EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Intra-frame cepstral sub-band weighting and histogram equalization for noise-robust speech recognition

Jeih-weih Hung* and Hao-teng Fan

Abstract

In this paper, we propose a novel noise-robustness method known as weighted sub-band histogram equalization (WS-HEQ) to improve speech recognition accuracy in noise-corrupted environments. Considering the observations that high- and low-pass portions of the intra-frame cepstral features possess unequal importance for noise-corrupted speech recognition, WS-HEQ is intended to reduce the high-pass components of the cepstral features. Furthermore, we provide four types of WS-HEQ, which partially refers to the structure of spatial histogram equalization (S-HEQ). In the experiments conducted on the Aurora-2 noisy-digit database, the presented WS-HEQ yields significant recognition improvements relative to the Mel-scaled filter-bank cepstral coefficient (MFCC) baseline and to cepstral histogram normalization (CHN) in various noise-corrupted situations and exhibits a behavior superior to that of S-HEQ in most cases.

Keywords: Sub-band division; Speech recognition; Robust speech features; Histogram equalization

1 Introduction

The performance of speech recognition systems is often degraded due to noise in application environments. A significant number of noise-robustness techniques have been proposed to address the noise problem, and one prevailing subset of these techniques is focused on reducing the statistical mismatch of speech features in the training and testing conditions of the recognizer. Typical examples are perceptual masking [1], empirical mode decomposition [2], optimally modified log-spectral amplitude estimation [3], wavelet packet decomposition with AR modeling [4], cepstral mean and variance normalization (MVN) [5], cepstral histogram normalization (CHN) [6,7], MVN with ARMA filtering (MVA) [8], higher order cepstral moment normalization (HOCMN) [9], and temporal structure normalization (TSN) [10]. In some of these methods, the compensation is performed on each individual cepstral channel sequence of an utterance by assuming that these channels are mostly uncorrelated [7].

Recently, certain studies have investigated the use of cepstral frame-based processing to compensate for the noise effect to achieve better recognition accuracy. For example, the work in [11] revealed that in the CHN method, even though each cepstral channel is processed by histogram equalization (HEQ), a significant histogram mismatch still exists among the training and testing cepstral features for the low-pass filtered (LPF) and high-pass filtered (HPF) portions of the intra-frame cepstra. Thus, the method of spatial HEQ in [11] further performs HEQ on the LPF and HPF portions to eliminate the aforementioned mismatch for the CHN-preprocessed cepstra. Compared with conventional CHN that processes each individual cepstral channel, spatial HEQ (S-HEQ) additionally takes the neighboring cepstral channels into consideration collectively and produces superior noise robustness. Furthermore, for a frame signal, the LPF and HPF portions of the cepstral vector just correspond to the logarithmic filter-bank (LFB) components at lower and higher frequencies, respectively. However, compensation performed directly on LPF and HPF is more helpful than that applied to the LFB components, most likely because the LFB components are significantly correlated [11].

*Correspondence: jwhung@ncnu.edu.tw
Department of Electrical Engineering, National Chi Nan University, Nantou 545, Taiwan

Partly inspired by S-HEQ, here we develop a novel scheme known as the weighted S-HEQ (WS-HEQ) to improve the recognition performance and operation efficiency of S-HEQ in three directions. First, because the LPF and HPF portions of the original or CHN-preprocessed cepstra possess different characteristics in noisy environments and provide unequal contributions to the recognition accuracy, we tune the portion of HPF produced in the original S-HEQ and show that this adjustment can outperform S-HEQ in recognition accuracy. Second, we change the order of the procedures in S-HEQ by first splitting the original intra-frame cepstra (not the CHN-preprocessed cepstra) into LPF and HPF, subsequently compensating LPF and HPF individually, and finally, normalizing the full-band cepstra. This new structure can reduce the effect of noise on the LPF and HPF portions in the plain cepstra more directly in comparison with S-HEQ. Finally, because S-HEQ requires three HEQ operations, we use the simpler process of MVN to replace any of the three HEQ processes in S-HEQ to improve the computational efficiency. The experimental results show that some variants of WS-HEQ, which require fewer HEQ operations, provide a similar or even better recognition accuracy relative to S-HEQ.

The remainder of this paper is organized as follows. Section 2 reviews S-HEQ, and the basic concept and detailed procedures of the proposed WS-HEQ are presented in Section 3. Section 4 describes the experimental setup, and Sections 5 and 6 contain a series of recognition experiments for WS-HEQ together with their corresponding discussions. Finally, the concluding remarks are summarized in Section 7.

2 Brief review of S-HEQ

If we consider using the Mel-scaled filter-bank cepstral coefficients (MFCC) as the baseline features for speech recognition, then the cepstral feature vector stream associated with an arbitrary utterance is represented by a matrix **C**:

$$\mathbf{C} = \{c(m,n); 0 \leq m \leq M-1, 0 \leq n \leq N-1\}, \qquad (1)$$

where m is the cepstral channel index within a frame and n is the frame index, and M and N are the total number of channels and frames within the utterance, respectively. In the temporal processing methods as MVN and CHN, the compensation is often directly performed on the individual channel stream (i.e., the sequence $\{c(\tilde{m}, n); 0 \leq n \leq N-1\}$ with respect to the $\tilde{m}$th channel), and therefore, all of the channel streams of the features are treated independently. According to the general concept that the cepstral coefficients within a frame are mostly uncorrelated [7], such a process is quite reasonable.

Recently, a novel method known as the spatial HEQ (S-HEQ) was suggested to decompose each frame of a CHN-preprocessed cepstral vector into two parts, a high-pass filtered and low-pass filtered portion (denoted hereafter as HPF and LPF), such that the temporal sequences of HPF and LPF can be processed separately and then the updated HPF and LPF can be combined to form the new feature vector stream. The work in [11] shows that S-HEQ outperforms the conventional CHN by providing better recognition accuracy. The overall procedure of S-HEQ is depicted in Figure 1.

3 Proposed approach: WS-HEQ

S-HEQ [11] offers additional insight into the possible distortions left unprocessed by CHN and a method for achieving even better noise robustness for speech features. In this section, we further examine S-HEQ to assess whether it can be further improved. The following two observations can be made about S-HEQ:

1. S-HEQ divides each CHN-preprocessed cepstral vector into HPF and LPF and subsequently treats the temporal stream of these two parts in the same manner (i.e., with HEQ processing). Therefore, S-HEQ does not consider the characteristic differences between HPF and LPF. According to [11], the plain HPF (from the original cepstra, not the CHN-preprocessed cepstra) is often more vulnerable to noise and displays more mismatch than the plain LPF, whereas S-HEQ compensates for the CHN-preprocessed HPF and LPF directly. Additionally, HPF and LPF possess unequal importance in speech recognition, which will be shown later.
2. In S-HEQ, the HEQ operation is repeated up to three times: one for the original feature stream set and the other two for the HPF and LPF stream sets. Thus, S-HEQ requires twice more computational effort than the conventional CHN method, which only processes the original stream set once via HEQ.

In this work, we design a simple experiment to evaluate the relative importance of different sub-bands of the cepstral features in speech recognition. With the Aurora-2

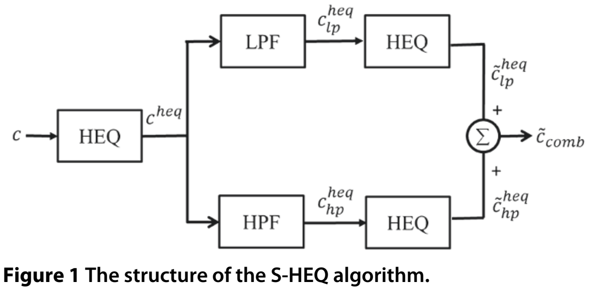


Figure 1 The structure of the S-HEQ algorithm.

database [12], we select 8,440 clean utterances for the clean-condition training task as the data used to train the acoustic models and 8,440 noisy utterances (corrupted by any of four types of noise at five signal-to-noise ratios) originally for the multi-condition training task as the testing data. Each utterance in the training and testing sets is first converted into a sequence of 13-dimensional cepstral vectors ($c0$, $c1$ to $c12$). The obtained cepstra are either kept unchanged or processed by CHN. Next, for each original/CHN-processed cepstral vector, we obtain its 'sub-band' version with the following two steps:

Step 1. Find the spectrum of the cepstral vector via discrete Fourier transform (DFT):
Let $\mathbf{c} = [c_0\, c_1\, c_2\, \ldots\, c_{12}]^T$ denote an arbitrary cepstral vector, and its spectrum is obtained by

$$C[k] = \sum_{m=0}^{12} c_m e^{-j\frac{2\pi mk}{13}}, 0 \le k \le 12. \qquad (2)$$

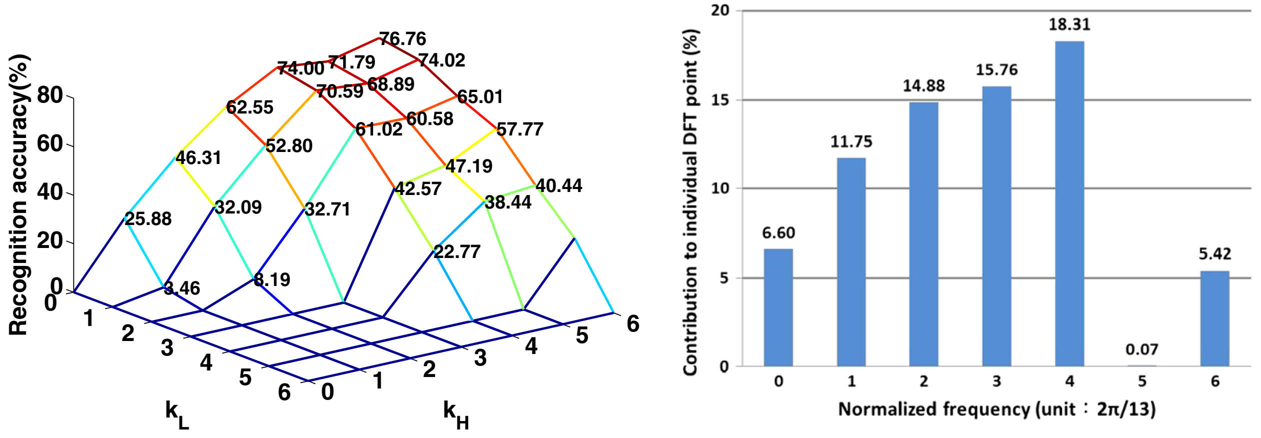
Due to the conjugate symmetry of $\{C[k]\}$, we only need to retain the first seven points, which correspond to $\{k\frac{2\pi}{13}; 0 \le k \le 6\}$ in normalized frequency.

