman filter*

1. Introduction

Automatic speech recognition (ASR) has been developed as a user interface for electronic devices such as smart phones, home appliances, car navigation systems, and so on [1]. In fact, the performance of modern ASR systems is quite satisfactory, and thus a considerable number of applications have been actively deployed in real-world environments. Furthermore, several attempts have been made in recent years to bring the convenience of ASR to disabled people who have severe constraints to their body movement, allowing only a narrow scope of physical activity [2]. In particular, some of these attempts have been focused on people with dysarthria who have paralysis of the articulator as well as most other parts of the body due to damage to their central nervous system. However, the performance of ASR systems degrades when it is applied to dysarthric speech. This is because dysarthric speech has particular characteristics, such as irregular sub-glottal pressure, loudness bursts, phoneme elongation, unexpected pauses during utterances, and pronunciation variations [3].

Diverse approaches have been proposed to improve the performance of dysarthric ASR systems [4]. However, most such approaches have focused on ASR in a clean environment. Thus, their performance was limited when the ASR system was deployed in real-world applications. To alleviate the effect of background noise on dysarthric ASR, various kinds of noise reduction methods, including spectral subtraction [5], minimum mean square error log-spectral amplitude (MMSE-LSA) [6], and Wiener filtering [7], have been applied to noisy dysarthric speech.

However, since these noise reduction methods have been developed for non-dysarthric speech rather than for dysarthric speech in noisy environments, they do not reflect the various characteristics of dysarthric speech [8]. Specifically, dysarthric speech is often accompanied by the imprecise articulation of consonants rather than vowels [9]. In addition, unvoiced consonants sound very similar to background noise, when compared to voiced consonants or vowels [10]. Thus, noise reduction methods such as Wiener filtering are apt to remove or highly distort unvoiced consonants.

In order to mitigate this problem, we have developed a noise reduction method based on Wiener filtering [11], where noise power for given frame is estimated differently depending on the classification of the frame as either a consonant or a vowel frame. This has been referred to as the consonant/vowel (CV)-dependent Wiener filter [11]. While the CV-dependent Wiener filter achieved better noise reduction performance than a conventional Wiener filter, its performance would be further improved if each consonant frame could be classified as an unvoiced or a voiced consonant frame.

Hence, in this paper, we newly propose a noise reduction method by incorporating the classification of phonemes. In other words, the proposed method first classifies each frame of noisy dysarthric speech to either a voiced or an unvoiced frame by using the pitch strength clustering method [12]. After that, voiced frames are further classified to either vowels or voiced consonants by using a vowel onset time estimated from the linear prediction (LP) residual signals. Even though voiced consonants have acoustic characteristics similar to vowels, Phatak and Allen found that the acoustic characteristics of consonants critically changed by small manipulation, thus humans tended to misrecognize consonants rather than vowels in noisy environments [13]. Based on this finding, we can apply a Wiener filter to the vowel or voiced consonant frames, where a noise power spectrum is estimated differently depending on the conso-

Manuscript received April 28, 2014.
Manuscript revised July 6, 2014.

[†]The authors are with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), 1 Oryong-dong, Buk-gu, Gwangju 500–712, Korea.
[††]The author is with the Visual Display R&D Office, Samsung Electronics, 416 Maetan 3-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do 443–742, Korea.
a) E-mail: hongkook@gist.ac.kr

DOI: 10.1587/transinf.2014EDP7130

nant/vowel classification. In other words, we make a noise power spectrum for a consonant frame underestimated compared to that of a vowel frame in order to reduce spectral distortion by Wiener filtering. By doing this, we can prevent the consonant frames from being distorted by noise reduction techniques. On the other hand, an unvoiced consonant is pronounced by coarticulating the shape of oral cavity, the position of the tongue, and the lip without vocal cord vibration. Thus, an unvoiced consonant can be represented by an autoregressive (AR) model [14] excited with white noise. From this reason, we apply an AR-based Kalman filter [14] to unvoiced consonant frames.

