EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Enhancing the magnitude spectrum of speech features for robust speech recognition

Jeih-weih Hung*, Hao-teng Fan and Wen-hsiang Tu

Abstract

In this article, we present an effective compensation scheme to improve noise robustness for the spectra of speech signals. In this compensation scheme, called magnitude spectrum enhancement (MSE), a voice activity detection (VAD) process is performed on the frame sequence of the utterance. The magnitude spectra of non-speech frames are then reduced while those of speech frames are amplified. In experiments conducted on the Aurora-2 noisy digits database, MSE achieves an error reduction rate of nearly 42% relative to baseline processing. This method outperforms well-known spectral-domain speech enhancement techniques, including spectral subtraction (SS) and Wiener filtering (WF). In addition, the proposed MSE can be integrated with cepstral-domain robustness methods, such as mean and variance normalization (MVN) and histogram normalization (HEQ), to achieve further improvements in recognition accuracy under noise-corrupted environments.

Keywords: Voice activity detection, Robust speech recognition, Speech enhancement

Introduction

The environmental mismatch caused by additive noise and/or channel distortion often seriously degrades the performance of speech recognition systems. Various robustness techniques have been proposed to reduce this mismatch, which can be roughly divided into two classes: model-based and feature-based approaches. In model-based approaches, compensation is performed on the pre-trained recognition model parameters so that the modified recognition models can more effectively classify the mismatched test speech features collected in the application environment. Typical examples of this class include noise masking [1-3], speech and noise decomposition (SND) [4], vector Taylor series (VTS) [5], maximum likelihood linear regression (MLLR) [6], model-based stochastic matching [7,8], model compensation based on non-uniform spectral compression (MC-SNSC) [9], statistical re-estimation (STAR) [10], and parallel model combination (PMC) [11-13] methods. In the feature-based approaches, a noise-robust feature representation is developed to reduce the sensitivity to various acoustic conditions and thereby alleviate the mismatch between those features used for training and testing. Examples of this class include spectral subtraction (SS) [14-17], Weiner filtering [18,19], short-time spectral amplitude estimation based on minimum mean-squared error criteria (MMSE-STSA) [20], MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [21], codeword-dependent cepstral normalization (CDCN) [22], SNR-dependent non-uniform spectral compression scheme (SNSC) [23], feature-based stochastic matching [7,8], multivariate Gaussian-based cepstral normalization (RATZ) [10], stereo-based piecewise linear compensation for environments (SPLICE) [24,25] methods, and a series of cepstral-feature statistics normalization techniques such as cepstral mean subtraction (CMS) [26], cepstral mean and variance normalization (MVN) [27], MVN plus ARMA filtering (MVA) [28], cepstral gain normalization (CGN) [29], histogram equalization (HEQ) [30,31], and cepstral shape normalization (CSN) [32]. A common advantage of the feature-based methods is their relative simplicity of implementation. This simplicity arises because all of these methods focus on front-end speech feature processing without any need to change the back-end model training and recognition schemes. Despite their simplicity, these methods usually improve recognition performance significantly in noise-corrupted application environments.

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

Springer

© 2012 Hung et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The mel-frequency cepstral coefficient (MFCC) is one of the most widely used speech feature representations due to its high recognition performance under clean conditions. However, MFCC is not very noise-robust, and thus many robustness techniques mentioned above can be applied in various domains of a speech signal when deriving MFCC. For example, SS, WF, MMSE-STSA, and MMSE log-STSA techniques are used in the spectral domain whereas CMS, MVN, MVA, and HEQ are often used in the cepstral domain. In particular, the method presented in this article is designed to compensate the spectrum of the speech signal to obtain more noise-robust MFCC.

In addition to MFCC features, the energy-related feature, i.e., the logarithmic energy (log E), is also effective in discriminating different phonemes. For this reason, it is often appended to the MFCC features to further enhance recognition performance. However, similar to MFCC, the log E feature is vulnerable to noise. In many recent studies [33-35], it has been found that compensating the log E feature can improve the recognition accuracy significantly under noisy conditions. For example, in our previously proposed method, silence feature normalization (SFN) [35], high-pass-filtered log E is used as the indicator for speech/non-speech frame classification, and the log E features of non-speech frames are set to be small, while those of speech frames are kept nearly unchanged. We have shown that SFN is very effective despite its simplicity in implementation.

Partially inspired by the concept of SFN, in our previous work [36] we presented another approach, called magnitude spectrum enhancement (MSE) to further process the magnitude spectra of speech frames. Initial experiments shown in [36] have indicated that MSE produced good results on the Aurora-2 evaluation task [37]. The main purpose of this article is to provide a rigorous investigation for the background of MSE, as well as a series of experiments to further show the effectiveness of MSE in reducing the effect of noise for speech recognition. In MSE, the noise-corrupted signal is processed in the linear spectral domain, with the hope that the resulting speech features are more noise-robust. Briefly speaking, in MSE, the magnitude spectrum of each non-speech frame is set to be small (as in SFN), whereas the magnitude spectrum of each speech frame is amplified by multiplying by a weighting factor that is related to the signal-to-noise ratio (SNR). The main purpose of MSE is to highlight the spectral difference between the speech and non-speech frames, not to re-construct the clean speech spectrum as SS and WF do. The experiments conducted on the Aurora-2 digit database show that our proposed MSE can provide a significant improvement in recognition accuracy in various noise-corrupted environments. MSE performs better than many spectral-domain methods, and it can be well integrated with cepstral-domain processing techniques, such as MVN, MVA, and HEQ. The best possible average accuracy rate for the Aurora-2 clean-condition training task with the proposed method can be as high as 83.80%.

The remainder of this article is organized as follows. Section 'Effect of additive noise to the linear and logarithmic magnitude spectrum of a speech signal' provides a mathematical analysis of a noise-corrupted speech signal as background knowledge for the presented MSE. Next, detailed MSE procedures are described in Section 'The magnitude spectrum enhancement (MSE) approach'. Section 'Experimental results and discussions, contains the experimental setup and a series of experimental results together with the corresponding discussions. Finally, the concluding remarks are given in Section 'Conclusions'.

Effect of additive noise to the linear and logarithmic magnitude spectrum of a speech signal

In this section, we provide a mathematical analysis for the effects of additive noise to the linear and logarithmic magnitude spectrum in a speech signal. Observing these effects will help us develop and present the new noise-robustness approach in Section 'The magnitude spectrum enhancement (MSE) approach'.

Effect of additive noise on the magnitude spectra of speech/non-speech frames

Assume that the signal for an arbitrary frame of a noise-corrupted utterance can be represented by

$$x_m[n] = s_m[n] + d_m[n], \quad 0 \le m \le M-1, \tag{1}$$

where m is the frame index, M is the total number of frames, and $s_m[n]$ and $d_m[n]$ are the speech and noise components of $x_m[n]$, respectively. Taking the discrete Fourier transform (DFT) on both sides of Equation (1), we have

$$X_m[k] = S_m[k] + D_m[k], \tag{2}$$

where $X_m[k]$, $S_m[k]$, and $D_m[k]$ represent the spectra of $x_m[n]$, $s_m[n]$, and $d_m[n]$, respectively, for the kth frequency bin. Obviously, the speech component $S_m[k]$ approaches zero in Equation (2) for a *non-speech* frame. Here, a parameter called the magnitude spectral ratio (MSR) is defined as

$$\gamma[k] = E\left(\frac{|S_p[k] + D_p[k]|}{|D_q[k]|}\right), \; p \neq q, \tag{3}$$

and represents the expectation of the ratio of a *speech* frame (frame p) to a *non-speech* frame (frame q) in the magnitude spectrum for the kth frequency bin. It can be shown that, under an additive white Gaussian noise (AWGN) environment and assuming that $S_p[k]$ is a constant, $|S_p[k] + D_p[k]|$ and $|D_q[k]|$ in Equation (3) are two

random variables with Rician and Rayleigh distributions [38], respectively. The parameter MSR in Equation (3) is then

$$\gamma[k] = \frac{\pi}{2}\exp\left(-\frac{|S_p[k]|^2}{4\sigma^2}\right)\left(\left(1+\frac{|S_p[k]|^2}{2\sigma^2}\right)I_0\left(\frac{|S_p[k]|^2}{4\sigma^2}\right) + \frac{|S_p[k]|^2}{2\sigma^2}I_1\left(\frac{|S_p[k]|^2}{4\sigma^2}\right)\right), \tag{4}$$

where σ^2 is the variance of the real- and imaginary- parts of the noise $D_m[k]$, $m = p, q$, and $I_0(.)$ and $I_1(.)$ are the modified Bessel functions of the first kind with orders zero and one, respectively. Furthermore, $\gamma[k]$ in Equation (4) is, in fact, *monotonically decreasing with respect to the noise variance* σ^2 (see Appendix 1 for a detailed analysis of the above results), indicating that speech frames become increasingly indistinguishable from non-speech frames based on their magnitude spectra as the signal-to-noise ratio (SNR) decreases.