Step 2. Retain a contiguous portion of the spectral points and transform them (together with their conjugate symmetric parts) into a new cepstral vector via inverse DFT. For example, if we retain the first to fifth spectral points unchanged and set the zeroth and the sixth spectral points to zero, then the resulting new cepstral vector is a sub-band version of the original cepstral vector and corresponds approximately to the band range of $[\frac{2\pi}{13}, \frac{10\pi}{13}]$.

The recognition accuracy rates for different cepstral features obtained from the above sub-band processing are shown in Figures 2a and 3a, the former being for the original cepstra and the later being for the CHN-processed cepstra (Please note that the testing data undergo the same process as the training data in the recognition experiment. Therefore, the original testing cepstra are recognized by the acoustic models trained from the original training cepstra, and the CHN-processed testing cepstra are recognized by the acoustic models trained from the CHN-processed training cepstra). The vertical axis in Figures 2a and 3a denotes the word accuracy rate, and the other two axes indicate the initial and final spectral points, k_L and k_H, of the assigned sub-band, respectively. Obviously, the CHN-processed cepstra outperform the original cepstra in recognition results. Besides, for both types of cepstra the full-band features are always able to achieve the highest accuracy, and decreasing the bandwidth of the sub-band worsens the accuracy. However, we can further evaluate the relative importance of different spectral points in the sub-band from the two figures using the following equation:

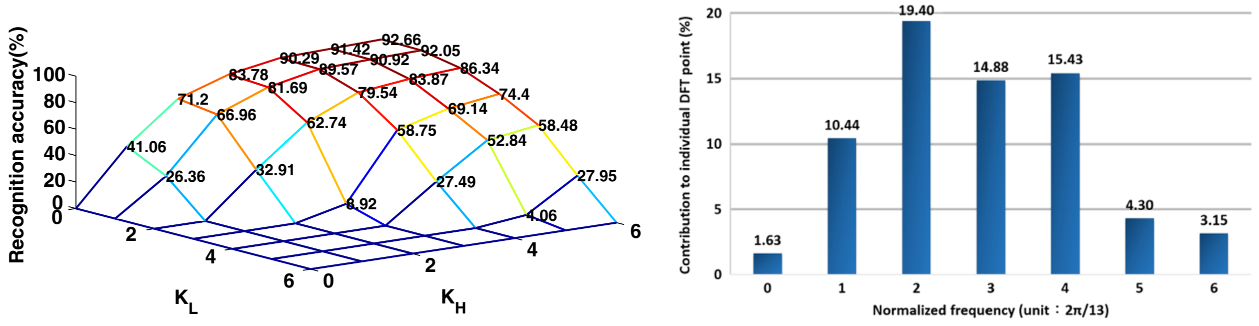
$$r_m = \frac{1}{N_r}\left(\sum_{k>m+1}(R_{m,k} - R_{m+1,k}) + \sum_{k<m-1}(R_{k,m} - R_{k,m-1})\right), \qquad (3)$$

where r_m denotes the averaged contribution of the mth spectral points, $R_{m,k}$ is the recognition rate using the cepstra within the sub-band including the mth to kth spectral points, and N_r is the total number of items in the summation of Equation 3. (The term 'relative importance' and its definition shown in Equation 3



(a) (b)

Figure 2 Some information about the DFT-based spectrum of cepstra without CHN processing. (a) Recognition rates for the band-pass filtered cepstra. **(b)** The contribution of each individual spectral point.



(a) (b)

Figure 3 Some information about the DFT-based spectrum of CHN-processed cepstra. (a) Recognition rates for the band-pass filtered cepstra. **(b)** The contribution of each individual spectral point.

are borrowed from [13], in which a series of band-pass filters are used to evaluate the various *modulation* spectral components in their contribution to the recognition accuracy.) The obtained results from the original and the CHN-processed cepstra are shown in Figures 2b and 3b, respectively. Note that in Equation 3, the number of spectral points in the assigned sub-band range is always greater than or equal to 2 because the cepstra associated with a single spectral point quite often result in a rather poor (even negative) recognition accuracy.

From Figures 2b and 3b, the seven spectral points possess unequal importance in noisy speech recognition. The middle and lower frequency points (except for the DC point) seem to contribute more to the recognition accuracy than the upper points. These results suggest that alleviating the higher frequency components in the cepstra more likely results in better recognition performance in a noisy environment. Besides, comparing Figure 3b with Figure 2b, we find that the CHN process helps the higher frequency points to reinforce their importance in speech recognition, especially for the point at frequency $\frac{10\pi}{13}$.

The spectrum of the cepstra in the aforementioned evaluation experiment is created via the DFT, with the main reason that low-pass and high-pass filters are to be applied to the cepstra in later discussions, and we often evaluate the effect of a filter on the processed signal in the Fourier-based frequency domain. Also, in most cases, the characteristics of a filter are investigated by its frequency response; the Fourier transform, of its impulse response. However, since each frame-wise cepstral vector is the truncated version of the inverse discrete cosine transform (IDCT) of the logarithmic spectrum of the corresponding frame, here we reconduct the preceding evaluation experiment based on the 'DCT-based' spectrum of the original/CHN-processed cepstra. That is, in step 1 of the experiment, we obtain the 13-point spectrum of any arbitrary cepstral vector $\mathbf{c} = [c_0\, c_1\, c_2\, \ldots\, c_{12}]^T$ via DCT:

$$\tilde{C}[k'] = \sum_{m=0}^{12} c_m \cos\left(\frac{\pi(m+\frac{1}{2})k'}{13}\right), \quad 0 \le k' \le 12, \tag{4}$$

and then in step 2, a contiguous portion of the DCT-based spectral points is retained and transformed into a new cepstral vector via IDCT.

Some differences between the DCT-based spectrum $\{\tilde{C}[k']\}$ in Equation 4 and DFT-based spectrum $\{C[k]\}$ in Equation 2 are as follows:

1. Unlike the DFT-based spectrum $\{C[k]\}$ which is complex-valued and conjugate symmetric, in general, the real-valued DCT-based spectrum $\{\tilde{C}[k']\}$ is not symmetric in any sense. Thus, we cannot discard the second half points of $\{\tilde{C}[k']\}$ as we do on $\{C[k]\}$.
2. $\{\tilde{C}[k']\}$ possesses a higher frequency resolution than $\{C[k]\}$. Comparing Equation 4 with Equation 2, the frequency difference between any two adjacent bins of $\{\tilde{C}[k']\}$ is $\frac{\pi}{13}$, while it is $\frac{2\pi}{13}$ for $\{C[k]\}$.
3. Referring to [14], the N-point DCT, $\{\tilde{C}[k']\}$, of a length-N sequence $\{c[n], 0 \le n \le N-1\}$ (here $N = 13$), can be computed via a $2N$ DFT of another length-$2N$ sequence $\{\tilde{c}[n], 0 \le n \le 2N-1\}$, denoted by $\{D[k']\}$, in which $\tilde{c}[n]$ is the even extension of $c[n]$ satisfying $\tilde{c}[n] = c[n]$ for $0 \le n \le N-1$ and $\tilde{c}[n] = x[2N-1-n]$ for $N \le n \le 2N-1$. $\{\tilde{C}[k']\}$ and $\{D[k']\}$ are related by:

$$\tilde{C}[k'] = 0.5e^{-j\frac{\pi k}{2N}} D[k'] \text{ for } 0 \le k \le N-1. \tag{5}$$

Generally speaking, the DCT-based spectrum $\{\tilde{C}[k']\}$ is more concentrated at low frequencies than the DFT-based spectrum $\{C[k]\}$, which is well known as the 'energy compaction property' of DCT. An underlying reason for this phenomenon is that

DFT implicitly assumes the periodic extension of the processed signal and often causes the artificial discontinuities at the signal boundary, which adds high frequency contents in the DFT-based spectrum. To show this, a length-N sequence $\{x[n], 0 \leq n \leq N-1\}$ is treated by N-point DFT as an N-periodic signal, denoted by $x_e[n]$, in which $x_e[n] = x[n]$ for $0 \leq n \leq N-1$ and $x_e[n+N] = x_e[n]$. Thus, $x_e[n]$ is generally discontinuous at the (original) boundary positions:

$$x_e[0] = x[0] \neq x[N-1] = x_e[-1]\,, \tag{6}$$

$$x_e[N-1] = x[N-1] \neq x[0] = x_e[N]\,. \tag{7}$$

However, as mentioned earlier, the N-point DCT of a length-N sequence $\{x[n]\}$ (starting at $n = 0$) can be obtained from the $2N$-point DFT of the even extension of $\{x[n]\}$, and the corresponding $2N$-periodic signal, denoted by $\tilde{x}_e[n]$, remains continuous at the boundary positions:

$$\tilde{x}_e[0] = \tilde{x}[0] = \tilde{x}[2N-1] = \tilde{x}_e[2N-1] = \tilde{x}_e[-1]\,, \tag{8}$$

$$\tilde{x}_e[2N-1] = \tilde{x}[2N-1] = \tilde{x}[0] = \tilde{x}_e[0] = \tilde{x}_e[2N]\,. \tag{9}$$

As a result, the (N-point) DCT-based spectrum does not contain the high frequency artifacts as the (N-point) DFT-based spectrum, and it appears more compact at low frequencies.

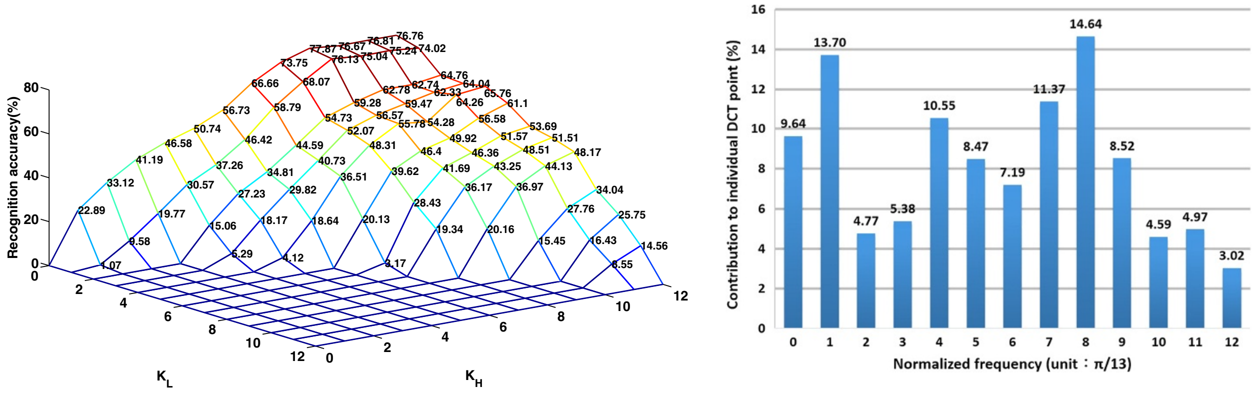
With the cepstra from the IDCT of sub-band DCT-based spectra, the corresponding evaluation experiment is performed to obtain the recognition accuracy rates, which are shown in Figures 4a and 5a, and the relative importance of different spectral points are shown in Figures 4b and 5b. Figure 4a,b is for the original cepstra and Figure 5a,b is for the CHN-processed cepstra. These two figures roughly reveal that the lower and middle DCT-based spectral points contribute to the recognition more than the upper ones in recognition, which somewhat coincides our observations from Figures 2a,b and 3a,b associated with the DFT-based spectra. In addition, comparing Figure 4b with Figure 2b and Figure 5b with Figure 3b, we find that the higher DCT-based spectral points reveal more importance than the higher DFT-based spectral points. which partially agrees with our previous statement that the DFT-based spectrum contains some artificial high frequency contents, which distort the higher spectral points and reduce the corresponding contribution.