2. Phoneme Classification for Noise Reduction

In this section, we describe phoneme classification for the proposed noise reduction method. First of all, we calculate the pitch strength, $p(l)$, voiced centroid, $c_v(l)$, and unvoiced centroid, $c_{uv}(l)$, for the speech segment starting from the l-th to $(l + N - 1)$-th sample, respectively [12]. In this paper, N is a frame length and it is set to 160 at a sampling rate of 8 kHz. Then, the voiced/unvoiced classification at the l-th sample is performed by computing the ratio defined as

$$R(l) = \frac{p(l) - c_{uv}(l)}{c_v(l) - c_{uv}(l)}. \tag{1}$$

Consequently, the l-th sample is declared as voiced if $R(l) > 0.5$. Otherwise, it is declared as unvoiced. After performing the sample-by-sample decision over all the samples within a frame, the frame is declared as voiced or unvoiced if the number of voiced samples is greater than that of unvoiced samples or vice versa.

Next, for a voiced consonant or vowel frame, CV classification is carried out [11]. To this end, a vowel onset time position is first estimated from LP residual signals. In other words, the first order difference (FOD) of LP residual at the n-th time sample of the m-th frame, $FOD(n; m)$, is defined as

$$FOD(n; m) = E(n; m) - E(n - 1; m), 1 \le n \le N - 1 \tag{2}$$

where $E(n; m)$ is the squared error signal obtained from the p-th order LP analysis. That is, $E(n; m) = (s(n; m) - \sum_{i=1}^{p} \alpha_i(m)s(n - i; m))^2$ where $s(n; m)$ and $\alpha_i(m)$ are the n-th sample of clean voiced speech and the i-th LP coefficient at the m-th frame, respectively. Then, by searching the local maxima of FOD, we determine whether one of them is greater than a pre-defined threshold, which is set to 0.5 in this paper. If there exists a local maximum, which implies that the m-th frame includes a vowel onset time, then this frame is declared as a vowel frame. On one hand, if the m-th frame does not include any vowel onset time but it is included within 10 frame intervals after the previously detected vowel onset frame, it is also declared as a vowel frame. Otherwise, the m-th frame is declared as a voiced consonant frame.

3. Proposed Noise Reduction Method Depending on Phoneme Class

This section proposes a noise reduction method depending on phoneme class. As shown in Fig. 1, we first classify each frame of noisy dysarthric speech to either a voiced or an unvoiced frame, where voiced frames are further divided to vowels or voiced consonants. Then, as shown in Fig. 2, we apply a Wiener filter to the vowel or voiced consonant frames. Notice here that we make a noise power spectrum for a consonant frame underestimated compared to that of a vowel frame in order to reduce spectral distortion by Wiener filtering. Otherwise, we apply a Kalman filter to the unvoiced consonant frames.

3.1 Wiener Filtering for Vowels and Voiced Consonants

Let $x_V(n; m)$ be a noisy voiced frame that can be represented as

$$x_V(n; m) = s_V(n; m) + w(n; m) \tag{3}$$

where $s_V(n; m)$ and $w(n; m)$ are the n-th sample of clean voiced speech and background noise at the m-th frame, respectively. The frequency domain representation of (3) is given by

$$X_V(k; m) = S_V(k; m) + W(k; m) \tag{4}$$

where $X_V(k; m)$, $S_V(k; m)$, and $W(k; m)$ are the k-th spectral

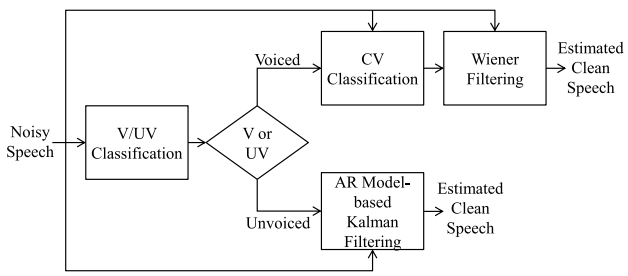

Fig. 1 Block diagram of the proposed phoneme class dependent noise reduction method for dysarthric speech in noise.

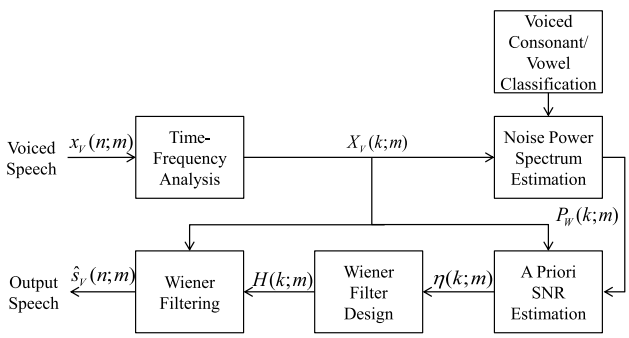

Fig. 2 Block diagram of the Wiener filtering method for vowels and voiced consonants.

component of $x_V(n;m)$, $s_V(n;m)$, and $w(n;m)$, respectively.