Effect of additive noise on the logarithmic magnitude spectrum in the frame sequences

First, we investigate the effect of noise on the *logarithmic* magnitude spectrum in an arbitrary frame within an utterance. According to Equation (2), we have

$$\begin{aligned} X_m^{(l)}[k] &= \log(|X_m[k]|) \\ &= 0.5\log(|X_m[k]|^2) \\ &= 0.5\log(|S_m[k]+D_m[k]|^2) \\ &\approx 0.5\log(|S_m[k]|^2+|D_m[k]|^2) \\ &= 0.5\log(\exp(2S_m^{(l)}[k])+\exp(2D_m^{(l)}[k])), \end{aligned} \tag{5}$$

where $X_m^{(l)}[k]$, $S_m^{(l)}[k]$, and $D_m^{(l)}[k]$ are the logarithmic magnitude spectra of $x_m[n]$, $s_m[n]$, and $d_m[n]$, respectively, from Equation (1). Thus, the difference between $X_m^{(l)}[k]$ (for the noise-corrupted speech) and $S_m^{(l)}[k]$ (for the embedded clean speech) is

$$\begin{aligned} \Delta[k] &= X_m^{(l)}[k] - S_m^{(l)}[k] \\ &\approx 0.5\log\left(1+\frac{\exp(2D_m^{(l)}[k])}{\exp(2S_m^{(l)}[k])}\right) \\ &= 0.5\log\left(1+\frac{|D_m[k]|^2}{|S_m[k]|^2}\right), \end{aligned} \tag{6}$$

From Equation (6), it is obvious that under the same noise magnitude level $|D_m[k]|$, the difference $\Delta[k]$ decreases as the speech magnitude $|S_m[k]|$ increases. Therefore, for a noise-corrupted utterance, the logarithmic magnitude spectrum of the speech frame is often less vulnerable to noise than that of the non-speech (noise-only) frame. However, this condition *does not hold* for the (linear) magnitude spectrum.

Next, let us consider the effect of noise on the frame sequence of logarithmic magnitude spectra, denoted by $\{X_m^{(l)}[k]\}_{m=0}^{M-1}$, for the utterance. Taking the Taylor series approximation of Equation (5) with respect to $(S_m^{(l)}[k], D_m^{(l)}[k]) = (0,0)$ up to order 2, we have

$$\begin{aligned} X_m^{(l)}[k] &= 0.5\log(\exp(2S_m^{(l)}[k])+\exp(2D_m^{(l)}[k])) \\ &\approx 0.5\log 2 + 0.5(S_m^{(l)}[k]+D_m^{(l)}[k]) + 0.25((S_m^{(l)}[k])^2 \\ &\quad + (D_m^{(l)}[k])^2 - 2S_m^{(l)}[k]D_m^{(l)}[k]). \end{aligned} \tag{7}$$

Thus the modulation spectrum $M_X(j\omega)$ of the sequence $\{X_m^{(l)}[k]\}_{m=0}^{M-1}$, computed by

$$M_X(j\omega) = \sum_{m=0}^{M-1} X_m^{(l)}[k]\, e^{-j\omega m}, \tag{8}$$

can be approximated as

$$\begin{aligned} M_X(j\omega) &\approx (\pi\log 2)\delta(\omega) + 0.5(M_S(j\omega)+M_D(j\omega)) \\ &\quad + \frac{1}{8\pi}(M_S(j\omega)*M_S(j\omega)+M_D(j\omega)*M_D(j\omega) \\ &\quad - 2M_S(j\omega)*M_D(j\omega)), \end{aligned} \tag{9}$$

where $M_X(j\omega)$, $M_S(j\omega)$, and $M_D(j\omega)$ are discrete-time Fourier transforms (DTFTs) of $\{X_m^{(l)}[k]\}_{m=0}^{M-1}$, $\{S_m^{(l)}[k]\}_{m=0}^{M-1}$, and $\{D_m^{(l)}[k]\}_{m=0}^{M-1}$ (along the frame axis with the index m, as in Equation (8)), respectively, and the symbol "$*$" denotes the convolution operation. If the two sequences, $\{S_m^{(l)}[k]\}_{m=0}^{M-1}$ and $\{D_m^{(l)}[k]\}_{m=0}^{M-1}$, are both low-pass and their bandwidths are B_s and B_d, respectively, then the terms $M_D(j\omega)*M_D(j\omega)$ and $M_S(j\omega)*M_D(j\omega)$ in Equation (9) have bandwidths of $2B_d$ and $B_s + B_d$, respectively. This finding implies that $\{X_m^{(l)}[k]\}_{m=0}^{M-1}$ has a wider bandwidth than $\{D_m^{(l)}[k]\}_{m=0}^{M-1}$. In other words, the logarithmic magnitude spectrum of the noise-corrupted speech segment possesses higher modulation frequency components than that of the noise-only segment in a noisy utterance. Again, this condition *does not hold* for the (linear) magnitude spectrum.

Note: it is easy to demonstrate that the above analysis of the logarithmic magnitude spectrum can be performed on the logarithmic energy ($\log E$) sequence in an utterance, obtaining the same conclusions [35]. That is,

1. The logarithmic energy is less distorted in a speech frame than in a non-speech frame.
2. For the logarithmic energy sequence of a noisy utterance, the speech segment possesses components of even-higher frequency than the non-speech segment.

The magnitude spectrum enhancement (MSE) approach

In this section, we describe a compensation scheme termed magnitude spectrum enhancement (MSE) [36] in order to improve the noise robustness of speech features. Briefly speaking, the magnitude spectra of the speech frames are enlarged in MSE whereas those of the non-speech frames are normalized to be very small. In addition, the speech/non-speech frame classification in this scheme is based on the logarithmic magnitude spectra and the logarithmic energy feature of the frames. Details of the MSE procedure are stated below.

Following the notations introduced in Section 'Effect of additive noise to the linear and logarithmic magnitude spectrum of a speech signal', here $\{x_m[n],\ 0 \le n \le N-1\}$ is the time-domain signal for the mth frame of an utterance and N is the frame length. The spectrum for this frame is calculated as

$$X_m[k]=\sum_{n=0}^{N-1} x_m[n]\, e^{-j\frac{2\pi nk}{K}},\quad 0 \le k \le \lfloor \frac{K}{2} \rfloor,\ 0 \le m \le M-1, \tag{10}$$

where K is the DFT size, and M is total number of frames in this utterance. Thus, $|X_m[k]|$ represents the magnitude spectrum for the kth frequency bin of the mth frame. In addition, the logarithmic energy ($\log E$) feature of the mth frame is given by

$$e_m = \log\left(\sum_{n=0}^{N-1} x_m^2[n]\right),\quad 0 \le m \le M-1. \tag{11}$$

The proposed magnitude spectrum enhancement (MSE) approach uses the following two steps to create the new magnitude spectrum.

Step I: Perform voice activity detection (VAD):

The VAD process that discriminates speech frames from non-speech frames in an utterance is based on two sources: the logarithmic magnitude spectrum (abbreviated as logMS) in Equation (10) and log E in Equation (11). Based on the observations made in Section 'Effect of additive noise on the logarithmic magnitude spectrum in the frame sequences', noise-corrupted speech segments possess a greater number of high (modulation) frequency components in the logMS and log E sequence than noise-only segments, and thus we expect that the high-pass-filtered logMS and log E sequences help to obtain more accurate VAD results.

As for the first source, we process the logMS sequence $\{\log(|X_m[k]|)\}_{m=0}^{M-1}$ with a high-pass IIR filter with an input-output relationship of

$$Y_m[k]=\log(|X_m[k]|)-\lambda Y_{m-1}[k],\ 0 \le k \le \lfloor \frac{K}{2} \rfloor,\ 0 \le m \le M-1, \tag{12}$$

where $0 \le \lambda < 1$ (the case $\lambda = 1$ leads to an unstable filter). The frequency response (magnitude part) of the high-pass IIR filter is depicted in Figure 1, showing that this filter emphasizes the higher frequency portions while *not* eliminating the near-DC components completely.