In light of the aforementioned discussions, we developed a novel method known as the WS-HEQ to enhance the speech features in noise robustness. The initial concept of WS-HEQ is to apply a weighting factor to the HPF portion in S-HEQ (as shown in Figure 1) to reduce the intra-frame higher frequency components, and we further provide several variations on the presented WS-HEQ. First, according to the order of the HEQ processing for the full-band cepstra and sub-band cepstra, we describe two structures:

Structure I. HEQ first operates on the plain (intra-frame) *full-band* cepstra and subsequently on the *sub-band* cepstra.

Structure II. HEQ first operates on the plain (intra-frame) *sub-band* cepstra and subsequently on the *full-band* cepstra.

Please note that in the above two structures, the two sub-band cepstral portions, LPF and HPF, are obtained with



(a)

(b)

Figure 4 Some information about the DCT-based spectrum of cepstra without CHN processing. (a) Recognition rates for the band-pass filtered cepstra. **(b)** The contribution of each individual spectral point.

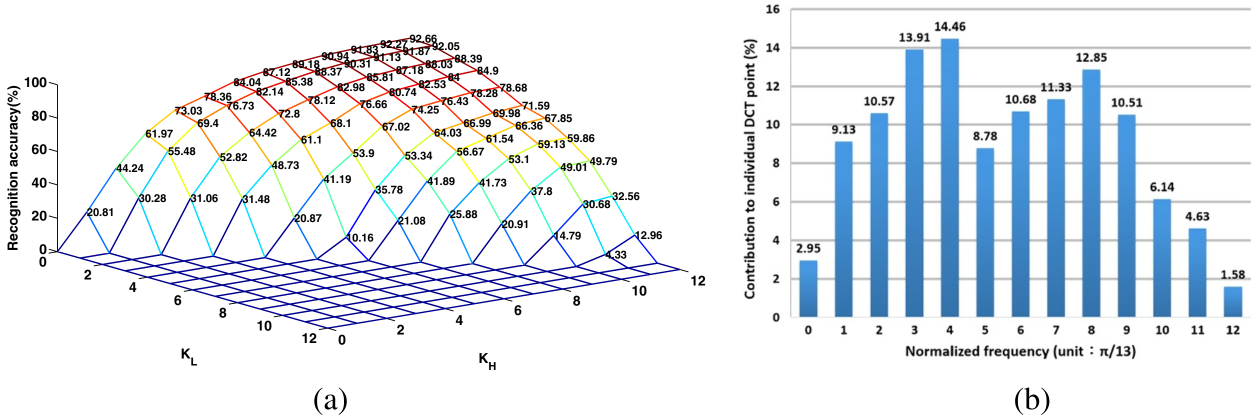


Figure 5 Some information about the DCT-based spectrum of CHN-processed cepstra. (a) Recognition rates for the band-pass filtered cepstra. **(b)** The contribution of each individual spectral point.

simple two-point FIR filters operating on the full-band cepstra [11]:

$$\text{LPF:} c_{lp}(m,n) = \frac{c(m,n) + c(m-1,n)}{2}, \tag{10}$$

$$\text{HPF: } c_{hp}(m,n) = \frac{c(m,n) - c(m-1,n)}{2}, \tag{11}$$

where $c_{lp}(m,n)$ and $c_{hp}(m,n)$ denote the low-pass and high-pass filtered parts of the nth cepstral frame.

Next, according to different treatments (i.e., the compensation methods, HEQ, and MVN) of LPF and HPF in Equations 10 and 11, each structure of WS-HEQ has the following four types of variations:

$$\text{Type 1:} \tilde{c}_{lp} = \mathbf{HEQ}[c_{lp}]\,, \tilde{c}_{hp} = \alpha\mathbf{HEQ}[c_{hp}]\,, \tag{12}$$

$$\text{Type 2:} \tilde{c}_{lp} = \mathbf{MVN}[c_{lp}]\,, \tilde{c}_{hp} = \alpha\mathbf{HEQ}[c_{hp}]\,, \tag{13}$$

$$\text{Type 3:} \tilde{c}_{lp} = \mathbf{HEQ}[c_{lp}]\,, \tilde{c}_{hp} = \alpha\mathbf{MVN}[c_{hp}]\,, \tag{14}$$

$$\text{Type 4:} \tilde{c}_{lp} = \mathbf{MVN}[c_{lp}]\,, \tilde{c}_{hp} = \alpha\mathbf{MVN}[c_{hp}]\,, \tag{15}$$

where $\mathbf{HEQ}[\cdot]$ and $\mathbf{MVN}[\cdot]$ denote the operators of the HEQ and MVN processes, respectively; $\tilde{c}_{lp}$ and $\tilde{c}_{hp}$ are the updated LPF and HPF, respectively (we omit the indices (m,n) for simplicity); and the parameter α with a range of $[0,1]$ is the scaling factor selected specifically for the HPF component. The flowcharts of the various structures and types of WS-HEQ are depicted in Figure 6a,b.

For clarity, in the following discussions, the term 'WS-HEQ' is written with an additional subscript of 'I' or 'II', and a superscript of '(1)', '(2)', '(3)', or '(4)' to identify different structures and different processing schemes for LPF and HPF in the presented WS-HEQ method. For example, $\text{WS-HEQ}_{\text{II}}^{(3)}$ indicates that the WS-HEQ method applying the second structure shown in Figure 6b and uses HEQ and MVN for the LPF and HPF portions, respectively. Additional discussions on the various forms of WS-HEQ are given:

1. Because the HEQ operation is nonlinear, $\text{WS-HEQ}_{\text{II}}^{(1)}$ with $\alpha = 1.0$ (no attenuation for HPF) as shown in Equation 12 in which HEQ is first performed on the sub-band cepstra and subsequently on the full-band cepstra, is different from S-HEQ (equivalent to $\text{WS-HEQ}_{\text{I}}^{(1)}$ with $\alpha = 1.0$) shown in Figure 1, in

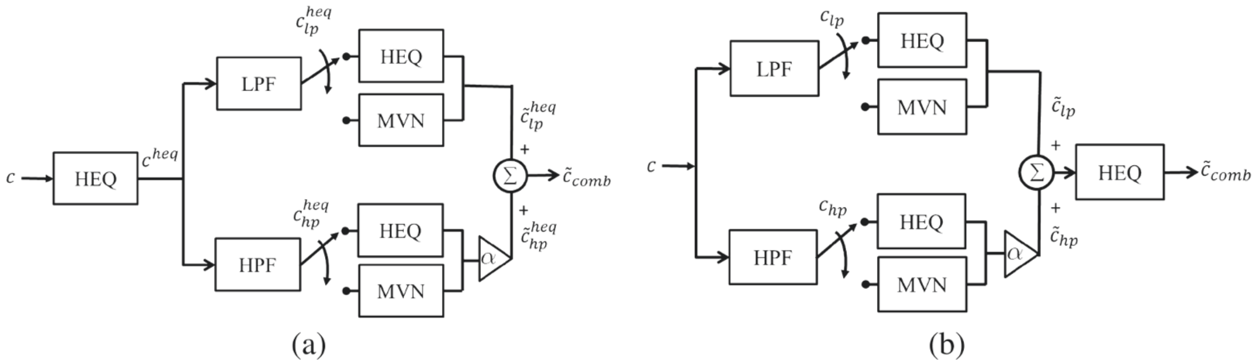


Figure 6 The flowcharts of two structures of WS-HEQ. (a) Structure I and **(b)** structure II.

which the full-band cepstra are HEQ-processed in advance.

2. In the first type of WS-HEQ$_{\text{II}}$, (*viz.* WS-HEQ$_{\text{II}}^{(1)}$), both LPF and HPF (of the original MFCC) are processed by HEQ. The resulting new HPF is attenuated by a factor of α and then combined with the new LPF to form the full-band cepstra, which are further processed by HEQ in the final stage. Therefore, WS-HEQ$_{\text{II}}^{(1)}$ requires three HEQ operations, the same as S-HEQ, demonstrating that S-HEQ and WS-HEQ$_{\text{II}}^{(1)}$ are similar in computational complexity.
3. The other three types of WS-HEQ as shown in Equations 13 to 15 differ from the first type in that they compensate either or both of the LPF and HPF portions via MVN instead of HEQ. MVN can be implemented more efficiently than HEQ because MVN involves only the operations of addition and multiplication, whereas a sorting algorithm is required in HEQ. We expect that the cost savings of HEQ on HPF/LPF will not affect the prospective recognition accuracy.

4 Experimental setup

The performance of our proposed WS-HEQ scheme is examined in two databases. One is the Aurora-2 database [12] corresponding to a connected English-digit recognition task, and the other is a subset of the TCC-300 database [15] for the recognition of 408 Chinese syllables. Briefly speaking, we conduct more comprehensive experiments with the Aurora-2 database for analysis and comparison upon the various forms of the presented WS-HEQ together with some other robustness algorithms, and a smaller number of experiments conducted on the subset of the TCC database are simply to examine if the presented WS-HEQ can be extended to work well in a median-size vocabulary recognition task which is more complicated than Aurora-2. Furthermore, in order to avoid the ambiguity and confusion in discussion, the remainder of this section and Section 5 are specially for the Aurora-2 evaluation task, while the detailed discussions about the TCC-300 subset task will given in Section 6.

As for the Aurora-2 database, the test data consist of 4,004 utterances, and three different subsets are defined for the recognition experiments: test sets A and B are both affected by four types of noise, and test set C is affected by two types. Each noise instance is artificially added to the clean speech signal at seven SNR levels (ranging from 20 to −5 dB). The signals in test sets A and B are filtered with a G.712 filter, and those in Set C are filtered with an MIRS filter. In the 'clean-condition training, multi-condition testing' evaluation task defined in [12], the training data consist of 8,440 noise-free clean utterances filtered with a G.712 filter. Thus, compared with the training data, test sets A and B are distorted by additive noise, and test set C is affected by additive noise and a channel mismatch.