Next, in order to construct a Wiener filter, the noise power spectrum, $P_W(k;m)$, is estimated differently according to the CV classification result, such as

$$P_W(k;m) = \begin{cases} P_W(k;m-1)D(\eta(k;m-1)) & \text{for a voiced consonant} \\ P_W(k;m-1) & \text{for a vowel} \end{cases} \quad (5)$$

where $D(\cdot)$ is introduced to control a noise power spectrum of a voiced consonant frame and it is defined as $D(x) = 1/\{1 + \exp(-a(x + b))\}$. It was found from the preliminary experiment that $0.2 \le a \le 0.3$ and $3 \le b \le 8$ in $D(\cdot)$ were proper for improved ASR performance. In particular, $a = 0.25$ and $b = 5$ provided a good compromise between spectral distortion and ASR performance. Moreover, $\eta(k;m-1)$ in (5) indicates the *a priori* signal-to-noise ratio (SNR) of the k-th frequency bin at the $(m-1)$-th frame, which is recursively estimated as [15]

$$\eta(k;m) = \beta(m)\frac{\hat{P}_S(k;m-1)}{P_W(k;m-1)} + (1-\beta(m))T\left[\gamma(k;m)-1\right] \quad (6)$$

where $\hat{P}_S(k;m-1)$ is an estimate of the k-th clean voiced speech power spectral component at the $(m-1)$-th frame. In (6), $\gamma(k;m)$ denotes the *a posteriori* SNR, which is computed by $\gamma(k;m) = P_X(k;m)/P_W(k;m)$, where $P_X(k;m) = |X_V(k;m)|^2$ is the k-th power spectral component of noisy voiced speech at the m-th frame. In addition, $T[x]$ is a half-wave rectifier such that $T[x] = x$ if $x \ge 0$, but $T[x] = 0$ otherwise. The $\beta(m)$ in (6) is a forgetting factor at the m-th frame, defined as

$$\beta(m) = \sqrt{1 - \frac{|E_X(m) - E_X(m-1)|}{\max(E_X(m), E_X(m-1))}} \quad (7)$$

where $E_X(m)$ is the sum of power spectra at the m-th frame over all the frequency bins, K, such that $E_X(m) = \sum_{k=0}^{K-1} P_X(k;m)$.

Consequently, the transfer function of the Wiener filter of the k-th frequency bin at the m-th frame, $H(k;m)$, is estimated as

$$H(k;m) = \frac{\eta(k;m)}{1+\eta(k;m)}. \quad (8)$$

Then, the k-th spectral component of clean voiced speech, $\hat{S}(k;m)$, can be estimated by $\hat{S}(k;m) = H(k;m)X_V(k;m)$, and we obtain an estimate of clean voiced speech, $\hat{s}(n;m)$, by applying an inverse discrete cosine transform to $\hat{S}(k;m)$.

3.2 Kalman Filtering for Unvoiced Consonants

In order to reduce noise for unvoiced consonants, the proposed method employs an AR model based Kalman filter [14] as shown in Fig. 3. Let $x_{UV}(n;m)$ be a noisy unvoiced frame, which is represented in the time and frequency domain as

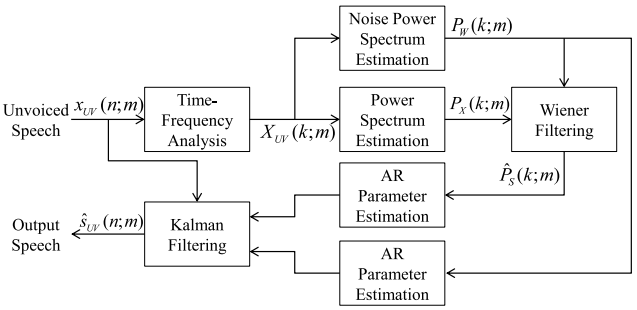

Fig. 3 Block diagram of the Kalman filtering method for noisy unvoiced consonants.