Next, we sum up the high-pass filtered logarithmic spectrum, $Y_m[k]$, over the entire frequency band for each frame:

$$z_m = \sum_{k=0}^{\lfloor \frac{K}{2} \rfloor} Y_m[k]. \tag{13}$$

Thus, z_m in Equation (13) is viewed as the cumulative high-pass-filtered logarithmic spectral magnitude of the mth frame. Finally, the first speech/non-speech decision parameter $d_{m,1}$ is obtained as follows:

$$d_{m,1} = \begin{cases} 1 \text{ if } z_m \ge \theta_z \\ 0 \text{ otherwise} \end{cases},\quad 0 \le m \le M-1, \tag{14}$$

where the threshold θ_z is simply set to the mean of the stream $\{z_m,\ 0 \le m \le M-1\}$.

As for the second source (the log E sequence) for the VAD process, we obtain the second speech/non-speech decision parameter $d_{m,2}$ for the mth frame,

$$d_{m,2} = \begin{cases} 1 \text{ if } e_m^{(h)} \ge \theta_e \\ 0 \text{ otherwise} \end{cases},\quad 0 \le m \le M-1, \tag{15}$$

where $e_m^{(h)}$ is the high-pass filtered version of e_m in Equation (11), in which the high-pass IIR filter is the same as that used in Equation (12). Again, the threshold θ_e is set to the mean of the stream $\{e_m^{(h)},\ 0 \le m \le M-1\}$.

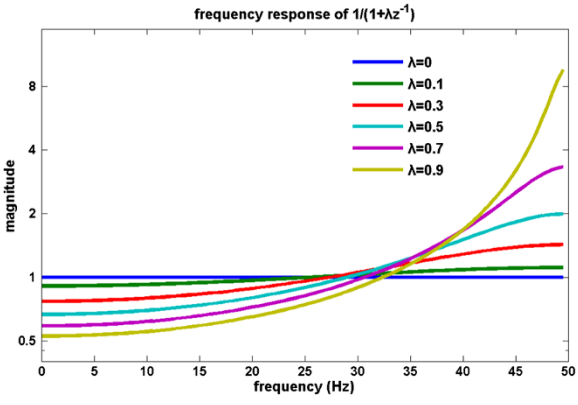

Figure 1 The frequency response (magnitude part) of the high-pass filter $H(z) = \frac{1}{1+\lambda z^{-1}}$ with different assignments of λ.

Finally, the result of the VAD process is obtained from the two parameters $d_{m,1}$ in Equation (14) and $d_{m,2}$ in Equation (15):

$$d_m = \begin{cases} 1 \text{ if } d_{m,1} = 1 \textbf{ or } d_{m,2} = 1 \\ 0 \text{ otherwise} \end{cases}, \quad 0 \le m \le M-1, \tag{16}$$

where d_m is the VAD indicator finally used. That is, the mth frame is classified as speech if either $d_{m,1}$ or $d_{m,2}$ is equal to unity. The main reason for using the "**or**" operation in Equation (16) is that the speech frames are likely to be misclassified as non-speech frames (i.e., a higher false-rejection rate) when we simply depend on either decision parameter $d_{m,1}$ or $d_{m,2}$ alone, especially when the SNR degrades.

Step II. Obtain the enhanced magnitude spectrum

This step amplifies the magnitude spectrum for the speech frames while diminishing it for the non-speech frames. The main purpose of this step is to enlarge the ratio of speech frames to non-speech frames in magnitude spectra to reduce the noise effect, as discussed in Section 'Effect of additive noise to the linear and logarithmic magnitude spectrum of a speech signal'. The magnitude spectra for the non-speech frames detected in Step I are first collected and then averaged to obtain the estimated noise (magnitude) spectrum for the utterance:

$$N[k] = \frac{\sum_{m=0}^{M-1}(1-d_m)|X_m[k]|}{\sum_{m=0}^{M-1}(1-d_m)}, \quad 0 \le k \le \lfloor \frac{K}{2} \rfloor. \tag{17}$$

Note that here, $N[k]$ is independent of the frame index m. Thus, the noise spectrum is estimated once for the utterance.

Next, a weighting factor for each magnitude spectral value $X_m[k]$ is defined as follows:

$$w_m[k] = \begin{cases} \left(\frac{|X_m[k]|}{N[k]+\delta}\right)^{\alpha} & \text{if } d_m = 1 \\ \varepsilon & \text{if } d_m = 0 \end{cases}, \quad 0 \le k \le \lfloor \frac{K}{2} \rfloor, \; 0 \le m \le M-1, \tag{18}$$

where α is a parameter within the range $[0, 1]$ that determines the degree of amplification, δ is a small positive constant that avoids the weighting factor becoming infinitely large as $N[k] \longrightarrow 0$, and ε is a very small positive random variable such that the magnitude spectra of detected non-speech frames are significantly reduced.

Thus, the weighting factor for a speech frame ($d_m = 1$) in Equation (18) is related to the SNR as follows:

$$w_m[k] \approx (\sqrt{SNR_m[k]} + 1)^{\alpha}, \tag{19}$$

where $SNR_m[k] = \left(\frac{|X_m[k]|^2}{N^2[k]}\right) - 1$ is the (estimated) SNR for the kth frequency bin of the mth frame.

Finally, the enhanced magnitude spectrum is obtained by multiplying the original magnitude spectrum with the weighting factor $w_m[k]$ in Equation (18):

$$|\tilde{X}_m[k]| = w_m[k]\,|X_m[k]|, 0 \le k \le \lfloor \frac{K}{2} \rfloor, 0 \le m \le M-1, \tag{20}$$

The proposed MSE has the following properties:

1. In MSE, the embedded VAD process uses the logarithmic magnitude spectrum rather than the linear magnitude spectrum. According to the discussions in Section 'Effect of additive noise to the linear and logarithmic magnitude spectrum of a speech signal', the logarithmic magnitude spectrum is less vulnerable to noise in speech frames, and its temporal-domain sequence exhibits a wider (modulation) spectral bandwidth in speech portions than in non-speech portions. Based on these two characteristics, the logarithmic magnitude spectrum is a more appropriate VAD indicator than the linear magnitude spectrum. The experimental results shown later will reveal that the logarithmic magnitude spectrum outperforms the linear magnitude spectrum for providing MSE with better recognition accuracy.
2. By assigning different weights to the magnitude spectra of speech and non-speech frames, the speech portions of an utterance are highlighted and the difference between the speech and non-speech portions in magnitude spectrum is strongly emphasized. This effect leads to a large magnitude spectral ratio (MSR) as defined in Equation (2) and implies that the effect of noise has been effectively reduced.
3. The idea of MSE is partially motivated by the matched filter theory in the field of communications [38]. For an observed signal denoted by $x[n] = s[n] + d[n]$, where $s[n]$ and $d[n]$ are the desired signal and additive noise, respectively, the magnitude (frequency) response of the matched filter which maximizes the output SNR is [39]:

$$|H(j\omega)| = \frac{|S(j\omega)|}{P_d(\omega)}, \tag{21}$$

where $|S(j\omega)|$ and $P_d(\omega)$ are the magnitude spectrum of $s[n]$ and the power spectral density of the noise $d[n]$, respectively. From Equation (21), we find $|H(j\omega)|$ is proportional to the input frequency domain SNR (defined by $\frac{|S(j\omega)|^2}{P_d(\omega)}$) provided the signal level $|S(j\omega)|^2$ or the noise level $P_d(\omega)$ is fixed. Thus, MSE shares the idea of the matched filter and uses a spectral weighting factor $w_m[k]$ in

Equation (18) which is positively correlated with the SNR. However, MSE differs from the matched filter in some aspects: First, MSE applies the magnitude spectrum of the noisy signal $x[n]$ rather than that of the clean signal $s[n]$, which is not available and requires estimation. Second, the magnitude spectrum of the noise $d[n]$ is used, which approximates the square root of the power spectral density of the noise. Finally, MSE additionally detects the non-speech regions and makes the corresponding spectra nearly zero, which is a non-linear operation and can further distinguish the speech and non-speech frames.

