SPEECH ENHANCEMENT USING A FREQUENCY-SPECIFIC COMPOSITE WIENER FUNCTION

Fei Chen and Philipos C. Loizou

Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA

ABSTRACT

This paper introduces a new speech enhancement approach based on the design of a frequency-specific composite gain function for Wiener filtering. Motivated by the recently established finding that the acoustic cues at low frequencies can improve speech recognition in noise by the combined electric and acoustic stimulation technique, a less aggressive gain function is applied to replace the Wiener filter’s conventional rigid gain function at low frequencies. With this modification, the proposed approach is able to recover more low frequency (LF) components and enhance the speech quality in noisy environments. An adaptive procedure is utilized to determine the low frequency boundary. Objective evaluation, based on the PESQ measure, revealed that the proposed approach with adaptive LF boundary improved significantly the PESQ scores compared to the scores obtained with the conventional Wiener filter in steady-state noise, white noise and babble interference conditions. Speech quality was also found to be significantly enhanced in the car noise conditions when the less aggressive gain function was selectively applied only to the consonants.

Index Terms – Speech enhancement, Wiener filtering, perceptual evaluation of speech quality (PESQ).

1. INTRODUCTION

In order to improve the quality and intelligibility of speech in noisy environments, speech enhancement techniques have been extensively studied for a long time. Conventional speech enhancement algorithms basically consist of four classes of algorithms, including spectral subtraction [1], subspace [2-3], statistical model based [4] and Wiener filter based algorithms [5]. More details on the above speech enhancement algorithms can be found in [5].

Many of these speech enhancement algorithms are characterized by a common rule of using a product of the noisy speech’s spectrum and a spectral gain function to suppress noise and recover the clean speech spectrum from the noisy background. Studies are now actively ongoing to develop different gain functions [6-9]. However, it has been shown that for the segmental signal-to-noise-ratio (SNR) estimate based noise suppression approach, its performance degrades substantially in the highly non-stationary noise environment. The rigid suppression rule used in Wiener filtering may exacerbate the effects of residual noise and/or musical noise. The suppression of speech components is especially noticeable in the unvoiced segments, where the weak speech components, such as consonants, are removed together along with noise.

Speech enhancement techniques have also been widely applied for cochlear implants (CI) [10-11]. Nowadays, CI is the only medical treatment to restore partial hearing to a severely-to-profoundly deafened person. One challenge for current CI technique is to improve the speech recognition in noisy environments. A recent development of CI technique is the combined electric and acoustic (EAS) stimulation, in which an electrode array is implanted only partially into the cochlea so as to preserve the residual acoustic hearing (20-60 dB HL up to 750 Hz and severe to profound hearing loss at 1000 Hz and above) that many patients still have at the low frequencies. The benefit of EAS in terms of better speech recognition in noise has been well documented [12].

Motivated by the EAS benefit of acoustic cues at the low frequencies for improving speech recognition in noise, the purpose of this paper is to propose a new frequency-specific gain function to recover the low frequency (LF) components and subsequently to enhance speech quality. In this approach, we modify the gain function of Wiener filtering by applying a less aggressive gain function at the low frequencies, so as to prevent the LF components from being over-attenuated.

2. PROPOSED ALGORITHM

2.1. Wiener Filtering

Let the noisy speech signal in time domain be denoted as $y(n) = x(n) + d(n)$, where $x(n)$ and $d(n)$ are clean speech and additive noise, respectively. In order to achieve noise suppression with a highly reduced musical noise effect, the concept of a priori SNR estimate has been introduced into the classical speech enhancement, including Wiener filtering [13]. The gain function $g_k$ in Wiener filtering can be expressed in terms of the a priori SNR $\xi_k$, as

$$g_k = \frac{\xi_k}{\xi_k + 1},$$

(1)

where $k$ denotes the $k$th frequency bin. The a priori SNR $\xi_k$ is estimated using the decision-directed method [4]. At frame $m$, $\xi_k(m)$ is estimated as

$$\xi_k(m) = \alpha \frac{|\hat{X}_k(m-1)|^2}{|D_k(m-1)|^2} + (1-\alpha) \cdot \max \left( \frac{|Y_k(m)|^2}{|D_k(m)|^2}, -1, 0 \right),$$

(2)

where $\alpha$ is a smoothing constant ($\alpha=0.98$), $\hat{X}_k(m-1)$ denotes the enhanced signal spectrum obtained at frame $(m-1)$, and $Y_k(m)$ and $D_k(m)$ denote the noisy speech and estimated noise spectra, respectively. This recursive relationship provides smoothness in the estimate of $\xi_k$, and consequently can eliminate the musical noise [14].
Following the gain function in Eq. (1), the enhanced signal spectrum at frame $m$ is obtained as

$$\hat{X}_k(m) = Y_k(m) \cdot g_k(m),$$

where $g_k(m)$ is the gain function at frame $m$.