$$x_{UV}(n;m) = s_{UV}(n;m) + w(n;m); \quad (9)$$

$$X_{UV}(k;m) = S_{UV}(k;m) + W(k;m) \quad (10)$$

where $s_{UV}(n;m)$ and $w(n;m)$ are the n-th samples of clean unvoiced speech and background noise at the m-th frame, respectively. In (10), $X_{UV}(k;m)$, $S_{UV}(k;m)$, and $W(k;m)$ are the k-th spectral components of $x_{UV}(n;m)$, $s_{UV}(n;m)$, and $w(n;m)$, respectively.

The power spectra of clean unvoiced speech and background noise are estimated as follows. First, the power spectrum of background noise, $P_W(k;m)$, is estimated by the weighted sum of a noise power spectrum at the previous frame, $P_W(k;m-1)$, and a power spectrum of noisy unvoiced speech at the current frame, $P_X(k;m) = |X_{UV}(k;m)|^2$, by

$$P_W(k;m) = \varepsilon P_W(k;m-1) + (1-\varepsilon)P_X(k;m) \quad (11)$$

where ε is a forgetting factor for noise spectrum estimation and is defined as

$$\varepsilon = \begin{cases} 1 - \frac{1}{m}, & \text{if } m < 100 \\ 0.99, & \text{otherwise} \end{cases}. \quad (12)$$

Second, in order to estimate the power spectrum of clean unvoiced speech, $\hat{P}_S(k;m)$, a Wiener filtering formula is used as

$$\hat{P}_S(k;m) = P_X(k;m)\frac{\omega\eta(k;m)}{1+\eta(k;m)} \quad (13)$$

where $\eta(k;m)$ is the *a priori* SNR defined in (6). In (13), ω is a weighting factor and it is set to 0.1 in this paper for reducing noise effect substantially in AR modeling.

Next, $P_W(k;m)$ in (11) and $\hat{P}_S(k;m)$ in (13) are converted into autocorrelation sequences to extract the AR parameters. In other words, the autocorrelation sequences of power spectra, $\gamma_{\hat{S}}(\tau;m)$ and $\gamma_W(\tau;m)$, can be obtained by using the Wiener-Khintchine theorem [16] as

$$\gamma_{\hat{S}}(\tau,m) = \frac{1}{K}\sum_{k=0}^{K-1}\hat{P}_S(k;m)e^{jk\tau}; \quad (14)$$

$$\gamma_W(\tau,m) = \frac{1}{K}\sum_{k=0}^{K-1}P_W(k;m)e^{jk\tau} \quad (15)$$

where τ is time lag for the autocorrelation. Then, the AR parameters for estimated clean unvoiced speech and noise signals are indirectly estimated by using Burg's method [17]. That is,

$$\gamma_{\hat{S}}(n;m) = \sum_{i=1}^{p} a_i(m)\gamma_{\hat{S}}(n-i;m) + u_1(n;m); \tag{16}$$

$$\gamma_{W}(n;m) = \sum_{i=1}^{q} b_i(m)\gamma_{W}(n-i;m) + u_2(n;m) \tag{17}$$

where $a_i(m)$ and $b_i(m)$ are the i-th AR parameters for estimated clean unvoiced speech and noise signals, respectively, at the m-th frame, and p and q are the orders of the AR models. In this paper, we set $p = 8$ and $q = 8$ by taking into account both computational complexity and model precision. In addition, $u_1(n;m)$ and $u_2(n;m)$ are zero-mean white Gaussian processes.

By using the AR parameters, a Kalman filter is designed for estimated clean unvoiced speech and background noise. From now on, we omit the subscript $_{UV}$ for s_{UV} for the sake of simplicity. First, the Kalman process and the Kalman measurement equations in the state-space domain for speech enhancement are given as

$$\bar{\mathbf{s}}(n;m) = \bar{\mathbf{F}}(m)\bar{\mathbf{s}}(n-1;m) + \bar{\mathbf{G}}\bar{\mathbf{u}}(n;m); \tag{18}$$

$$x(n;m) = \bar{\mathbf{C}}^T\bar{\mathbf{s}}(n;m) \tag{19}$$

where T is the transpose operator, and $\bar{\mathbf{s}}(n;m)$ and $\bar{\mathbf{u}}(n;m)$ are state vectors constructed as

$$\bar{\mathbf{s}}(n;m) = \begin{bmatrix} s(n-p+1;m) & \cdots & s(n-1;m) & s(n;m) \\ n(n-q+1;m) & \cdots & n(n-1;m) & n(n;m) \end{bmatrix}; \tag{20}$$

$$\bar{\mathbf{u}}(n;m) = \begin{bmatrix} u_1(n;m) \\ u_2(n;m) \end{bmatrix}. \tag{21}$$