4. Compared to the SFN method [35], the magnitude spectrum in MSE for the features in non-speech portions is set to be small. However, in speech portions of the utterance, MSE further amplifies the magnitude spectrum, whereas in SFN the energy-related feature is kept nearly unchanged.
5. Like spectral compensation techniques, spectral subtraction (SS) [14-16] and Weiner filtering (WF) [18,19], MSE attempts to reduce the effect of noise in the spectral domain of speech signals. However, the main purpose of SS and WF is to restore a clean spectrum from the noise-corrupted utterance. This situation contrasts with MSE, where the (magnitude) spectrum of the speech portions is amplified, possibly making the resulting spectrum quite different from the clean spectrum. In general, the updated magnitude spectra using SS and WF are often presented as follows:

$$\text{SS:}\ \ |\tilde{X}_m[k]| \approx |X_m[k]| \left(1 + \frac{1}{SNR_m[k]}\right)^{-\frac{1}{2}}, \tag{22}$$

$$\text{WF:}\ \ |\tilde{X}_m[k]| \approx |X_m[k]| \left(1 + \frac{1}{SNR_m[k]}\right)^{-1}. \tag{23}$$

For MSE, the new magnitude spectrum is:

$$\text{MSE:}\ \ |\tilde{X}_m[k]| \approx |X_m[k]| (\sqrt{SNR_m[k]} + 1)^{\alpha} \times (\text{for speech frames}). \tag{24}$$

In addition, the speech and non-speech portions are treated quite differently in MSE (as shown in Equations (18) and (20)), while they are not explicitly treated differently in SS and WF.

6. In MSE, the VAD procedure used in Step I is quite simple to implement and can be replaced with any other VAD method. In addition, the cepstral features derived from the MSE-processed spectrum can be further compensated using any cepstral-domain robustness techniques such as MVN, MVA, and HEQ to achieve further improvements in recognition performance, which will be shown in Section 'Experimental results and discussions'.

Experimental results and discussions

We use two sets of experimental environments in this article. In the first environment, the Aurora-2 connected US-digit database [37] is the platform for evaluating the proposed MSE and other various techniques. It is used to explore the resulting spectrograms of the speech signals processed by MSE and some other spectral-domain processes, to analyze the possible improvements achievable by each approach, and to discuss the comparisons among different techniques. On the other hand, in the second environment, the NUM-100A continuous Mandarin speech database [40] is used. This database contains microphone-recoded Mandarin digit strings produced by Mandarin adults. We perform the proposed MSE on this data set to further investigate if MSE is still effective in processing the noisy speech that belongs to a different language.

Experiments for the Aurora-2 database

Here, the presented MSE scheme has been tested with the AURORA Project Database Version 2.0 (Aurora-2), the details of which are described in [37]. In short, the testing data consist of 4004 utterances from 52 female and 52 male speakers, and three different subsets are defined for the recognition experiments: Test Sets A and B are each affected by four types of noise, and Set C is affected by two types. Each noise instance is 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. G.712 and MIRS are two standard frequency characteristics defined by the ITU [41].

The Aurora-2 task has the following two training modes [37]:

1. In the first mode, "clean-condition training", the training data consist of 8440 *clean* speech utterances from 55 female and 55 male adults.
2. In the second mode, "multi-condition training", the clean training data in the first mode are equally split into 20 subsets. These 20 subsets are added with four different types of noise at five different SNRs. The four noise types are suburban train, babble, car and exhibition hall, which are the same as the noise types in Test Set A. The SNRs are 20, 15, 10, and 5 dB and the clean condition.

Therefore, in the first mode, "clean-condition training", the obtained clean acoustic models contain no information about the possible distortions. This mode can help us

evaluate the degree of robust capability of the speech features (associated with the robustness algorithm) against noise. As for the second mode, "multi-condition training", the corresponding results can reveal the impact of a different type of noise or a different SNR than seen during training [37]. In our following experiments and discussions, we will primarily focus on the first mode in order to observe the presented MSE in the reduction of noise effect. However, we will also provide the experimental results for the second mode together with relatively brief discussions.

Results for the task of clean-condition training and multi-condition testing

With the Aurora-2 database under the mode of "clean-condition training", we perform the MSE method and a series of robustness methods to compare the recognition accuracy. As for the cepstral-domain methods, each utterance in the clean training set and three testing sets is directly converted to 13-dimensional MFCC ($c1$–$c12$, $c0$) sequence according to the feature settings in [37]. Next, the MFCC features are processed using MVN, MVA or HEQ. The spectral-domain methods used here include our MSE, spectral subtraction (SS), Wiener filtering (WF) and MMSE-based log-spectral amplitude estimation (MMSE log-STSA). Each utterance is first processed in the linear spectral domain. The updated spectra are converted to a sequence of 13-dimensional MFCC (($c1$–$c12$, $c0$)). The resulting 13 new features, plus 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 silence are trained with the demo scripts provided by the Aurora-2 CD set [42]. Each digit HMM has 16 states, with 3 Gaussian mixtures per state.

Detailed information about some of the methods used follows:

1. We apply three versions of spectral subtraction (SS) proposed in [14-16]. For the purposes of clarity, they are denoted by SS_{Boll}, $SS_{Berouti}$, and SS_{Kamath}, respectively, in which the author names are represented by the subscripts.
2. As with spectral subtraction, three versions of the Wiener filtering (WF) methods proposed in [18,19] are tested here. The first method is based on *a priori* signal to noise ratio (PSNR) estimation, and the latter two WF methods apply a two-step noise reduction (TSNR) procedure and a harmonic regeneration noise reduction (HRNR) scheme, respectively. Thus, these methods are abbreviated as WF_{PSNR}, WF_{TSNR}, and WF_{HRNR} for later discussions.
3. For the proposed MSE, the parameters δ in Equation (18) is set to 0.001, and the positive random number ε in Equation (18) is uniformly distributed within the range $(0, 10^{-5})$. In order to obtain a proper selection of the filter coefficient λ in Equation (12) and the weight parameter α in Equation (18), we use the 8440 noise-corrupted training utterances for the mode of "multi-condition training" in the Aurora-2 database as *the development set*. The averaged recognition accuracy rates with respect to different assignments of λ and α (both from 0.1 to 0.9 with an interval of 0.2) are shown in Table 1. As a result, we set λ and α to 0.7 and 0.5, respectively, since such a setting gets the optimal accuracy rate for the development set.
4. For MVA, the order of the ARMA filter is set to 3.
5. For HEQ, each feature stream in the utterance is normalized to approach a Gaussian distribution with zero mean and unity variance.

Comparison of various noise robustness approaches

Table 2 presents the individual set recognition accuracy rates averaged over five SNR conditions (0–20 dB at 5 dB intervals) for Test Sets A, B, and C, achieved using various approaches. Figure 2 shows the accuracy rates for spectral-domain methods under different SNR conditions, which are obtained by averaging over all ten noise types contained in the three Test Sets. Based on Table 2 and Figure 2, we make the following observations:

1. Compared to baseline processing, most approaches provide significant recognition accuracy improvement in almost all cases. All three SS methods give better results than the baseline for Test Sets A and B, while the improvement for Test Set C is relatively insignificant. A possible explanation of this finding is that SS is particularly designed to alleviate additive noise and thus does not handle the channel mismatch in the utterances of Test Set C

Table 1 The averaged recognition accuracy rates (%) of the development set (the multi-condition training data in the AURORA-2 database) achieved by MSE with different filter coefficients λ in Equation (12) and exponents α in Equation (18)

The exponent α in Equation (18)	The filter coefficient λ in Equation (12)				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
0.1	88.22	88.44	88.73	88.96	88.90
0.3	89.21	89.49	89.71	90.03	89.54
0.5	89.53	89.66	89.86	**90.38**	89.78
0.7	89.98	89.94	90.21	90.28	89.54
0.9	89.93	89.80	90.06	90.27	89.31

Table 2 Recognition accuracy (%) achieved by various approaches for Aurora-2 clean-condition training task averaged across the SNRs between 0 and 20 dB, where AVG (%) and RR (%) are the averaged accuracy rate and the relative error rate reduction over the baseline

Method	Set A	Set B	Set C	AVG	RR
MFCC baseline	59.24	56.37	67.53	59.75	–
Spectral-domain methods					
SS_{Boll}	61.81	64.09	60.09	62.38	6.53
$SS_{Berouti}$	69.76	70.47	69.39	69.97	25.40
SS_{Kamath}	66.91	67.50	67.19	67.20	18.52
WF_{PSNR}	71.78	73.66	70.37	72.25	31.05
WF_{TSNR}	51.12	55.90	45.64	51.94	-19.41
WF_{HRNR}	56.20	59.65	53.47	57.03	-6.74
MMSE log-STSA	72.71	73.58	71.99	72.91	32.71
MSE	**77.76**	**79.89**	**69.42**	**76.94**	**42.72**
Cepstral-domain methods					
MVN	73.81	75.02	75.08	74.55	36.77
HEQ	81.42	83.34	81.51	82.21	55.80
MVA	78.15	79.17	79.12	78.75	47.21

very well. On the other hand, WF_{PSNR} performs the best among the three Wiener filtering approaches, while WF_{TSNR} and WF_{HRNR} result in poorer accuracy rates relative to the MFCC baseline. Furthermore, WF_{PSNR} behaves better than SS and is also very helpful with Test Set C. Finally, the method "MMSE log-STSA" performs quite well, and its corresponding averaged recognition accuracy is slightly better than that of WF_{PSNR}.