### 2.2. Frequency-Specific Composite Gain Function

Fig. 1 shows the Wiener gain function Eq. (1) as a function of the a priori SNR. Apart from this widely-used gain function, several other types of gain functions have also been proposed or modified based on Eq. (1). Amehraye et al. [8] recently proposed a generalized strategy to design the gain function, which summarized several perceptually motivated modification of standard Wiener filtering. In [9], Plapous et al. used the following less aggressive gain function:

$$g_k = \frac{\sqrt{\xi_k}}{\sqrt{\xi_k} + 1}.$$  

The less aggressive gain function, shown in Fig. 1, would recover spectral components, especially those with lower a priori SNRs, without over-attenuation. This study replaces Wiener filter’s conventional rigid gain function at low frequencies with the less aggressive gain function in Eq. (4), and proposes a new frequency-specific composite gain function.

### 2.3. Determination of Low-Frequency Boundary

An adaptive procedure was used to determine the low-frequency boundary below which the less aggressive gain function is applied and above which the rigid gain function at high frequencies is applied. The adaptive LF boundary determination is based on a thresholding procedure using a priori SNRs. The a priori SNR level $\xi_k = 0.48$ dB is empirically selected as the threshold and compared with a priori SNRs of the frequency bins below 3 kHz at each frame. The max frequency bin with a priori SNR larger than the threshold is decided as the LF boundary ($f_{\text{LF bound}}$). If there is no frequency bin with a priori SNR larger than the threshold, 600 Hz is selected as the LF boundary. Therefore, the LF boundary varies below 3 kHz. The above adaptive thresholding procedure can be summarized as follows:

$$\max_{f \in (0,3000) \text{ Hz}} (\xi_k(m)) \geq 0.48 \text{ dB}, \quad f_{\text{LF bound}} = \arg \max_{f \in [0,3000] \text{ Hz}} (\xi_k(m)) \geq 0.48 \text{ dB}, \quad f_{\text{LF bound}} = 600 \text{ Hz}$$

Fig. 2 shows the spectrograms of the enhanced speech by the proposed approach and Wiener filtering. As can be seen in Fig. 2(c), the Wiener filtering aggressively attenuates the spectral components at low frequencies. However, the proposed approach recovers more LF spectral components (e.g. around 1.25s and 2.5s in Fig. 2(d)) by using the less aggressive gain function (Eq. (4)).

### 3. RESULTS

We evaluated the quality of the enhanced speeches by using perceptual evaluation of speech quality (PESQ) measure for Wiener filtering and the proposed frequency-specific composite gain function based Wiener filtering, which we denote as com_Wiener. The PESQ measure, described in ITU-T Recommendation P.862, has been proven to be more reliable and correlated with Mean Opinion Score and speech intelligibility than other traditional objective measures [15].
Fig. 3. PESQ scores obtained using Wiener filtering and the proposed methods when speech was corrupted by: (a) steady-state noise, (b) white noise, (c) babble and (d) car interferences. The ‘x’ sign means that the PESQ score is lower than that from Wiener filtering. The ‘*’ sign indicates that the PESQ score is significantly improved when compared with that from Wiener filtering. The ‘+’ sign above the adaptive LF boundary based variant of com_Wiener means that the PESQ score is significantly improved when compared with that from its paired fixed LF boundary based variant of the com_Wiener estimator.

The implementation of PESQ measure is based on the code included in [5].

3.1. Experimental Conditions

The speech material consisted of phonetically-balanced sentences taken from the IEEE database [16]. All the sentences were produced by a male speaker, and recorded at a 25 kHz sampling rate in a sound-proof booth (Acoustic Systems) in our lab. Four types of maskers were used to corrupt the IEEE sentences. The first was continuous steady-state noise (SSN), which had the same long-term spectrum as the test sentences in the IEEE corpus. The second was white noise. The third and fourth were non-stationary babble and car interferences. The test sentences were corrupted by the stationary SSN and white noise at -10, -5, 0, 5 and 10 dB SNR, and by babble and car interferences at 0, 5, 10 and 15 dB SNR. For the purpose of PESQ computation, the enhanced speech was down-sampled to the sampling rate of 16 kHz. All the programs were implemented in the Matlab environment, and the PESQ computation code and car and babble interference materials are accessible from [5]. A number of noise estimation algorithms have been proposed. This study used a noise estimation algorithm suitable for highly non-stationary noisy environments [17], for estimating the noise spectrum in Eq. (2).