In (18) and (19), $\bar{\mathbf{F}}(m)$, $\bar{\mathbf{G}}$, and $\bar{\mathbf{C}}$ denote a transition matrix, a system excitation matrix, and a measurement matrix, respectively. Specifically, the transition matrix, $\bar{\mathbf{F}}(m)$, is obtained as

$$\bar{\mathbf{F}}(m) = \begin{bmatrix} \mathbf{F}_s(m) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_n(m) \end{bmatrix} \tag{22}$$

where

$$\mathbf{F}_s(m) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ a_p(m) & a_{p-1}(m) & \cdots & a_1(m) \end{bmatrix}; \tag{23}$$

$$\mathbf{F}_n(m) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ b_q(m) & b_{q-1}(m) & \cdots & b_1(m) \end{bmatrix}; \tag{24}$$

and the system excitation matrix, $\bar{\mathbf{G}}$ is represented by

$$\bar{\mathbf{G}} = \begin{bmatrix} \mathbf{G}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_n \end{bmatrix} \tag{25}$$

where

$$\mathbf{G}_s = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T_{p\times 1}; \tag{26}$$

$$\mathbf{G}_n = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T_{q\times 1}. \tag{27}$$

The measurement matrix, $\bar{\mathbf{C}}$ is given as

$$\bar{\mathbf{C}} = \begin{bmatrix} \mathbf{C}_s \\ \mathbf{C}_n \end{bmatrix} \tag{28}$$

where

$$\mathbf{C}_s = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T_{p\times 1}; \tag{29}$$

$$\mathbf{C}_n = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T_{q\times 1}. \tag{30}$$

Consequently, the estimate of the unvoiced speech signal, $\hat{s}_{UV}(n;m)$, can be obtained by Kalman filtering as

$$\hat{s}_{UV}(n;m) = \mathbf{C}_1^T\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_n;m) \tag{31}$$

where $\mathbf{C}_1 = \begin{bmatrix} \mathbf{C}_s^T & 0 & \cdots & 0 \end{bmatrix}^T_{(p+q)\times 1}$, and $\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_n;m)$ is an estimate of $\bar{\mathbf{s}}(n;m)$ given the noisy speech vectors from the first to the n-th frame, $\mathbf{x}_n$. Note that the detailed explanation of (31) is described in the literature [18].

4. Performance Evaluation

In order to evaluate the performance of the proposed noise reduction method, we applied the proposed method as a preprocessing step of an ASR system. First of all, we trained an ASR system using isolated words of 18,240 utterances of the Korean speech corpus [19]. The acoustic models were based on triphones, where a three-state left-to-right hidden Markov model (HMM) with four Gaussian mixtures was used for each state. For the language model, the lexicon size was 100 words, and a finite state network grammar was employed. As a test database, we used 100 utterances of Korean command words for device control [20]. Each command word was spoken by 31 dysarthric speakers in mild and moderate dysarthric groups, containing 18 and 13 participants, respectively. In this paper, the degree of dysarthria was determined according to the percentage of consonants correct (PCC) index [21] as shown in Table 1. In particular, we chose two degree of dysarthria classification such as mild and moderate. Note here that the acoustic models were trained with non-dysarthric clean speech utterances, while they were tested with noisy dysarthric speech utterances obtained by artificially adding a babble noise and an office noise with SNRs of 10 and 15 dB.

Table 2 compares the average word error rates (WERs) of a baseline ASR system and three ASR systems employing a conventional Wiener filter [7], the CV-dependent Wiener filter [11], and the proposed noise reduction method. As

Table 1 Classification of dysarthric speakers in a testing database according to the percentage of consonants correct (PCC) index.