In the experiments, each utterance in the clean training set and the three testing sets is first converted to a 13-dimensional MFCC ($c0$, $c1$ to $c12$) sequence. Next, the MFCC features are processed by either S-HEQ [11] or the various forms of WS-HEQ noted in Section 3. In addition, the selected target distribution of the HEQ operation applied to any of the full-band, LPF, and HPF cepstra is the standard normal (Gaussian), with a zero mean and unity variance. (Please note that, given the full-band cepstral sequences being standard normal and approximately mutually uncorrelated, the corresponding LPF and HPF via the operations in Equations 10 and 11 are also standard normal. Similarly, if the HPF and LPF are both standard normal and approximately mutually uncorrelated, then the corresponding full-band cepstra are normally distributed with a zero mean and a variance of less than 1 since we scale down the HPF portion.)

The resulting 13 new features, in addition to their first- and second-order derivatives, are the components of the final 39-dimensional feature vector. With the new feature vectors in the clean training set, the hidden Markov models (HMMs) for each digit and for silence are trained with the scripts provided by the Aurora-2 CD set [16]. Each HMM digit contains 16 states, with three Gaussian mixtures per state.

In particular, the 8,440 noisy utterances (corrupted by four types of noise at five signal-to-noise ratios) originally for the multi-condition training task [12], which has been mentioned earlier in Section 3, are served as the *development set* here in order to obtain an appropriate selection of the scaling factor α for the HPF portion in Equations 12 to 15. The value of α is varied from 0.0 to 1.0 with an interval of 0.1 in each form of WS-HEQ, and then the one that achieves the optimal recognition accuracy for the development set is chosen for the corresponding WS-HEQ in practice. The selected values of α for different forms of WS-HEQ are listed in Table 1.

Table 1 Scaling factor α for each type of WS-HEQ

Structure I		Structure II	
Method	**Optimal α**	**Method**	**Optimal α**
WS-HEQ$_{\text{I}}^{(1)}$	0.6	WS-HEQ$_{\text{II}}^{(1)}$	0.6
WS-HEQ$_{\text{I}}^{(2)}$	0.6	WS-HEQ$_{\text{II}}^{(2)}$	0.6
WS-HEQ$_{\text{I}}^{(3)}$	0.5	WS-HEQ$_{\text{II}}^{(3)}$	0.7
WS-HEQ$_{\text{I}}^{(4)}$	0.7	WS-HEQ$_{\text{II}}^{(4)}$	0.6

It gives the optimal recognition accuracy for each WS-HEQ variant in the development set as to the Aurora-2 database.

Table 2 The recognition accuracy results (%) of the MFCC baseline, CHN, S-HEQ, and WS-HEQ with structure I

Method		Set A	Set B	Set C	Average	RR
MFCC		59.24	56.37	67.53	59.75	-
CHN		79.28	81.53	79.98	80.32	51.11
WS-HEQ$_{\mathrm{I}}^{(1)}$	$\alpha = 1.0$ (S-HEQ)	81.56	84.51	80.78	82.58	56.73
	$\alpha = 0.6$	83.36	85.37	83.89	84.27	60.92
WS-HEQ$_{\mathrm{I}}^{(2)}$	$\alpha = 1.0$	80.88	83.64	80.46	81.90	55.04
	$\alpha = 0.6$	82.29	83.22	82.82	82.76	57.16
WS-HEQ$_{\mathrm{I}}^{(3)}$	$\alpha = 1.0$	79.66	82.51	79.33	80.73	52.13
	$\alpha = 0.5$	83.57	85.15	83.93	84.27	60.92
WS-HEQ$_{\mathrm{I}}^{(4)}$	$\alpha = 1.0$	80.20	82.82	80.18	81.24	53.39
	$\alpha = 0.7$	82.88	84.70	82.78	83.59	59.23

They are for different test sets while averaged over five SNR conditions (20 to 0 dB) as to the Aurora-2 database. RR (%) is the relative error rate reduction compared with the MFCC baseline.

5 Experimental results and discussions for the Aurora-2 task

5.1 Recognition accuracy

The presented WS-HEQ is evaluated in terms of recognition accuracy. Tables 2 and 3 show the individual set recognition accuracy rates averaged over five SNR conditions (0 to 20 dB, with a 5-dB interval) for the MFCC baseline, CHN, S-HEQ (equivalent to WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 1.0$), and various forms of the presented WS-HEQ, while Table 4 further lists the recognition accuracy rates for each individual SNR situations but averaged over ten noise situations. In addition, Figure 7 depicts the overall averaged word error rates achieved by several methods, including MVA, HOCMN, TSN, CHN, S-HEQ, WS-HEQ$_{\mathrm{I}}^{(1)}$($\alpha = 0.6$), and WS-HEQ$_{\mathrm{II}}^{(1)}$($\alpha = 0.6$). From Tables 2,3,4 and Figure 7, we have the following findings:

1. Compared with the MFCC baseline, all of the HEQ-related methods provide very similar accuracy rates for the clean situation, and they are able to provide significant improvement in recognition accuracy for various noise-corrupted situations, showing that HEQ is quite helpful for speech features in terms of noise robustness.
2. S-HEQ (WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 1.0$) outperforms CHN by around 2.3% in the averaged accuracy, and thus, further manipulation of the mismatch in LPF and HPF with two extra HEQ operations can benefit the recognition performance.
3. WS-HEQ$_{\mathrm{II}}^{(1)}$ with $\alpha = 1.0$ produces results similar to those of S-HEQ, and thus, the proposed structure II (shown in Figure 6b) performs quite well. Additionally, provided that no attenuation exists for HPF by setting $\alpha = 1.0$, using structure II in the other three types of WS-HEQ, i.e., WS-HEQ$_{\mathrm{II}}^{(2)}$, WS-HEQ$_{\mathrm{II}}^{(3)}$, and WS-HEQ$_{\mathrm{II}}^{(4)}$ as shown in Equations 13 to 15, outperforms the respective

Table 3 The recognition accuracy results (%) of the MFCC baseline, CHN, and WS-HEQ with structure II

Method		Set A	Set B	Set C	Average	RR
MFCC		59.24	56.37	67.53	59.75	-
CHN		79.28	81.53	79.98	80.32	51.11
WS-HEQ$_{\mathrm{II}}^{(1)}$	$\alpha = 1.0$	81.75	84.61	80.81	82.70	57.03
	$\alpha = 0.6$	84.13	86.16	84.39	84.99	62.71
WS-HEQ$_{\mathrm{II}}^{(2)}$	$\alpha = 1.0$	82.59	84.93	82.03	83.41	58.79
	$\alpha = 0.6$	83.54	85.75	83.83	84.48	61.44
WS-HEQ$_{\mathrm{II}}^{(3)}$	$\alpha = 1.0$	80.55	83.99	79.80	81.77	54.72
	$\alpha = 0.7$	83.25	85.10	83.50	84.04	60.35
WS-HEQ$_{\mathrm{II}}^{(4)}$	$\alpha = 1.0$	80.83	83.44	80.37	81.78	54.73
	$\alpha = 0.6$	82.30	83.45	82.90	82.88	57.47

They are for different test sets while averaged over five SNR conditions (20 to 0 dB) as to the Aurora-2 database. RR (%) is the relative error rate reduction compared with the MFCC baseline.

Table 4 The recognition accuracy results (%) of the MFCC baseline, CHN, and eight forms of WS-HEQ

Method		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
MFCC		99.12	95.33	86.62	65.93	36.01	14.86	8.16
CHN		98.97	96.30	93.89	88.48	74.81	48.10	19.94
WS-HEQ$_{\mathrm{I}}^{(1)}$	$\alpha = 1.0$ (S-HEQ)	99.02	97.24	94.99	90.09	77.88	52.73	22.68
	$\alpha = 0.6$	98.99	97.60	95.72	91.54	80.62	55.87	23.26
WS-HEQ$_{\mathrm{I}}^{(2)}$	$\alpha = 1.0$	99.02	97.34	95.14	90.10	77.38	49.57	18.40
	$\alpha = 0.6$	99.09	97.68	95.64	91.30	79.23	49.97	17.69
WS-HEQ$_{\mathrm{I}}^{(3)}$	$\alpha = 1.0$	98.96	96.89	94.45	89.02	75.55	47.78	17.26
	$\alpha = 0.5$	98.99	97.51	95.71	91.59	80.59	55.95	23.88
WS-HEQ$_{\mathrm{I}}^{(4)}$	$\alpha = 1.0$	98.93	97.20	95.01	89.75	76.57	47.68	16.92
	$\alpha = 0.7$	99.07	97.87	96.18	91.98	79.99	51.94	19.83
WS-HEQ$_{\mathrm{II}}^{(1)}$	$\alpha = 1.0$	98.84	96.86	94.61	89.76	78.00	54.30	24.91
	$\alpha = 0.6$	99.05	97.53	95.75	91.99	81.35	58.35	26.98
WS-HEQ$_{\mathrm{II}}^{(2)}$	$\alpha = 1.0$	98.87	97.41	95.50	91.15	79.06	53.96	22.29
	$\alpha = 0.6$	99.04	97.60	95.70	91.68	80.83	56.61	24.45
WS-HEQ$_{\mathrm{II}}^{(3)}$	$\alpha = 1.0$	98.89	96.76	94.29	88.97	76.42	52.44	23.69
	$\alpha = 0.7$	99.05	97.87	95.98	91.86	80.62	53.88	20.51
WS-HEQ$_{\mathrm{II}}^{(4)}$	$\alpha = 1.0$	98.92	97.13	94.84	90.07	76.92	49.96	19.89
	$\alpha = 0.6$	98.96	97.49	95.54	91.17	79.00	51.23	19.37

They are for different SNR cases while averaged over ten noise situations as to the Aurora-2 database.

methods under structure I. In particular, WS-HEQ$_{\mathrm{II}}^{(2)}$ behaves better than WS-HEQ$_{\mathrm{II}}^{(1)}$, whereas WS-HEQ$_{\mathrm{I}}^{(2)}$ behaves worse than WS-HEQ$_{\mathrm{I}}^{(1)}$, revealing that applying structure II can make WS-HEQ less costly in computation and can obtain improved recognition results simultaneously.

4. Reducing the HPF component by setting the factor α as less than 1.0 as in Table 1 significantly improves the recognition accuracy, regardless of the different structures and types of WS-HEQ. WS-HEQ$_{\mathrm{II}}^{(1)}$ gives an averaged accuracy of 84.99%, which is optimal among all of the methods and corresponds to error reduction rates of 62.71%, 23.73%, and 13.83% relative to the MFCC baseline, CHN, and S-HEQ, respectively. These results support the aforementioned observations that HPF is more extensively contaminated by noise and that lo wering HPF is beneficial.
5. Among the four types of WS-HEQ listed in Equations 12 to 15, by assigning α as less than 1.0, WS-HEQ$^{(1)}$, which requires three HEQ operations, displays the best behavior, regardless of the selected structure. However, the two types that require only two HEQ operations (i.e., WS-HEQ$^{(2)}$ and

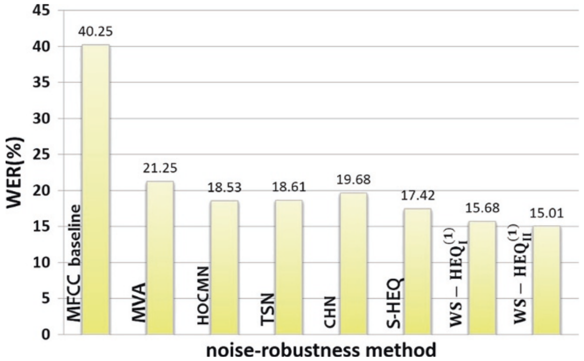


Figure 7 Overall word error rate (%) averaged over all noise types and levels achieved by different noise-robustness methods.

