A Noise Reduction Method Incorporating Consonant and Vowel Characteristics for Dysarthric Speech Recognition

Woo Kyeong Seong, Ji Hun Park, and Hong Kook Kim
School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{wkseong, jh_park, hongkook}@gist.ac.kr

Abstract. In this paper, a noise reduction method is proposed which incorporates consonant and vowel characteristics for dysarthric automatic speech recognition (ASR). Due to the noise-like acoustic characteristics of unvoiced consonants, Wiener filtering approaches may provide more distorted spectra of unvoiced consonants of dysarthric speech than those of voiced consonants. Thus, the proposed method selectively applies a Wiener filter or a Kalman filter depending on the voiced or unvoiced classification of consonants, respectively. In order to demonstrate the effectiveness of the proposed noise reduction method, ASR experiments are carried out on a database of mild and mild-to-moderate dysarthric speeches under different noise conditions. Consequently, it is shown that the proposed noise reduction method achieves relative average word error rate reductions of 20.45% and 8.10% for the mild and mild-to-moderate dysarthric groups, respectively, compared to that using a Wiener filter.

Keywords: Noise reduction, voiced/unvoiced classification, Kalman filter, Wiener filter, autoregressive model

1 Introduction

Dysarthric speech is different from normal speech because of its poor articulation resulting from damage of articulation tissue [1]. Some features of dysarthric speech imply phoneme elongation, irregular intensity of utterance, and pronunciation variations, which degrades the performance of dysarthric automatic speech recognition (ASR) [2]. There have been several research works conducted to improve dysarthric ASR performance [3][4]. However, they have focused on a clean environment that considers noiseless conditions.

In a real-world environment, a background noise is mostly involved, which causes severely degraded performance of an ASR [5]. Furthermore, conventional noise reduction techniques such as Wiener filtering may be inappropriate for dysarthric speech because of the characteristics of dysarthric speech. In other words, dysarthric speech is often accompanied by imprecise articulation of consonants rather than vowels [6]. In addition, unvoiced consonants are much similar to background noise, compared to voiced consonants or vowels. Thus, unvoiced consonants are apt to be removed or highly distorted by the conventional Wiener filtering techniques. In order
to mitigate this problem, there is a need to apply different noise reduction techniques depending on the voiced and unvoiced (V/UV) classification.

Therefore, in this paper, a noise reduction method is proposed by incorporating different characteristics of voiced and unvoiced consonants for the performance improvement of dysarthric ASR. To this end, V/UV classification is first conducted using the pitch strength clustering method [7]. Then, a Kalman filter is applied to unvoiced consonants, while a Wiener filter is applied to voiced consonants and vowels.

2 Proposed Noise Reduction Method

This section proposes a new noise reduction method for dysarthric speech. Fig. 1 shows a block diagram of the proposed noise reduction method. As shown in the figure, V/UV classification is first performed on the noisy dysarthric speech based on the pitch strength clustering method [7]. In particular, each frame of noisy speech is declared as either a voiced or an unvoiced frame by maximizing the distance between the centroids of the pitch strength of each class, where the pitch strength is a spectral similarity between the original signal and a sawtooth waveform constructed by missing non-prime harmonics and the estimated pitch. In other words, if the frame is declared as a voiced frame, then a Wiener filter [8] is applied as

\[
\eta(n,k) = \frac{\eta(n,k)}{1 + \eta(n,k)} | X(n,k) |^2
\]

where \( \hat{S}(n,k) \) and \( X(n,k) \) are the estimated clean and noisy speech spectral components at the \( n \)-th time frame and \( k \)-th frequency bin, respectively. In (1), \( \eta(n,k) \) indicates a priori signal-to-noise ratio (SNR) at the \( n \)-th time frame and \( k \)-th frequency bin. On the other hand, a Kalman filter is designed if the frame is an unvoiced frame. That is, the Kalman filter used in this paper is based on the AR model [9], such as

\[
\hat{s}(n) = C_1^T \hat{s}(n | x_n), \quad C_1 = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^T
\]

where \( \hat{s}(n) \) is the estimated clean speech signal and \( C_1 \) is a measurement matrix where \( p \) and \( q \) are the order of the AR model for clean speech and noise, respec-
A Noise Reduction Method Incorporating Consonant and Vowel Characteristics for Dysarthric Speech Recognition

### Table 1. Comparison of average WERs (%) between an ASR system employing the proposed method and that using the conventional Wiener filter.

<table>
<thead>
<tr>
<th>Dysarthric Group</th>
<th>Noise Reduction Method</th>
<th>Relative WER Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wiener Filtering</td>
<td>Proposed</td>
</tr>
<tr>
<td>Mild</td>
<td>48.06</td>
<td>38.23</td>
</tr>
<tr>
<td>Mild-to-Moderate</td>
<td>82.24</td>
<td>75.58</td>
</tr>
</tbody>
</table>

\[ \tilde{x}(n) \] is an estimate of a state vector, \( \tilde{s}(n) \), given the noisy speech vectors from the first to the \( n \)-th frame, \( x_s \), by following a recursive equation found in [8]. Here, \( \tilde{s}(n) \) is given as

\[
\tilde{s}(n) = \begin{bmatrix} s(n-p+1) \cdots s(n-1) \\ v(n-q+1) \cdots v(n-1) \end{bmatrix} \nonumber
\]

where \( s(n) \) and \( v(n) \) are clean speech and noise, respectively, and are represented as AR models, as follows:

\[
s(n) = \sum_{i=1}^{p} a_i s(n-i) + w(n), \quad (4)
\]

\[
v(n) = \sum_{i=1}^{q} b_i v(n-i) + u(n). \quad (5)
\]

In (4) and (5), both \( w(n) \) and \( u(n) \) are all zero-mean white Gaussian processes. Consequently, a noise-reduced dysarthric speech is finally obtained by applying the combined approach of the Wiener and Kalman filters to a noisy dysarthric speech.

### 3 Performance Evaluation

In this section, the performance of the proposed noise reduction method was evaluated as a means of average word error rate on a dysarthric speech database. To this end, a hidden Markov model based ASR was constructed by using a Korean speech database that was composed of 18,240 isolated words [10]. A three-state left-to-right HMM with four Gaussian mixtures was trained as an acoustic triphone model, and a finite state network grammar was constructed with a lexicon size of 100 words. Next, 100 utterances of Korean command words for device control were prepared as a test database [11], where each command word was spoken by 31 dysarthric speakers in the mild and mild-to-moderate groups. To simulate different noise conditions, babble noise was artificially added to each utterance at 10 and 15 dB SNRs.

Table 1 compares average word error rates (WERs) of an ASR system employing the proposed noise reduction method with that employing the conventional Wiener filter. As shown in the table, relative WER reductions of 20.45% and 8.10% were achieved by the proposed method for the mild and mild-to-moderate groups, respectively, compared to that using the conventional Wiener filter.
4 Conclusion

In this paper, a noise reduction method was proposed by incorporating consonant and vowel characteristics for the performance improvement of dysarthric ASR. In other words, two different noise reduction techniques, such as Wiener and Kalman filters, were applied according to the voiced or unvoiced classification. It was shown from the ASR experiments that an ASR system employing the proposed method achieved average word error rate reductions of 20.45% and 8.10% for the mild and mild-to-moderate groups under 10 and 15 dB SNRs, respectively, compared to that using the conventional Wiener filter.

Acknowledgments. This work was supported in part by the Technology Innovation program (10036461, Development of an embedded key-word spotting speech recognition system individually customized for disabled persons with dysarthria) by the Ministry of Trade, industry & Energy (MI, Korea), and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2012–0006777).

References