2 Class	Mild	Moderate		
4 Class	Mild	Mild-to-Moderate	Moderate-to-Severe	Severe
PCC (%)	85-100	65-84.9	50-64.9	-49.9
Male	9	5	4	0
Female	9	4	0	0
Total	18	9	4	0

Table 2 Comparison of average WERs (%) between a baseline ASR system and ASR systems using the conventional Wiener filter, the CV-dependent Wiener filter, and the proposed noise reduction method, where the numbers in parenthesis are WER reductions (%) relative to the baseline ASR system.

Noise Type	Dergee of Dysarthria	Noise Reduciton Method			
		Baseline	Conventional Wiener Filter	CV-Dependent Wiener Filter	Proposed Method
Babble	Mild	62.39	48.06 (22.97)	40.33 (35.36)	38.23 (38.72)
	Moderate	87.93	82.24 (6.47)	77.85 (11.46)	75.58 (14.05)
	Avg.	73.10	62.39 (14.65)	56.06 (23.31)	53.89 (26.28)
Office	Mild	60.08	40.08 (33.29)	34.03 (43.36)	30.42 (49.37)
	Moderate	85.08	80.08 (5.88)	76.85 (9.67)	75.08 (11.75)
	Avg.	70.56	56.85 (19.43)	51.98 (26.33)	49.15 (30.34)
Avg.	Mild	61.24	44.07 (28.04)	37.18 (39.29)	34.33 (43.94)
	Moderate	86.51	81.16 (6.18)	77.35 (10.59)	75.33 (12.92)
	Avg.	71.84	59.62 (17.01)	54.02 (24.81)	51.52 (28.29)

shown in the table, an ASR system using the proposed noise reduction method provided the lowest WERs for both dysarthric groups. In particular, the conventional Wiener filter and the CV-dependent Wiener filter achieved relative WER reductions of 17.01% and 24.81%, respectively, compared to the baseline ASR system. By applying the proposed method, we could further reduce average WER, resulting in a relative WER reduction of 28.29% compared to the baseline ASR system. However, it seemed to be that the proposed method was less effective for moderate dysarthric speech than mild dysarthric speech. This was because unvoiced consonant segments were likely to be classified as voiced segments in case of moderate dysarthric speech due to less accurate consonant articulation.

Figure 4 compares the spectrogram of clean dysarthric speech with those of estimated clean dysarthric speeches by the CV-dependent Wiener filter and by the proposed noise reduction method. Note that the speeches were uttered by a male speaker who was classified as the mild dysarthric class

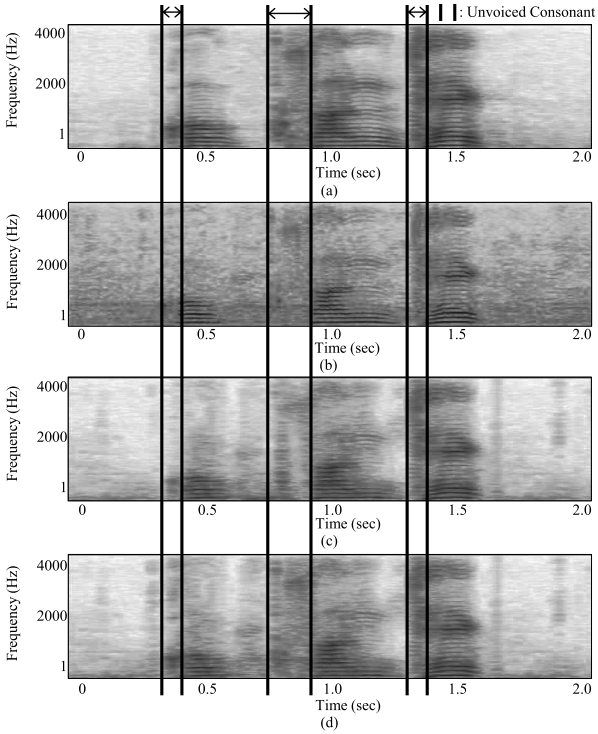

Fig. 4 Spectrograms of (a) clean dysarthric speech, (b) noisy dysarthric speech under a babble noise condition of 15 dB SNR, (c) estimated clean dysarthric speech by the CV-dependent Wiener filter, and (d) that by the proposed noise reduction method.