2. Among the spectral-domain methods studied, the proposed MSE method outperforms MMSE log-STSA and various versions of SS and WF in almost all cases. Furthermore, MSE leads to a relative error reduction rate of 49.82% for additive-noise conditions (Test Sets A and B) and 42.72% for all conditions (Test Sets A, B and C) compared with baseline results. The results show that MSE effectively enhances the robustness of MFCC in various noise-corrupted environments.
3. The proposed MSE method provides very promising recognition accuracy rates for all SNR conditions. In particular, MSE outperforms WF_{PSNR} and MMSE log-STSA for higher SNR cases (20 and 15 dB), and the three methods deliver very similar accuracy rates for lower SNR cases.
4. Among the three cepstral-domain methods, HEQ behaves the best, followed by MVA and then MVN. In addition, the three cepstral-domain methods perform better than most spectral-domain methods, with the exception that MVN performs worse than MSE for Test Sets A and B. This finding leads to the concept of integrating these cepstral-domain methods with the proposed MSE as discussed below. It will be shown that such integration can offer further improvements in performance.
5. In order to examine if the presented MSE gives rise to a statistically significant improvement in recognition accuracy relative to the other methods, the one-proportion z-test [43] is performed as follows: Let p and p_0 denote the accuracy rates provided by MSE and the method for comparison, respectively. We set the null hypothesis as $H_0 : p = p_0$ and the alternative hypothesis $H_1 : p > p_0$, and the test statistic for the hypothesis is:

$$z = \frac{p - p_0}{\sqrt{p_0(1 - p_0)/N}}, \tag{25}$$

where N is the number of words in the test and here $N = 214465$ for the Aurora-2 evaluation task [37]. If the test statistic z in Equation (25) is larger than about 2.326, then the null hypothesis H_0 is rejected and the improvement is statistically significant with a confidence level of 99% (since

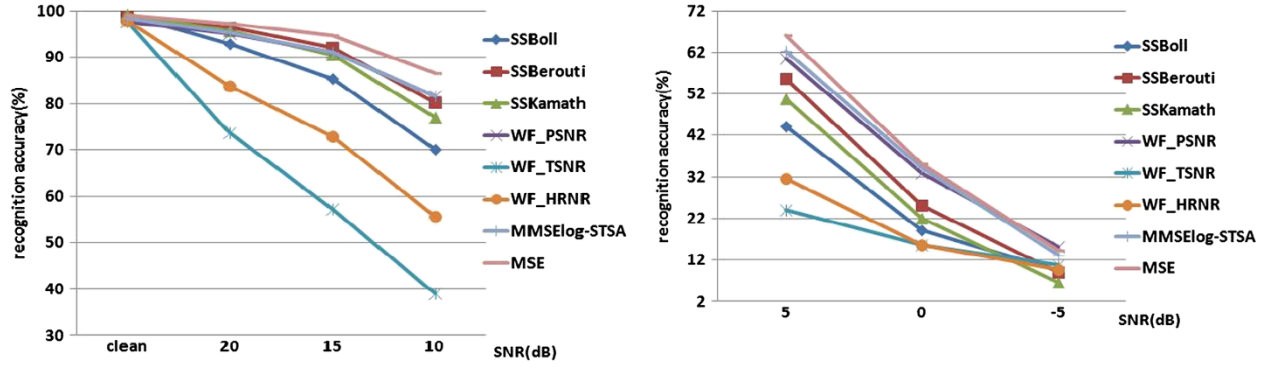

Figure 2 Recognition accuracy (%) achieved by various spectral-domain methods for different SNR conditions averaged over all noise types in three Test Sets for Aurora-2 *clean-condition* training task.

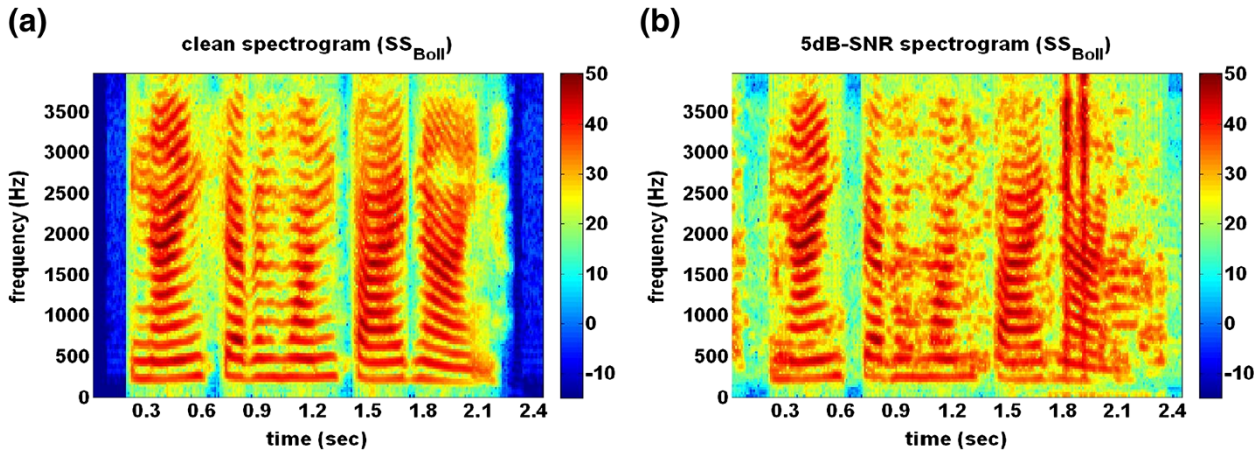

Figure 3 The SS_{Boll}–processed spectrogram of an utterance under two SNR levels: **(a)** clean, **(b)** 5 dB.

$\int_{2.326}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \approx 1\% = 1 - 99\%$). According to the obtained test statistic z in Equation (25), we find that the improvement brought by MSE relative to the other spectral-domain methods is statistically significant. For example, when the method for comparison is MMSE log-STSA, the corresponding test statistic z in Equation (25) is 41.99, far larger than the threshold 2.326.

In addition to the recognition accuracy, we also examine the various spectral-domain methods' capabilities of reducing the spectrogram mismatch caused by additive noise. Figures 3, 4, 5, 6, 7, 8, 9, and 10 show the spectrograms of a digit utterance ("FLJ_97159A.08" in the Aurora-2 database) for two SNR levels, clean and 5 dB (with babble noise), obtained by SS_{Boll}, $SS_{Berouti}$, SS_{Kamath}, WF_{PSNR}, WF_{TSNR}, WF_{HRNR}, MMSE log-STSA and the proposed MSE, respectively. First, the figures show that for the clean case, the voiced portions and the short pauses between any two consecutive digits or syllables are clearly revealed using almost all approaches. Second, for the noise-corrupted case, WF_{PSNR}, MMSE log-STSA, and MSE highlight the short pauses more than the other approaches, and they preserve the voiced segments better with less distortion (especially in the region [0.7 s, 1.3 s]). Thus, the similar treatment of these short pauses under clean and noise-corrupted conditions using the three methods may result in a relatively insignificant mismatch between the two SNR conditions, causing the higher recognition accuracy shown previously. Finally, the detected speech segments are quite obviously separated in the MSE-processed spectrogram, and this fact may be one reason why MSE performs very well.