WS-HEQ$^{(3)}$) perform quite similarly to WS-HEQ$^{(1)}$ when structure II is used. Finally, WS-HEQ$^{(4)}$ performs worse than the other three types, possibly because it applies only one HEQ operation. Even so, WS-HEQ$_{\text{I}}^{(4)}$ and WS-HEQ$_{\text{II}}^{(4)}$ with $\alpha = 0.6$ can behave very close to S-HEQ (WS-HEQ$_{\text{I}}^{(1)}$ with $\alpha = 1.0$).

6. The presented WS-HEQ$_{\text{II}}^{(1)}$ with $\alpha = 0.6$ behaves better in the overall averaged word error rate when compared with several well-known noise-robustness methods: TSN, HOCMN, MVA, CHN, and S-HEQ. The absolute error rate reduction of WS-HEQ$_{\text{II}}^{(1)}$ with $\alpha = 0.6$ relative to the MFCC baseline is as high as 25.24%.

Taking a step further, among the methods used for comparison, MVA and TSN explicitly applies a temporal filter, and in most cases, the used filter is low pass so as to perform a 'temporal' smoothing on the cepstral time series. In contrast, the presented WS-HEQ lowers HPF (the high-pass filtered portion) of each cepstral vector and is analogous to a 'spatial' smoothing operation. Such an observation leads to the idea of combining either MVA or TSN with WS-HEQ in order to achieve a two-dimensional smoothing. To realize this idea, the cepstra are first processed with any of the eight forms of WS-HEQ and then further compensated by MVA or TSN. The obtained recognition results are shown in Tables 5 and 6, in which the applied WS-HEQ uses the scaling factor α listed in Table 1. As we look into the results shown in Tables 5 and 6, it can be found that the pairing of WS-HEQ and MVA/TSN consistently achieves better performance than the individual component method, regardless of the various forms of WS-HEQ. For example, the method 'WS-HEQ$_{\text{II}}^{(1)}$+TSN' obtains the averaged accuracy of 86.25%, better than TSN (81.39%) and

Table 5 The recognition accuracy results (%) achieved by the combination of MVA and WS-HEQ

Method	Set A	Set B	Set C	Average
MVA	78.15	79.17	79.12	78.75
WS-HEQ$_{\text{I}}^{(1)}$	83.36	85.37	83.89	84.27
WS-HEQ$_{\text{I}}^{(1)}$+MVA	84.95	85.60	84.97	85.21
WS-HEQ$_{\text{I}}^{(2)}$	82.29	83.22	82.82	82.76
WS-HEQ$_{\text{I}}^{(2)}$+MVA	85.34	85.89	85.69	85.63
WS-HEQ$_{\text{I}}^{(3)}$	83.57	85.15	83.93	84.27
WS-HEQ$_{\text{I}}^{(3)}$+MVA	84.84	85.59	85.38	85.25
WS-HEQ$_{\text{I}}^{(4)}$	82.88	84.70	82.78	83.59
WS-HEQ$_{\text{I}}^{(4)}$+MVA	84.62	85.70	84.90	85.11
WS-HEQ$_{\text{II}}^{(1)}$	84.13	86.16	84.39	84.99
WS-HEQ$_{\text{II}}^{(1)}$+MVA	85.08	86.43	85.40	85.69
WS-HEQ$_{\text{II}}^{(2)}$	83.54	85.75	83.83	84.48
WS-HEQ$_{\text{II}}^{(2)}$+MVA	85.81	86.69	86.17	86.23
WS-HEQ$_{\text{II}}^{(3)}$	83.25	85.10	83.50	84.04
WS-HEQ$_{\text{II}}^{(3)}$+MVA	84.18	85.69	84.48	84.84
WS-HEQ$_{\text{II}}^{(4)}$	82.30	83.45	82.90	82.88
WS-HEQ$_{\text{II}}^{(4)}$+MVA	85.14	85.41	85.87	85.40

They are for different test sets while averaged over five SNR conditions (20 to 0 dB) as to the Aurora-2 database. The scaling factor α listed in Table 1 is adopted for each WS-HEQ.

Table 6 The recognition accuracy results (%) achieved by the combination of TSN and WS-HEQ

Method	Set A	Set B	Set C	Average
TSN	80.86	82.39	80.46	81.39
WS-HEQ$_{\text{I}}^{(1)}$	83.36	85.37	83.89	84.27
WS-HEQ$_{\text{I}}^{(1)}$+TSN	85.32	86.77	85.01	85.84
WS-HEQ$_{\text{I}}^{(2)}$	82.29	83.22	82.82	82.76
WS-HEQ$_{\text{I}}^{(2)}$+TSN	85.06	86.21	84.75	85.46
WS-HEQ$_{\text{I}}^{(3)}$	83.57	85.15	83.93	84.27
WS-HEQ$_{\text{I}}^{(3)}$+TSN	85.35	86.54	85.15	85.78
WS-HEQ$_{\text{I}}^{(4)}$	82.88	84.70	82.78	83.59
WS-HEQ$_{\text{I}}^{(4)}$+TSN	84.41	85.81	83.96	84.88
WS-HEQ$_{\text{II}}^{(1)}$	84.13	86.16	84.39	84.99
WS-HEQ$_{\text{II}}^{(1)}$+TSN	85.59	87.41	85.26	86.25
WS-HEQ$_{\text{II}}^{(2)}$	83.54	85.75	83.83	84.48
WS-HEQ$_{\text{II}}^{(2)}$+TSN	85.77	87.08	85.34	86.21
WS-HEQ$_{\text{II}}^{(3)}$	83.25	85.10	83.50	84.04
WS-HEQ$_{\text{II}}^{(3)}$+TSN	84.68	86.75	84.25	85.42
WS-HEQ$_{\text{II}}^{(4)}$	82.30	83.45	82.90	82.88
WS-HEQ$_{\text{II}}^{(4)}$+TSN	84.59	85.32	84.59	84.88

They are for different test sets while averaged over five SNR conditions (20 to 0 dB) as to the Aurora-2 database. The scaling factor α listed in Table 1 is adopted for each WS-HEQ.

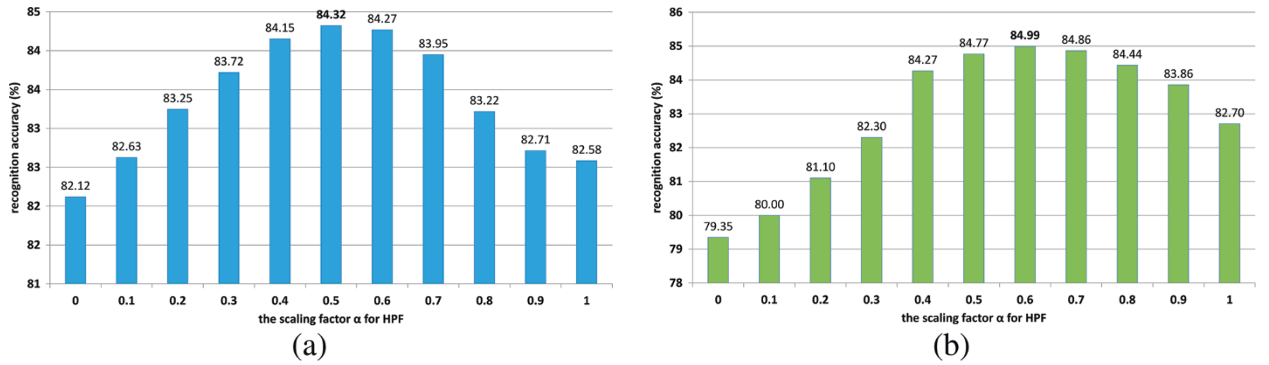


Figure 8 The overall recognition accuracy achieved by two WS-HEQ methods with different α. (a) WS-HEQ$_{\mathrm{I}}^{(1)}$. **(b)** WS-HEQ$_{\mathrm{II}}^{(1)}$.

WS-HEQ$_{\mathrm{II}}^{(1)}$ (84.99%). These results indicate that the joint spatial-temporal smoothing can provide the cepstral features with better noise robustness in comparison with either spatial smoothing or temporal smoothing in isolation. In particular, different forms of WS-HEQ behave very similar and can give around 85% in averaged accuracy when TSN/MVA is integrated, implying that when employing TSN/MVA as a post-processing technique, simpler versions of WS-HEQ, such as WS-HEQ$_{\mathrm{I}}^{(4)}$ and WS-HEQ$_{\mathrm{II}}^{(4)}$, are relatively more appropriate in practical applications due to their high recognition performance and relatively low computation complexity in comparison with S-HEQ, WS-HEQ$_{\mathrm{I}}^{(1)}$, and WS-HEQ$_{\mathrm{II}}^{(1)}$.

5.2 The influence of the parameter α in WS-HEQ

As stated previously, the parameter α in WS-HEQ determines the degree of attenuation for the HPF portion of the processed cepstra. Here, we would like to investigate how the value of α in WS-HEQ influences the recognition accuracy of the test sets. For simplicity, we vary the parameter α from 0.0 to 1.0 in two types of WS-HEQ: WS-HEQ$_{\mathrm{I}}^{(1)}$ and WS-HEQ$_{\mathrm{II}}^{(1)}$, and the corresponding recognition accuracy rates averaged over all noise types and levels in three test sets are shown in Figure 8a,b. These two figures reveal that

1. Lowering the HPF part by tuning α from 1.0 to 0.4 in both WS-HEQ$_{\mathrm{I}}^{(1)}$ and WS-HEQ$_{\mathrm{II}}^{(1)}$ achieves better results consistently relative to these two WS-HEQ methods using $\alpha = 1.0$. However, further reducing the HPF part can ruin the recognition accuracy, which implies that the HPF part also contains information helpful for recognition.
2. The optimal accuracy for WS-HEQ$_{\mathrm{I}}^{(1)}$ and WS-HEQ$_{\mathrm{II}}^{(1)}$ occurs when α is assigned to 0.5 and 0.6, respectively, while the results from the development

Table 7 Recognition accuracy results (%) of WS-HEQ$_{\mathrm{I}}^{(1)}$ using the optimal scaling factor α (in parentheses)