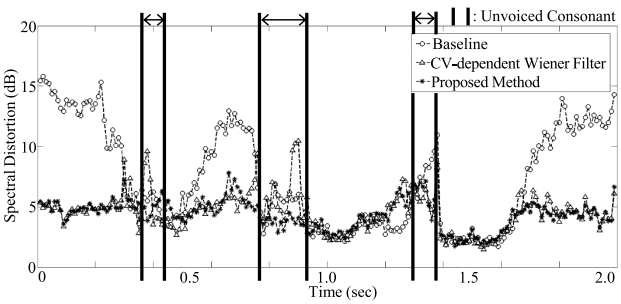

Fig. 5 Comparison of spectral distortions between clean dysarthric speech and noisy speech as well as estimated clean dysarthric speeches by the CV-dependent Wiener filter and the proposed noise reduction method.

and noisy speech was obtained by artificially adding a babble noise of 15 dB SNR as mentioned previously. In addition, the segments marked by the vertical bars represented the unvoiced consonant segments.

Finally, Fig. 5 compares the spectral distortions [22] between clean dysarthric speech, as shown in Fig. 4 (a), and estimated clean dysarthric speeches by the CV-dependent Wiener filter, as shown in Fig. 4 (c), and that by the proposed noise reduction method, as shown in Fig. 4 (d). As a reference, we also compared the spectral distortion between the clean and noisy speeches. Note that the segments marked by the vertical bars were identical to those shown in Fig. 4. It was clearly shown from Fig. 4 that for those unvoiced consonant segments, the proposed noise re-

duction method provided lower spectral distortion than the CV-dependent Wiener filter. Consequently, it could be concluded from Figs. 4 and 5 that the proposed noise reduction method outperformed the CV-dependent Wiener filter, especially on the unvoiced consonant segments.

5. Conclusion

In this paper, we proposed a phoneme class dependent noise reduction method to improve the performance of dysarthric speech recognition in noisy environments. To this end, the proposed method classified each speech frame to either a voiced or an unvoiced frame by using the pitch strength clustering method. After that, a voiced frame is further separated to a voiced consonant or a vowel. Then, we applied a Wiener filter to the voiced frames by estimating the transfer function according to the voiced consonant and vowel classification. Otherwise, we applied an AR model based Kalman filter to the unvoiced consonants. We carried out the performance evaluation of the proposed noise reduction method under simulated babble and office noise conditions for mild and moderate dysarthric speaker groups. As a result, an ASR system with the proposed noise reduction method achieved relative WER reductions of 43.94% and 12.92% for the mild and moderate groups, respectively, compared to a baseline ASR system. However, as mentioned in Sect. 4, it was shown that the proposed method was less effective for moderate dysarthric speech than mild dysarthric speech due to inaccurate consonant articulation. Thus, we are going on extending the proposed method to improve a consonant classification accuracy of moderate dysarthric speech as a future work.

Acknowledgments

This work was supported in part by the NRF grant funded by the government of Korea (MSIP) (No. 2012-010636), and by the MSIP under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1019) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] Y.R. Oh, J.S. Yoon, H.K. Kim, M.B. Kim, and S.R. Kim, "A voiced-driven scene-model recommendation service for portable digital imaging devices," IEEE Trans. Consum. Electron., vol.55, no.4, pp.1739–1747, Nov. 2009.

[2] M.S. Hawley, S.P. Cunningham, P.D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neil, "A voice-input voice-output communication aid for people with severe speech impairment," IEEE Trans. Neural Systems and Rehabilitation Engineering, vol.21, no.1, pp.23–31, Jan. 2013.

[3] M. Hasegawa-Johnson, J. Gunderson, A. Penman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing, pp.1060–1063, Toulouse, France, May 2006.

[4] H. Tolba and A.S. El Torgoman, "Towards the improvement of automatic recognition of dysarthric speech," Proc. 2nd IEEE International Conf. Computer Science and Information Technology, pp.277–281, Beijing, China, Aug. 2009.

[5] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," IEEE Trans. Audio Speech Language Process., vol.20, no.7, pp.2080–2094, Sept. 2012.

[6] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," IEEE Trans. Audio Speech Language Process., vol.21, no.9, pp.1913–1928, Sept. 2013.