Integration of MSE with cepstral feature processing techniques MSE, which is performed on the spectral

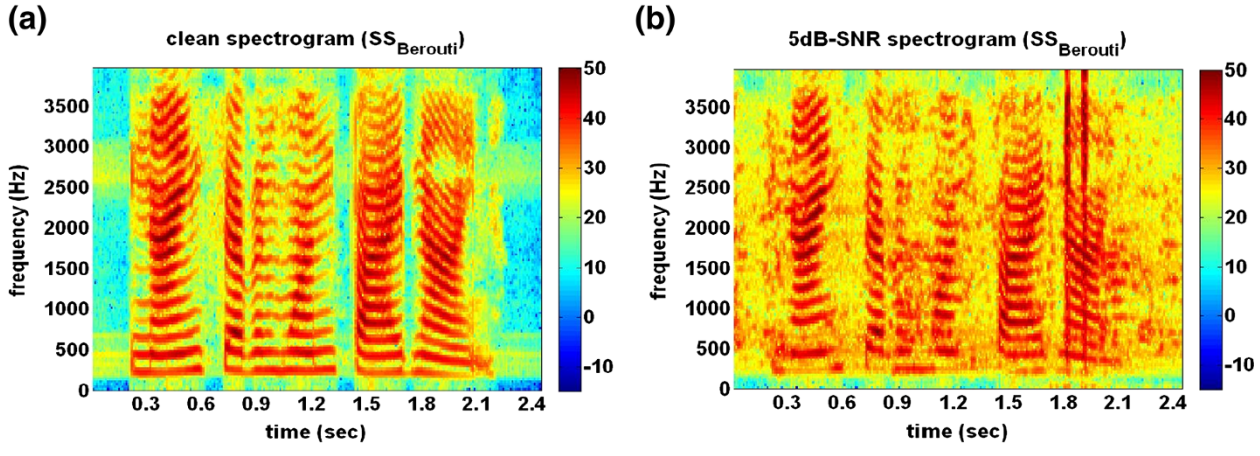

Figure 4 The $SS_{Berouti}$–processed spectrogram of an utterance under two SNR levels: **(a)** clean, **(b)** 5 dB.

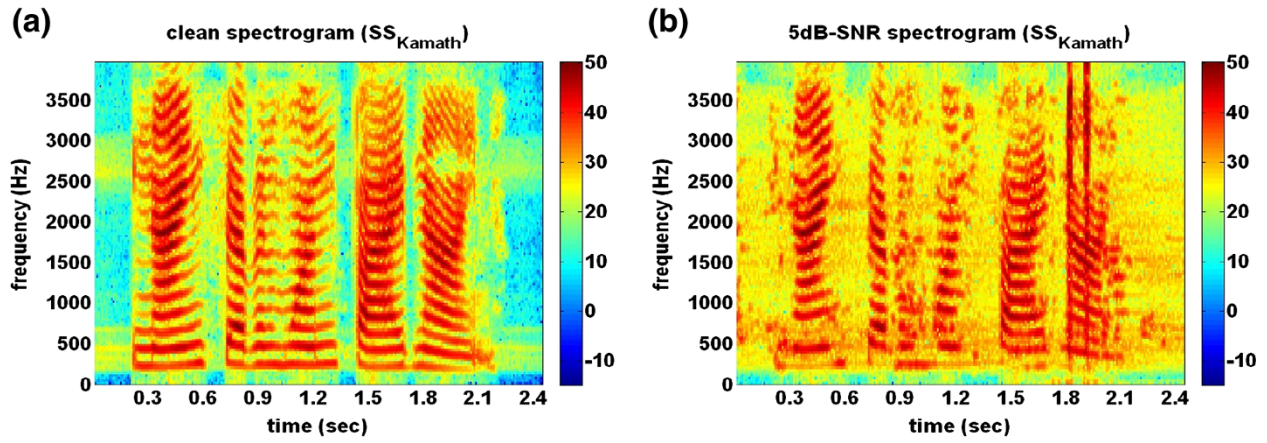

Figure 5 The SS$_{Kamath}$–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

domain of features, can be easily integrated with cepstral-domain processing techniques. Here, we test whether such integration brings about further recognition performance. MFCC features are first derived from the MSE-processed spectra and then processed using MVN, HEQ, or MVA. For a more complete comparison, we also integrate any of the spectral-domain methods, SS$_{Berouti}$, WF$_{PSNR}$, and MMSE log-STSA, with the cepstral-domain method. The corresponding recognition results are shown in Table 3. For the comparison purposes, the accuracy rates for MSE, SS$_{Berouti}$, WF$_{PSNR}$, MMSE log-STSA, MVN, HEQ, and MVA are relisted from Table 2. Several findings are reported in Table 3:

1. The combination of MSE and the cepstral-domain method produces better results than the individual component methods in most cases. For example, MSE plus MVA (82.37%) is better than MSE (76.94%) and MVA (78.75%) in recognition accuracy averaged over ten noise types among the three Test Sets and results in a relative error reduction rate of 56.20%. Similar results are achieved with MSE plus MVN and MSE plus HEQ. These results clearly indicate that MSE can be successfully added to cepstral-domain approaches to further improve noise robustness.
2. For the channel-distorted signals in Test Set C, MSE performs worse than the cepstral-domain methods alone. However, combining MSE with either of MVN and MVA can yield better recognition rates with regard to Test Set C. For example, MSE plus MVA (80.58%) is better than MVA alone (79.12%) in averaged recognition accuracy. Therefore, MSE enhances MVN and MVA in processing channel-distorted signals even though it is primarily designed for additive-noise conditions.
3. Different from MSE, combining any of the three spectral-domain methods, SS$_{Berouti}$, WF$_{PSNR}$, MMSE log-STSA, with any cepstral-domain method