SNR	Set A				Set B				Set C	
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train	MIRS subway	MIRS street
Clean	99.14 (0.5)	99.06 (0.5)	98.96 (0.9)	99.04 (0.9)	99.14 (0.5)	99.06 (0.5)	98.96 (0.9)	99.04 (0.9)	99.14 (0.6)	99.21 (1.0)
20 dB	96.87 (0.4)	97.61 (0.8)	98.51 (1.0)	97.41 (0.5)	97.73 (0.7)	98.07 (0.5)	98.27 (0.7)	97.90 (0.6)	97.33 (0.5)	97.91 (0.5)
15 dB	94.69 (0.4)	95.77 (0.6)	96.87 (0.6)	94.79 (0.5)	95.70 (0.7)	96.46 (0.5)	96.99 (1.0)	96.51 (0.5)	94.87 (0.4)	96.43 (0.7)
10 dB	89.84 (0.4)	91.38 (0.6)	92.93 (0.7)	89.57 (0.4)	91.74 (0.8)	92.53 (0.6)	93.89 (0.8)	93.58 (0.6)	90.24 (0.4)	92.08 (0.5)
5 dB	79.80 (0.3)	79.66 (0.6)	82.11 (0.5)	78.90 (0.4)	80.32 (0.7)	81.17 (0.5)	84.61 (0.7)	82.91 (0.5)	80.29 (0.2)	81.17 (0.5)
0 dB	57.69 (0.3)	50.00 (0.6)	55.41 (0.7)	57.42 (0.4)	55.30 (0.7)	56.77 (0.5)	62.69 (0.7)	58.04 (0.7)	58.12 (0.3)	56.77 (0.4)
−5 dB	26.74 (0.4)	20.16 (1.0)	20.16 (0.0)	28.85 (0.0)	23.76 (1.0)	24.76 (0.4)	27.65 (1.0)	21.97 (1.0)	26.71 (0.1)	24.15 (0.2)

This is with respect to each noise type and level (SNR) as to the Aurora-2 database.

Table 8 The recognition accuracy results (%) of WS-HEQ$_{\mathrm{II}}^{(1)}$ using the optimal scaling factor α (in parentheses)

SNR	Set A				Set B				Set C	
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train	MIRS subway	MIRS street
Clean	99.36 (0.7)	99.06 (0.8)	99.05 (0.8)	99.11 (0.4)	99.36 (0.7)	99.06 (0.8)	99.05 (0.8)	99.11 (0.4)	99.17 (0.7)	99.15 (0.4)
20 dB	96.56 (0.7)	97.52 (0.6)	98.12 (0.8)	96.98 (0.8)	97.54 (0.7)	97.94 (0.8)	98.42 (0.5)	98.36 (0.8)	96.96 (0.4)	97.76 (0.8)
15 dB	94.60 (0.4)	95.71 (0.7)	97.17 (0.7)	94.60 (0.5)	95.52 (0.6)	96.55 (0.6)	96.90 (0.6)	97.25 (0.8)	94.41 (0.6)	96.28 (0.6)
10 dB	89.65 (0.5)	92.14 (0.6)	93.83 (0.7)	89.48 (0.5)	91.83 (0.7)	93.05 (0.6)	94.72 (0.8)	94.32 (0.6)	89.75 (0.4)	92.62 (0.6)
5 dB	79.89 (0.5)	80.11 (0.6)	84.28 (0.8)	78.22 (0.5)	80.07 (0.7)	82.56 (0.5)	85.51 (0.8)	84.33 (0.6)	79.83 (0.4)	81.65 (0.6)
0 dB	58.27 (0.4)	53.42 (0.7)	61.65 (0.7)	57.27 (0.4)	56.77 (0.7)	58.92 (0.5)	65.49 (0.7)	61.65 (0.8)	57.72 (0.4)	58.74 (0.6)
−5 dB	28.43 (0.5)	22.76 (0.8)	28.81 (0.9)	30.48 (0.4)	25.85 (0.9)	27.33 (0.6)	31.26 (0.9)	29.28 (0.9)	29.60 (0.5)	27.42 (0.8)

This is with respect to each noise type and level (SNR) as to the Aurora-2 database.

set suggest the parameter α to be 0.6 for these two WS-HEQ methods (as shown in Table 1). However, WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 0.6$ gives the recognition rate of 84.27%, very close to the optimal one (84.32%). Therefore, it assures us that the development set can help to determine the nearly optimal parameter in the test sets.

3. The performance of WS-HEQ$_{\mathrm{I}}^{(1)}$ and WS-HEQ$_{\mathrm{II}}^{(1)}$ is not very sensitive to the parameter α, which is based on the observation that the accuracy difference is below 1.0% provided the value of α is within the range $[0.4, 0.7]$.

Next, we explore the best possible recognition results for each testing situation achieved by WS-HEQ with various assignments of the scaling parameter α. Please note that, in the preceding experiments the scaling parameter α in WS-HEQ is determined by the development set and then uniformly applied to the every test set. Here, we would like to investigate whether the optimal choice of α (which gives rise to the highest recognition accuracy) depends on the noise type and level (*viz.* the SNR) of the testing utterances. To do this, we vary the value of α from 0.0 to 1.0 with an interval of 0.1 in each form of WS-HEQ to process the features in the training and testing sets and then perform the experiment. The optimal recognition accuracy rate and the associated α with respect to each noise type and level in the testing set achieved by WS-HEQ$_{\mathrm{I}}^{(1)}$ and WS-HEQ$_{\mathrm{II}}^{(1)}$ are respectively shown in Tables 7 and 8. Some contents of the tables together with the data obtained from the other six forms of WS-HEQ (which are not listed here due to their huge amount) are further summarized in

Table 9 The recognition accuracy results (%) of various forms of WS-HEQ for different test sets

Method		Set A	Set B	Set C	Average
WS-HEQ$_{\mathrm{I}}^{(1)}$	$\alpha = 0.6$	83.36	85.37	83.89	84.27
	Optimal α	83.86	85.56	84.52	84.67
WS-HEQ$_{\mathrm{I}}^{(2)}$	$\alpha = 0.6$	82.29	83.22	82.82	82.76
	Optimal α	83.04	84.08	83.32	83.51
WS-HEQ$_{\mathrm{I}}^{(3)}$	$\alpha = 0.5$	83.57	85.15	83.93	84.27
	Optimal α	83.86	85.46	84.43	84.62
WS-HEQ$_{\mathrm{I}}^{(4)}$	$\alpha = 0.7$	82.88	84.70	82.78	83.59
	Optimal α	83.52	84.86	83.85	84.12
WS-HEQ$_{\mathrm{II}}^{(1)}$	$\alpha = 0.6$	84.13	86.16	84.39	84.99
	Optimal α	84.47	86.39	84.57	85.26
WS-HEQ$_{\mathrm{II}}^{(2)}$	$\alpha = 0.6$	83.54	85.75	83.83	84.48
	Optimal α	84.25	86.04	84.52	85.02
WS-HEQ$_{\mathrm{II}}^{(3)}$	$\alpha = 0.7$	83.25	85.10	83.50	84.04
	Optimal α	83.84	85.85	84.05	84.69
WS-HEQ$_{\mathrm{II}}^{(4)}$	$\alpha = 0.6$	82.30	83.45	82.90	82.88
	Optimal α	82.92	84.18	83.17	83.47

These results are obtained by using (1) the scaling factor α listed in Table 1 (2) the scaling factor α that achieves the optimal recognition accuracy with respect to the individual noise type and level (SNR), both of which are for different Test Sets while averaged over 5 SNR conditions (20 dB to 0 dB) as to the Aurora-2 database.

Tables 9 and 10, which also contain a portion of the data in Tables 2 and 3 for the purpose of comparison. Observing these tables, we find that the value of the factor α that achieves the optimal recognition accuracy indeed depends on the noise type and level of the utterances. However, there seems no general rule for selecting a better α with respect to any specific noise situation. Furthermore, as seen in Table 9, in most cases, the accuracy rates obtained with the optimal α associated with the individual noise situation are very close to the accuracy rates using a fixed α which gets the optimal results for the development set. The maximum difference between the above two types of accuracy rates is 0.75%, which occurs at the method of WS-HEQ$_{\mathrm{I}}^{(2)}$. As a result, we can roughly conclude that using the α recommended by the development set suffices to provide WS-HEQ with nearly optimal performance.

5.3 The feature distortion reduced by WS-HEQ

Apart from the recognition performance, in this subsection, we evaluate WS-HEQ in the capacity of reducing the feature distortion caused by noise. The incoherent feature distortion [7] defined by

$$\varphi = \frac{\Sigma_k(|X[k]| - |\tilde{X}[k]|)^2}{\Sigma_k|\tilde{X}[k]|^2} \tag{16}$$

is measured for the feature streams processed by the noise-robustness method, where $\tilde{X}[k]$ and $X[k]$ denote the DFT of the noise-free clean feature stream and its noise-corrupted counterpart, respectively. Figure 9 depicts the feature distortion associated with any cepstral channel at the SNR of 10 dB, averaged over the 1,001 utterances in test set A of the Aurora-2 database, with respect to the feature streams processed by any of S-HEQ, WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 0.6$, and WS-HEQ$_{\mathrm{II}}^{(1)}$ with $\alpha = 0.6$. From Figure 9, two observations are made: first, WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 0.6$ results in smaller distortions than S-HEQ irrespective of the cepstral channel, implying that to lower the HPF portion of the cepstra can further reduce the

Table 10 The recognition accuracy results (%) of various forms of WS-HEQ at different SNRs

Method		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
WS-HEQ$_{\mathrm{I}}^{(1)}$	$\alpha = 0.6$	98.99	97.60	95.72	91.54	80.62	55.87	23.26
	Optimal α	99.08	97.76	95.91	91.78	81.09	56.82	24.29
WS-HEQ$_{\mathrm{I}}^{(2)}$	$\alpha = 0.6$	99.09	97.68	95.64	91.30	79.23	49.97	17.69
	Optimal α	99.10	97.86	95.94	91.66	79.87	52.22	19.62
WS-HEQ$_{\mathrm{I}}^{(3)}$	$\alpha = 0.5$	98.99	97.51	95.71	91.59	80.59	55.95	23.88
	Optimal α	99.08	97.67	95.88	91.81	81.03	56.69	24.64
WS-HEQ$_{\mathrm{I}}^{(4)}$	$\alpha = 0.7$	99.07	97.87	96.18	91.98	79.99	51.94	19.83
	Optimal α	99.10	97.87	96.23	92.14	80.51	53.86	23.00
WS-HEQ$_{\mathrm{II}}^{(1)}$	$\alpha = 0.6$	99.05	97.53	95.75	91.99	81.35	58.35	26.98
	Optimal α	99.15	97.62	95.90	92.14	81.65	58.99	28.12
WS-HEQ$_{\mathrm{II}}^{(2)}$	$\alpha = 0.6$	99.04	97.60	95.70	91.68	80.83	56.61	24.45
	Optimal α	99.13	97.80	96.09	92.15	81.44	56.68	24.40
WS-HEQ$_{\mathrm{II}}^{(3)}$	$\alpha = 0.7$	99.05	97.87	95.98	91.86	80.62	53.88	20.51
	Optimal α	99.13	97.51	95.68	91.66	80.94	57.64	27.09
WS-HEQ$_{\mathrm{II}}^{(4)}$	$\alpha = 0.6$	98.96	97.49	95.54	91.17	79.00	51.23	19.37
	Optimal α	99.08	97.58	95.69	91.30	79.41	53.39	21.84

These results are obtained using (1) the scaling factor α listed in Table 1 (2) the scaling factor α that achieves the optimal recognition accuracy with respect to the individual noise type and level (SNR), both of which are for different SNR conditions while averaged over ten noise types as to the Aurora-2 database.