Four variants of the com_Wiener were tested in this study. The first used the adaptive procedure to determine the LF boundary, and the second utilized 600 Hz as the fixed LF boundary. These two variants applied the less aggressive gain function and the rigid gain function at low frequencies and high frequencies, respectively, in both voiced and unvoiced segments. We also examined performance when the less aggressive gain function was applied only to the weak unvoiced consonants. The voiced/unvoiced boundaries for IEEE sentence are available from [5]. The low and upper frequency boundaries for consonants in unvoiced segment were also determined using the adaptive procedure described in the previous Section. The third and fourth variants applied the adaptive and fixed LF boundaries in voiced segments, respectively. We will refer to the above-mentioned four variants of the proposed approach as: 1) com_Wiener(A), 2) com_Wiener(F), 3) com_Wiener_C(A), and 4) com_Wiener_C(F). ‘(A)’ and ‘(F)’ mean the adaptive and fixed LF boundary based variants, respectively. The suffix ‘_C’ denotes the use of the less aggressive gain function to the consonants only.

The paired t-test was used to assess significant differences, at the 0.05 significance level, between the various speech enhancement algorithms. The significance level of 0.05 was selected to accept or reject the null hypothesis that there was no
performance difference between a given pair of enhancement algorithms.

3.2. Experimental Results

The results of PESQ scores are shown in Fig. 3. For comparative purposes, we also show the PESQ scores obtained using the Wiener filtering algorithm (the same gain function is applied to all frequencies). Firstly, as can be seen from Fig. 3 nearly all PESQ scores obtained by the four variants of the com_Wiener estimator improved compared to those obtained using the conventional Wiener estimator, except in one condition with com_Wiener(F) at 10 dB SNR and car interference (marked by ‘x’ in Fig. 3). Secondly, when speech was corrupted by stationary SNR and white noise, nearly all improvements were found to be statistically significant, as indicated by ‘*’ in Fig. 3, except in one condition with com_Wiener(F) at 10 dB. For babble interference, both adaptive LF boundary based variants (com_Wiener(A) and com_Wiener_C(A)) produced significant improvements at all SNR levels. When the less aggressive gain function was applied only to the consonants, the fixed LF boundary based variant (com_Wiener_C(F)) produced significant improvement at all SNR levels. Similarly, for car interference, when the less aggressive gain function was applied only to the consonants, the adaptive and fixed LF boundary based variants (com_Wiener_C(A) and com_Wiener_C(F)) produced significant improvement at SNR levels of 5, 10 and 15 dB.

In brief, the above results indicate that: 1) com_Wiener(A) produced significant PESQ improvements for steady-state noise, white noise and babble interference, and 2) com_Wiener_C(A) and com_Wiener_C(F) yielded statistically significant improvements in PESQ scores for the four maskers tested. Though not shown in Fig. 3, it is noted that all the four variants of com_Wiener produced significantly improved speech quality compared to that of unprocessed (noisy) corrupted sentences.

Fig. 3 also shows the difference in PESQ scores for two pairs of adaptive and fixed LF boundary based variants, i.e. 1) com_Wiener(A) with com_Wiener(F), and 2) com_Wiener_C(A) with com_Wiener_C(F). As can be seen, the adaptive LF boundary based variant significantly improved the PESQ scores compared with its paired fixed LF boundary based variant when speech was corrupted by white noise, babble and car interferences.

4. CONCLUSIONS

This paper presented a perceptually motivated design for the gain function of Wiener filtering. A less aggressive gain function is incorporated into Wiener filtering to replace its conventional rigid gain function at low frequencies. The less aggressive gain function was shown to recover more low frequency components by alleviating over attenuation. An adaptive procedure was used to determine the low frequency boundary. When the LF boundary was computed adaptively, the proposed speech enhancement approach produced significantly higher PESQ scores, compared to the conventional Wiener filtering for steady-state noise, white noise, and babble interferences. When the less aggressive gain function was applied selectively to consonants only, an improvement in speech quality was noted in the car interference condition.

5. ACKNOWLEDGEMENT

This research was supported in part by Grant R01-DC07527 from NIDCD/NIH.

6. REFERENCES