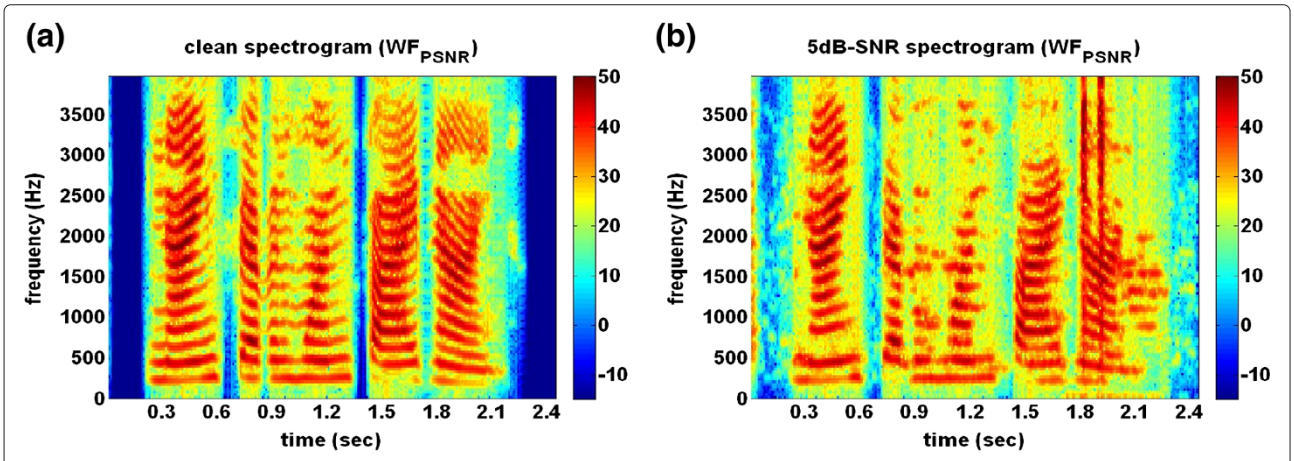

Figure 6 The WF$_{PSNR}$–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

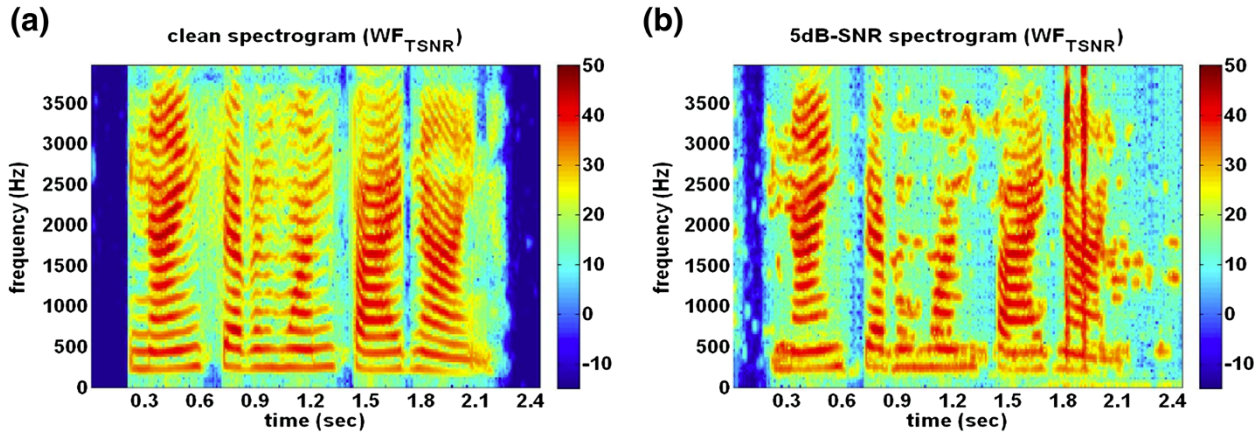

Figure 7 The WF$_{TSNR}$–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

performs worse than the component cepstral-domain method alone. For example, MMSE log-STSA plus HEQ achieves an averaged accuracy of 79.58%, less than 82.21% obtained by the single HEQ. These results again imply that the presented MSE outperforms the other three spectral-domain methods used here.

The influence of the VAD error for MSE in speech recognition In this section, we first investigate the effect of the VAD error on recognition performance in MSE. For this purpose, we perform MSE under the "oracle condition". That is, the VAD results for each clean utterance are directly applied to its various noise-corrupted counterparts to implement the magnitude spectrum enhancement. This process is referred to as "MSE$^{(o)}$" here. Assuming that the VAD error of MSE for a clean utterance is small and negligible, the recognition accuracy difference between MSE$^{(o)}$ and MSE for noise-corrupted utterances can be viewed as a consequence of the VAD error due to noise.

The recognition accuracy rates for MSE$^{(o)}$ and MSE are listed in Table 4. As expected, MSE$^{(o)}$ always performs better than MSE because it contains no VAD errors. However, the difference in accuracy is not very significant. In the worst case (SNR = 0 dB), the performance degradation is 4.96% (1.64% for Set A, 8.77% for Set B, and 4.00% for Set C), and on average, it is 2.90% (1.57% for Set A, 3.92% for Set B, and 3.49% for Set C). These results indicate that the performance of MSE is somewhat influenced by the the error of the embedded VAD process.

Next, we select different VAD indicators for MSE to see the corresponding effect. According to the analysis in Section 'Effect of additive noise on the logarithmic magnitude spectrum in the frame sequences', the high-pass filtered logarithmic magnitude spectrum (log *MS*) and the logarithmic energy (log *E*) can emphasize the difference of

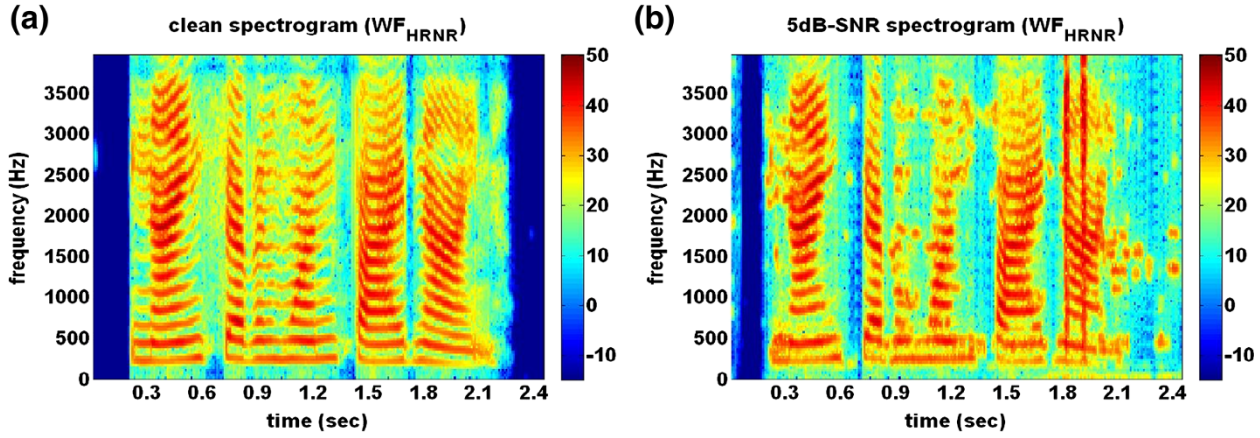

Figure 8 The WF$_{HRNR}$–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

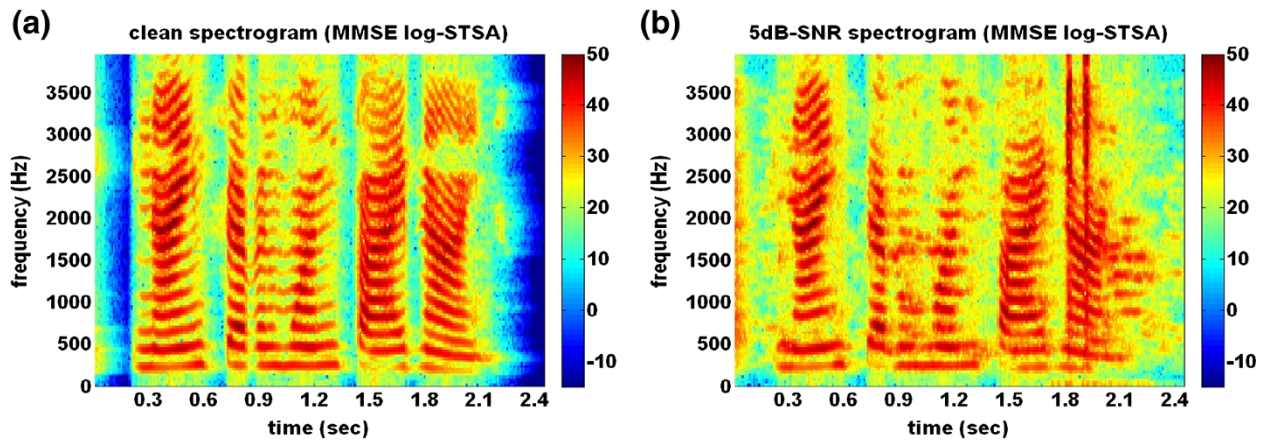

Figure 9 The MMSE log-STSA–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

the speech and non-speech frames, and thus they are chosen to be the VAD indicators of MSE. Here, we adopt the following two alternatives as the VAD indicators:

1. the original linear magnitude spectrum ($X_m[k]$ in Equation (10)) and the energy (the exponent of e_m Equation (11)),
2. the high-pass filtered linear magnitude spectrum and the high-pass filtered energy,

and the corresponding two MSE processes are denoted by $\text{MSE}^{(L_1)}$ and $\text{MSE}^{(L_2)}$, respectively, for simplicity. Figure 11 shows the recognition accuracy rates for $\text{MSE}^{(L_1)}$ and $\text{MSE}^{(L_2)}$ under different SNR conditions for the three Test Sets, and we add the results of the original MSE in this figure for comparison. From this figure, we find that when the SNR is high (clean and 20 dB), there is no substantial performance difference among the three MSE methods. However, when the noise level becomes larger, the original MSE significantly outperforms the other two versions of MSE, $\text{MSE}^{(L_1)}$, and $\text{MSE}^{(L_2)}$. As a result, compared with the linear magnitude spectrum and energy, the high-pass filtered logarithmic magnitude spectrum (as well as the logarithmic energy) can provide more accurate VAD under noisy conditions and achieve better recognition results for the subsequent MSE processing.

Further issues regarding MSE processing Several issues relating to the new proposed MSE scheme are further investigated in this Section.