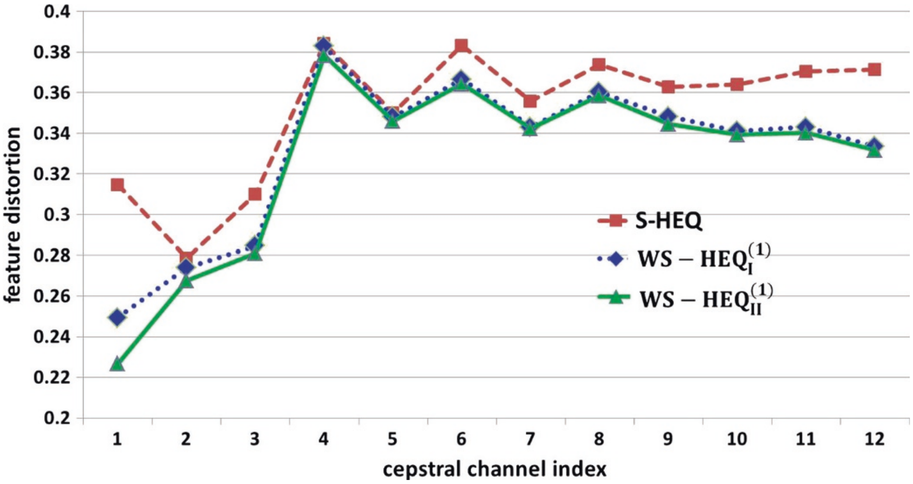


Figure 9 Feature distortion averaged over the 1,001 utterances of test set A. It is achieved by S-HEQ, WS-HEQ$_{\mathrm{I}}^{(1)}$ ($\alpha = 0.6$), and WS-HEQ$_{\mathrm{II}}^{(1)}$ ($\alpha = 0.6$). The DFT size used in Equation 16 is set to 512.

effect of noise; second, by setting the parameter α to 0.6, the distortions provided by WS-HEQ$_{\mathrm{II}}^{(1)}$ are slightly smaller than those by WS-HEQ$_{\mathrm{I}}^{(1)}$ for most of the cepstral channels, which agrees with the finding that WS-HEQ$_{\mathrm{II}}^{(1)}$ slightly outperforms WS-HEQ$_{\mathrm{I}}^{(1)}$ in recognition accuracy.

5.4 The effect of lowering HPF in different schemes

In order to further examine the effect of attenuating HPF in recognition accuracy, here we additionally design three schemes to process the cepstra in training and testing sets:

Scheme 1. The original (full-band) cepstra is split into LPF and HPF, and then the HPF portion is scaled by a factor α. Finally, the original LPF and the attenuated HPF are combined to constitute the new cepstra. This scheme is to remove all three HEQ processes in WS-HEQ$_{\mathrm{I}}^{(1)}$ or WS-HEQ$_{\mathrm{II}}^{(1)}$ shown in Figure 6a,b, and its flowchart is depicted in Figure 10a.

Scheme 2. The original (full-band) cepstra is preprocessed by HEQ, and then the HEQ-preprocessed cepstra is split into LPF and HPF. We scale LPF by a factor α and finally combine LPF and attenuated HPF to obtain the new cepstra. This scheme is to remove the two HEQ processes for LPF and HPF in WS-HEQ$_{\mathrm{I}}^{(1)}$ shown in Figure 6a, and its flowchart is depicted in Figure 10b.

Scheme 3. The original cepstra is split into LPF and HPF, and then the HPF portion is tuned with the scaling factor α. Next, we combine LPF and attenuated HPF to obtain the full-band cepstra, which are further post-processed by HEQ to obtain the new cepstra. This scheme is to remove the two HEQ processes for LPF and HPF in WS-HEQ$_{\mathrm{II}}^{(1)}$ shown in Figure 6b, and its flowchart is depicted in Figure 10c.

The scaling factor α in the above three schemes is varied from 0.6 to 1.4 with an interval of 0.2. Please note that the case with $\alpha > 1$ corresponds to amplifying HPF and thus reducing the proportion of LPF in the overall cepstra. The recognition results for the three schemes are

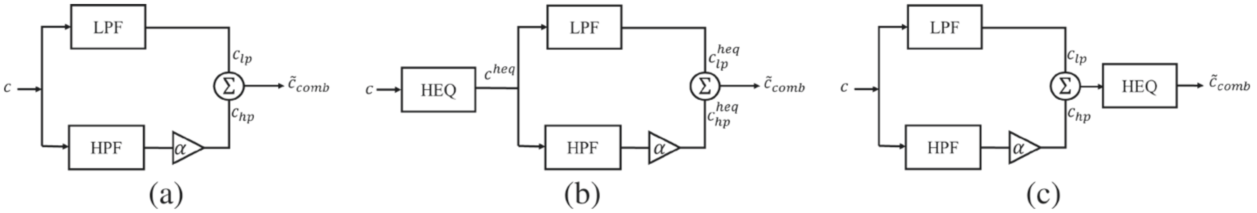


Figure 10 Flowcharts of three schemes defined in defined in Section 5.4. (a) Scheme 1. **(b)** Scheme 2. **(c)** Scheme 3.

shown in Table 11. From this table, we have the following observations:

1. At the clean noise-free case in all three schemes, the recognition accuracy remains as high as around 99% nearly irrespective of the varied scaling factor α, which implies that neither lowering nor raising the HPF portion of the cepstra can significantly influence the recognition performance. The possible explanation for this result is that the back-end acoustic modeling with HMMs compensates well for the variation of the front-end speech features.
2. From the results for scheme 1, reducing HPF (using $\alpha < 1$) without pre- or post-processing with HEQ produces degraded performance under noise-corrupted situations compared with the case using $\alpha = 1$, which disagrees with the results for various forms of WS-HEQ as shown in the preceding sub-sections. Under the same situations, setting $\alpha > 1$ to amplify HPF (and thus to reduce the proportion of LPF) cannot improve the accuracy, either. Therefore, the relative importance of LPF and HPF in noise-corrupted cepstra discussed in Section 3 cannot be reflected in recognition accuracy when there is no noise-robust processing such as HEQ. In other words, merely emphasizing LPF or HPF fails to result in more noise-robust cepstra and produces worse recognition accuracy.
3. Different from the results for scheme 1, the results associated with schemes 2 and 3 show that when the cepstra are pre- or post-processed by HEQ, reducing the HPF part by setting $\alpha < 1$ can promote the recognition accuracy under noise-corrupted situations (except for the case of -5-dB SNR). On the other hand, the cases corresponding to $\alpha > 1$ in which HPF is raised produce worse results. The underlying reason is probably that the noise effect of HPF is relatively difficult to alleviate, and simply lowering HPF can benefit HEQ to give better performance. Similar situations can be also found in Tables 2 and 3 by comparing the results of WS-HEQ$_{\mathrm{I}}^{(2)}$, WS-HEQ$_{\mathrm{I}}^{(3)}$, WS-HEQ$_{\mathrm{II}}^{(2)}$ and WS-HEQ$_{\mathrm{II}}^{(3)}$ with $\alpha = 1$. WS-HEQ$_{\mathrm{I}}^{(2)}$ and WS-HEQ$_{\mathrm{II}}^{(2)}$ outperform WS-HEQ$_{\mathrm{I}}^{(3)}$ and WS-HEQ$_{\mathrm{II}}^{(3)}$, respectively, indicating a stronger normalization strategy like HEQ is required to compensate the distortion in HPF, while a relatively simple MVN process suffices to improve LPF well. Furthermore, comparing Table 11 with Tables 2 and 3 we find that the effect of lowering HPF in recognition accuracy appears a lot more significant when we further compensate the sub-band cepstra (*viz.* HPF and LPF) by HEQ/MVN, again in agreement with the statements about S-HEQ [11] that additionally normalizing HPF and LPF can reduce the environmental mismatch caused by noise.

6 The experiment on the TCC-300 Mandarin dataset

Besides the evaluation on the Aurora-2 dataset as described in the previous two sections, here the recognition experiments with the presented WS-HEQ are further carried out in another dataset, the eleventh group of the TCC-300 microphone speech database from the Association for Computational Linguistics and Chinese Language Processing in Taiwan [15]. This dataset includes 7,009 Mandarin character strings uttered by 50 male and 50 female adult speakers. The corresponding read

Table 11 The recognition accuracy results (%) of the three schemes defined in Section 5.4

	SNR	The scaling factor α				
		0.6	**0.8**	**1.0**	**1.2**	**1.4**
Scheme 1 (MFCC)	Clean	99.08	99.16	99.12	99.14	99.09
	$20 \sim 0$ dB	52.84	59.15	59.75	59.42	56.48
	-5 dB	6.13	7.67	8.16	7.70	6.87
Scheme 2 (pre-HEQ)	clean	98.95	99.00	98.97	98.92	99.00
	$20 \sim 0$ dB	81.84	81.41	80.32	80.16	80.04
	-5 dB	17.58	17.27	19.94	15.12	14.39
Scheme 3 (post-HEQ)	clean	98.97	99.00	98.97	98.99	98.84
	$20 \sim 0$ dB	81.19	80.42	80.32	79.72	79.35
	-5 dB	16.73	14.62	19.94	14.05	18.34

They are for the clean condition, the average of five SNR conditions (20, 15, 10, 5, and 0 dB), and the -5 dB SNR condition but averaged over the ten noise types as to the Aurora-2 database.

speaking-style speech signals were recorded with a microphone at the sampling rate of 16 kHz. The Mandarin characters included in the utterances of this dataset correspond to 408 different Mandarin syllables. In the experiment, the syllable recognition is performed on this dataset without any language model or grammar constraint at the back end so that the recognition performance can be more related to the used front-end acoustic features. As a result, in comparison with the 11-digit recognition on the Aurora-2 telephone-band dataset in the previous sections, here we conduct a more complicated task of medium-vocabulary recognition (408 syllables) on the broad-band speech data. Among the 7,009 Mandarin utterances in the TCC-300 subset, 6,509 strings are selected in acoustic model training, while the other 500 are in testing. The utterances in the training set are kept noise-free, while the utterances in the testing set are artificially added with noise at four SNR levels (20, 15, 5, and 0 dB) to produce noise-corrupted speech data. The noise types include white (broad-band) and pink (narrow-band), both taken from the NOISEX 92 database [17]. These utterances for training and testing are first converted into 13-dimensional MFCCs ($c0$, $c1 - c12$), and then processed by various kinds of noise-robustness algorithms. Similar to the feature parameter settings for the Aurora-2 database, the resulting 13 new features plus their first- and second-order derivatives constitute the finally used 39-dimensional feature vector.