[7] A. Ispas, M. Dorpinghaus, G. Ascheid, and T. Zemen, "Characterization of non-stationary channels using mismatched Wiener filtering," IEEE Trans. Signal Process., vol.61, no.2, pp.274–288, Jan. 2013.

[8] W.K. Seong, J.H. Park, and H.K. Kim, "Effects of noise suppression on consonant pronunciation variations of dysarthric speech," Proc. 4th International Symp. Quality of Life Technology, pp.1–2, Incheon, Korea, Oct. 2012.

[9] L.J. Platt, G. Andrews, and P.M. Howei, "Dysarthria of adult cerebral palsy: II. phonemic analysis of articulation errors," J. Speech and Hearing Research, vol.23, no.1, pp.41–55, March 1980.

[10] S. Adams, A. Dykstra, M. Jenkins, and M. Jog, "Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease," J. Medical Speech-Language Pathology, vol.16, no.4, pp.165–172, Dec. 2008.

[11] J.H. Park, W.K. Seong, and H.K. Kim, "Preprocessing of dysarthric speech in noise based on CV-dependent Wiener filtering," Proc. Paralinguistic Information and Its Integration in Spoken Dialogue System Workshop, pp.41–47, Granada, Spain, Sept. 2011.

[12] A. Camacho, "Detection of pitched/unpitched sound using pitch strength clustering," Proc. 9th International Conf. Music Information Retrieval, pp.533–537, Philadelphia, PA, Sept. 2008.

[13] S.A. Phatak and J.B. Allen, "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am., vol.121, no.4, pp.2312–2326, April 2007.

[14] A. Yasmin, Speech Enhancement Using Voice Source Model, Ph. D. Thesis, University of Waterloo, Canada, 1999.

[15] 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, pp.1109–1121, Dec. 1984.

[16] R. Johansson, System Modeling and Identification, Prentice Hall, 1994.

[17] J.P. Burg, A new analysis technique for time series data, Chapter in Modern Spectrum Analysis, D.G. Childers (Eds.), IEEE Press, 1978.

[18] C.H. You, S.N. Koh, and S. Rahardja, "Kalman filtering speech enhancement incorporating masking properties for mobile communication in a car environment," Proc. IEEE International Conf. Multimedia and Expo, pp.1343–1346, Taipei, Taiwan, June 2004.

[19] S. Kim, S. Oh, H.Y. Jung, H.B. Jeong, and J.S. Kim, "Common speech database collection," J. Acoustic Society of Korea, vol.21, no.1, pp.21–24, July 2002.

[20] D.L. Choi, B.W. Kim, Y.J. Lee, and M.H. Chung, "Design and construction of dysarthric speech database," Proc. Korean Society of Speech Sciences, pp.171–172, Iksan, Korea, June 2011.

[21] L.D. Shirberg and J. Kwiatkowski, "Phonological disorders III: A procedure for assessing severity of involvement," J. Speech and Hearing Disorders, vol.47, no.3, pp.256–270, Aug. 1982.

[22] A.C.R. Nandasena, P.C. Nguyen, and M. Akagi, "Spectral stability based event localizing temporal decomposition," Computer Speech and Language, vol.15, no.4, pp.381–401, Oct. 2001.

Woo Kyeong Seong received his B.S. degree in Electronics Engineering from Inha University, Korea in 2010. He is currently pursuing a combined-MS-Ph.D degree at Gwangju Institute of Science and Technology (GIST). His current research interests include pronunciation variation modeling for speech recognition.

Ji Hun Park received his B.S. degree in Electronics Engineering from Kwangwoon University, Korea in 2006. He then received both M.S. and Ph.D degrees in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2008 and 2013. He is currently a senior engineer at Samsung Electronics, Co. Ltd., Korea.

Hong Kook Kim received a B.S. degree in Control and Instrumentation Engineering from Seoul National University, Korea in 1988. He then received both M.S. and Ph.D degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology (SAIT), Kiheung, Korea, from 1990 to 1998. During 1998–2003, he was a senior technical staff member with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, NJ. Since August 2003, he has been with the School of Information and Communications at GIST as a Professor. His current research interests include speech recognition and coding, audio coding and 3D audio, and embedded algorithms and solutions for speech and audio processing for handheld devices.