As for the acoustic modeling, we train the HMMs of INITIAL and FINAL units, which corresponds to the semi-syllables in Mandarin Chinese. In most cases, a Mandarin Chinese syllable can be split into INITIAL/FINAL parts analogous to the consonant/vowel pair in English. There are totally 112 right-context-dependent INITIAL HMMs and 38 context-independent FINAL HMMs to be trained. Each INITIAL HMM consists of five states and eight Gaussian mixtures per state, while each FINAL HMM contains ten states and eight

Table 12 Recognition accuracy results (%) of WS-HEQ for different SNR conditions at *white* noise environment

Method		Clean	20 dB	15 dB	10 dB	5 dB	Average
MFCC		76.38	30.08	14.72	6.23	3.52	13.64
CHN		76.52	55.22	43.56	30.90	18.84	37.13
WS-HEQ$_{\text{I}}^{(1)}$	$\alpha = 1.0$	77.15	56.48	46.81	33.05	19.40	38.94
	$\alpha = 0.6$	76.94	59.28	49.81	36.80	22.74	42.16
WS-HEQ$_{\text{I}}^{(2)}$	$\alpha = 1.0$	77.50	56.30	47.62	33.75	19.82	39.37
	$\alpha = 0.6$	76.07	59.63	49.88	36.73	23.74	42.50
WS-HEQ$_{\text{I}}^{(3)}$	$\alpha = 1.0$	77.08	55.92	44.94	32.09	19.36	38.08
	$\alpha = 0.6$	76.84	58.68	48.44	36.22	23.11	41.61
WS-HEQ$_{\text{I}}^{(4)}$	$\alpha = 1.0$	76.91	54.71	45.29	32.02	19.22	37.81
	$\alpha = 0.6$	76.59	58.14	48.95	35.89	23.11	41.52
WS-HEQ$_{\text{II}}^{(1)}$	$\alpha = 1.0$	77.96	56.88	47.46	33.82	20.94	39.78
	$\alpha = 0.6$	76.54	59.75	49.79	36.89	23.72	42.54
WS-HEQ$_{\text{II}}^{(2)}$	$\alpha = 1.0$	77.31	57.79	47.83	33.91	20.85	40.10
	$\alpha = 0.6$	76.33	59.86	49.72	37.24	24.37	42.80
WS-HEQ$_{\text{II}}^{(3)}$	$\alpha = 1.0$	77.92	57.18	46.36	33.70	20.64	39.47
	$\alpha = 0.6$	77.01	59.98	49.11	36.26	23.53	42.22
WS-HEQ$_{\text{II}}^{(4)}$	$\alpha = 1.0$	77.10	57.11	46.46	34.40	20.90	39.72
	$\alpha = 0.6$	77.24	60.07	49.91	36.85	23.65	42.62

These recognition accuracy results (%) of the MFCC baseline, CHN, and eight forms of WS-HEQ are for different SNR conditions at the white noise environment as to the subset of the TCC database.

Gaussian mixtures per state. The HMM for each of the 408 Mandarin syllables is then constructed by concatenating the associated INITIAL and FINAL HMMs.

Tables 12 and 13 list the syllable recognition accuracy rates of the MFCC baseline and the various robustness methods including CHN, S-HEQ (equivalent to WS-HEQ$_{\mathrm{I}}^{(1)}$ with $\alpha = 1.0$), and seven forms of the presented WS-HEQ for the white and pink noise environments, respectively. The scaling parameter α in WS-HEQ is set to 0.6, which is not optimized but just to clarify whether lowering HPF can give rise to performance improvement. From these two tables, we have the following findings:

1. Due to the simple free-syllable decoding framework in the recognition procedure, the recognition accuracy of MFCC baseline features at the clean noise-free condition is just around 75%. Besides, the noise robustness methods used here result in similar or even better performance compared with the MFCC baseline when the testing utterances contain no noise.
2. Both types of noise degrade the performance of MFCC seriously as the SNR gets worse, while CHN and all of the other HEQ-related algorithms benefit the recognition accuracy significantly. In particular, the various forms of WS-HEQ with $\alpha = 1$ outperforms CHN, indicating that additionally processing LPF and HPF with HEQ or MVN can further enhance CHN to produce better results.
3. Reducing the scaling factor α from 1.0 to 0.6 in the eight forms of WS-HEQ consistently brings about better results by significant margins in all noise-corrupted situations. This result reconfirms the capability of the presented HPF lowering operation in boosting noise robustness of CHN-processed features. Furthermore, when α is set to 0.6, the performance difference among various

Table 13 Recognition accuracy results (%) of WS-HEQ for different SNR condition at the *pink* noise environment

Method		Clean	20 dB	15 dB	10 dB	5 dB	Average
MFCC		76.38	59.44	44.24	22.34	5.85	32.97
CHN		76.52	61.40	52.71	38.06	24.30	44.12
WS-HEQ$_{\mathrm{I}}^{(1)}$	$\alpha = 1.0$	77.15	62.83	53.96	39.62	25.09	45.38
	$\alpha = 0.6$	76.94	63.48	54.99	40.76	27.10	46.58
WS-HEQ$_{\mathrm{I}}^{(2)}$	$\alpha = 1.0$	77.50	63.62	54.45	39.79	25.91	45.94
	$\alpha = 0.6$	76.07	63.11	55.22	40.35	26.94	46.41
WS-HEQ$_{\mathrm{I}}^{(3)}$	$\alpha = 1.0$	77.08	61.38	51.89	38.06	24.44	43.94
	$\alpha = 0.6$	76.84	62.55	54.34	40.60	26.61	46.03
WS-HEQ$_{\mathrm{I}}^{(4)}$	$\alpha = 1.0$	76.91	62.38	53.15	38.43	23.81	44.44
	$\alpha = 0.6$	76.59	63.15	54.50	40.14	26.14	45.98
WS-HEQ$_{\mathrm{II}}^{(1)}$	$\alpha = 1.0$	77.96	63.04	54.15	39.83	25.63	45.66
	$\alpha = 0.6$	76.54	63.62	55.27	41.51	26.84	46.81
WS-HEQ$_{\mathrm{II}}^{(2)}$	$\alpha = 1.0$	77.31	63.27	54.24	39.65	25.75	45.73
	$\alpha = 0.6$	76.33	62.62	55.29	40.81	26.82	46.39
WS-HEQ$_{\mathrm{II}}^{(3)}$	$\alpha = 1.0$	77.92	62.50	53.19	39.86	24.98	45.13
	$\alpha = 0.6$	77.01	63.48	54.90	41.02	27.08	46.62
WS-HEQ$_{\mathrm{II}}^{(4)}$	$\alpha = 1.0$	77.10	62.55	52.80	38.18	24.91	44.61
	$\alpha = 0.6$	77.24	63.13	54.52	40.88	26.91	46.36

These recognition accuracy results (%) of the MFCC baseline, CHN, and eight forms of WS-HEQ are for different SNR conditions at the pink noise environment as to the subset of the TCC database.

forms of WS-HEQ becomes relatively small in comparison with that under the condition of $\alpha = 1.0$.

7 Conclusions

In this paper, we explored the relative importance of different frequency components of the intra-frame speech features and subsequently presented a novel algorithm, WS-HEQ, to improve noisy speech recognition. WS-HEQ mainly reduces the intra-frame high-pass filtered component of the speech features, which appears more vulnerable to noise. Compared with the well-known S-HEQ method, WS-HEQ can achieve superior recognition accuracy, higher computational efficiency, or both. In future work, we will pursue new filter structures for obtaining the LPF and HPF components for WS-HEQ to achieve better results. Additionally, we will investigate how to tune the intra-frame speech features more flexibly in the corresponding DFT or DCT domains for further noise reduction.

Competing interests

The authors declare that they have no competing interests.

Received: 1 August 2013 Accepted: 27 November 2013
Published: 23 December 2013

References

1. HK Maganti, M Matassoni, A perceptual masking approach for noise robust speech recognition. EURASIP J. Audio Speech Music Process. **2012**(29) (2012)
2. K Wu, C Chen, B Yeh, Noise-robust speech feature processing with empirical mode decomposition. EURASIP J. Audio Speech Music Process. **2011**(9) (2011)
3. I Cohen, B Berdugo, Speech enhancement for non-stationary noise environments. Signal Process. **81**(11), 2403–2418 (2001)
4. B Kotnik, Z Kačič, A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and AR modeling of speech signals. Signal Process. **87**(6), 1202–1223 (2007)
5. S Tibrewala, H Hermansky, Multi-band and adaptation approaches to robust speech recognition, in *5th Eurospeech Conference on Speech Communications and Technology* (Eurospeech, Rhodes, 22–25 Sept 1997)
6. F Hilger, H Ney, Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. Audio Lang. Process. **14**, 845–854 (2006)
7. J Benesty, MM Sondhi, Y Huang (eds.), *Springer Handbook of Speech Processing* (Springer, Heidelberg, 2008)
8. C Chen, J Bilmes, MVA processing of speech features. IEEE Trans. Audio Speech Lang. Process. **15**, 257–270 (2007)
9. C-W Hsu, L-S Lee, Higher order cepstral moment normalization for improved robust speech recognition. IEEE Trans. Audio Speech Lang. Process. **17**, 205–220 (2009)
10. X Xiao, ES Chng, H Li, Normalization of the speech modulation spectra for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. **16**, 1662–1674 (2008)
11. V Joshi, R Bilgi, S Umesh, L García, MC Benítez, Sub-band level histogram equalization for robust speech recognition, in *12th International Conference on Spoken Language Processing* (Interspeech, Florence, 27–31 Sept 2011)
12. HG Hirsch, D Pearce, The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millenium* (ISCA ITRW ASR, Paris, 18–20 Sept 2000)
13. N Kanedera, T Arai, H Hermansky, M Pavel, On the importance of various modulation frequencies for speech recognition, in *5th European Conference on Speech Communication and Technology* (Eurospeech, Rhodes, 22–25 Sept 1997)
14. X Huang, A Acero, H-W Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development* (Prentice Hall, New Jersey, 2001)
15. ACLCLP (1990). [http://www.aclclp.org.tw/corp.php], Accessed 10 Aug 2013
16. ELDA (1995). [http://www.elda.org/article52.html], Accessed 8 Aug 2013
17. AP Varga, HJM Steeneken, M Tomlinson, D Jones, The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit (1992)

doi:10.1186/1687-4722-2013-29
Cite this article as: Hung and Fan: **Intra-frame cepstral sub-band weighting and histogram equalization for noise-robust speech recognition.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:29.