The effect of the exponent α in MSE One of the central ideas of MSE is to amplify the spectral magnitude for speech frames, and from Equations (18) and (24), the amplification factor (for speech frame) is

$$w_m[k] = \left(\frac{|X_m[k]|}{N[k]+\delta}\right)^{\alpha} \approx \left(\sqrt{SNR_m[k]}+1\right)^{\alpha}, \quad (26)$$

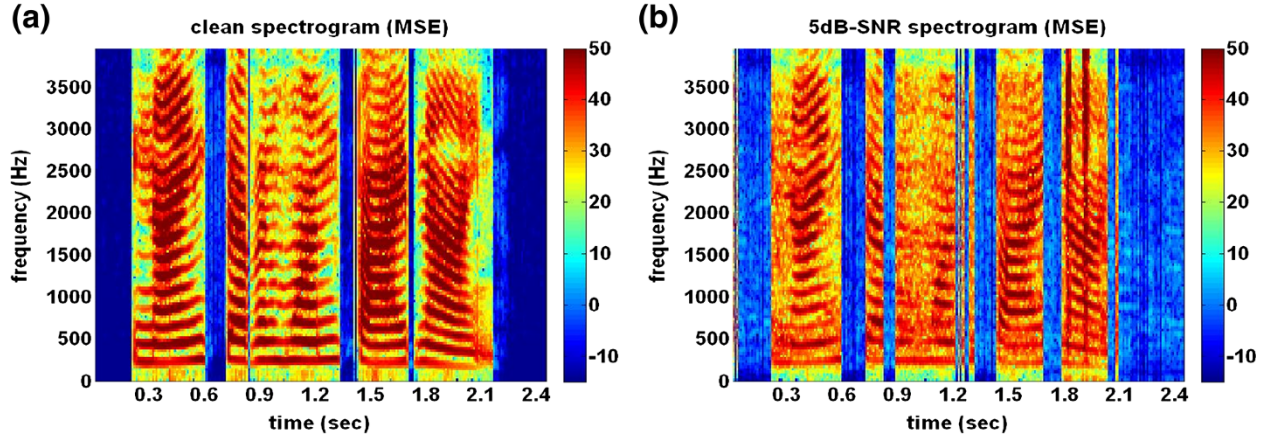

Figure 10 The MSE–processed spectrogram of an utterance under two SNR levels: (a) clean, (b) 5 dB.

Table 3 Recognition accuracy (%) achieved by various approaches for Aurora-2 clean-condition training task averaged across the SNRs between 0 and 20 dB, where AVG (%) and RR (%) are the averaged accuracy rate and the relative error rate reduction over the baseline

Method	Set A	Set B	Set C	AVG	RR
MFCC baseline	59.24	56.37	67.53	59.75	–
$SS_{Berouti}$	69.76	70.47	69.39	69.97	25.40
WF_{PSNR}	71.78	73.66	70.37	72.25	31.05
MMSE log-STSA	72.71	73.58	71.99	72.91	32.71
MSE	77.76	79.89	69.42	76.94	42.72
MVN	73.81	75.02	75.08	74.55	36.77
$SS_{Berouti}$+MVN	65.71	70.39	66.94	67.83	20.07
WF_{PSNR}+MVN	67.33	70.26	67.35	68.50	21.75
MMSE log-STSA+MVN	73.55	75.67	73.33	74.35	36.28
MSE+MVN	**81.85**	**82.15**	**76.23**	**80.85**	**52.42**
HEQ	81.42	83.34	81.51	82.21	55.80
$SS_{Berouti}$+HEQ	73.04	76.99	73.52	74.71	37.18
WF_{PSNR}+HEQ	74.95	77.30	74.92	75.88	40.08
MMSE log-STSA+HEQ	79.13	80.81	77.99	79.58	49.26
MSE+HEQ	**84.19**	**83.20**	**78.20**	**83.80**	**59.75**
MVA	78.15	79.17	79.12	78.75	47.21
$SS_{Berouti}$+MVA	71.07	75.05	72.05	72.86	32.56
WF_{PSNR}+MVA	69.02	71.70	69.01	70.09	25.69
MMSE log-STSA+MVA	74.39	76.79	74.50	75.37	38.81
MSE+MVA	**83.58**	**85.02**	**80.58**	**82.37**	**56.20**

Examining Equation (26), the exponent value α controls the degree of amplification. Increasing the value of α enlarges the difference between the speech and non-speech frames in magnitude spectrum and may also lead to a greater mismatch among the speech frames for the same syllable or phoneme under different SNR conditions. As a result, a larger α in MSE does not always bring about improved recognition accuracy, even if the VAD contains no errors. Here, we assign the exponent α to different values within the range [0, 1] and then proceed with MSE to investigate the corresponding recognition accuracy.

Figure 12 shows the recognition results averaged over five SNR conditions ($0 \sim 20$ dB) and all ten noise types in the three Test Sets for different values of α for MSE (the filter coefficient λ in Equation (12) is fixed as 0.7). As shown in Figure 12, we find that

1. The case $\alpha = 0$, where the magnitude spectrum is kept unchanged in MSE, yields an averaged recognition accuracy of 72.26%, significantly better than the MFCC baseline result (59.75%). This result shows that simply setting the magnitude spectrum of the detected non-speech frames to be nearly zero is beneficial to the recognition performance.
2. The recognition accuracy improves as the value α is increased from 0 to 0.6, and the additional

Table 4 Recognition accuracy (%) achieved by $MSE^{(o)}$ and MSE for Aurora-2 clean-condition training task, where $MSE^{(o)}$ is MSE employing nearly error-free VAD results (MSE in the oracle condition)

SNR	Set A		Set B		Set C		Average	
	$MSE^{(o)}$	MSE	$MSE^{(o)}$	MSE	$MSE^{(o)}$	MSE	$MSE^{(o)}$	MSE
20 dB	97.70	97.31	98.19	97.55	96.91	96.19	97.74	97.18
15 dB	95.76	94.47	96.78	95.54	93.60	90.98	95.73	94.20
10 dB	89.36	87.25	92.76	89.54	82.86	78.67	89.42	86.45
5 dB	72.43	69.97	79.23	73.47	60.47	54.53	72.76	68.28
0 dB	41.44	39.80	52.10	43.33	30.73	26.73	43.56	38.60
average	79.33	77.76	83.81	79.89	72.91	69.42	79.84	76.94

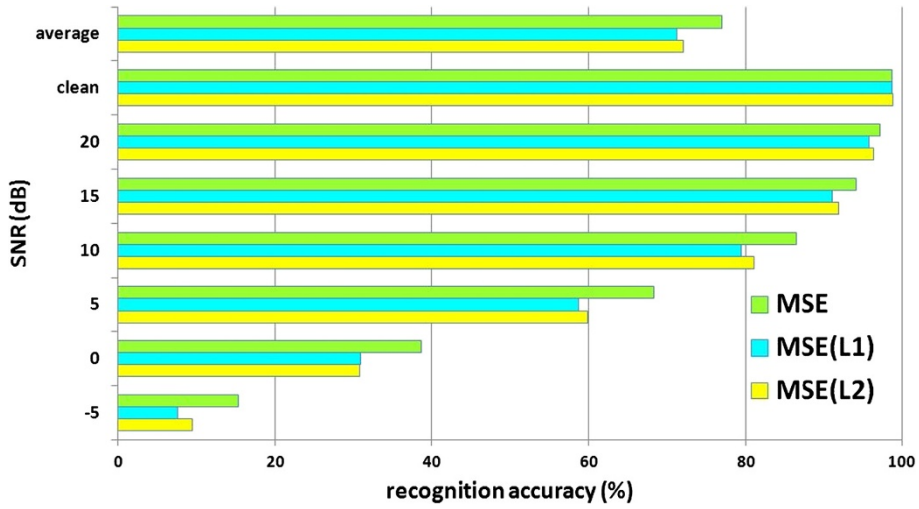

Figure 11 Recognition accuracy (%) (averaged over all the ten noise types in three Test Sets) achieved by MSE, MSE$^{(L_1)}$, and MSE$^{(L_2)}$ for different SNR conditions.

improvement in accuracy is 4.80% (from 72.26% to 77.06%). Therefore, amplifying the magnitude spectrum of the speech frames correctly is helpful.

3. When the exponent α is further increased from 0.6 to 1, the recognition rates worsen, possibly due to the enlarged mismatch among the speech frames mentioned previously. However, the decrease in maximum accuracy is just 0.62% (from 77.06% at $\alpha = 0.6$ to 76.44% at $\alpha = 0.8$), implying that the recognition accuracy is relatively insensitive to α (provided that α is within the range [0.6, 1]).

The effect of the filter coefficient λ in MSE

As stated in Section 'The magnitude spectrum enhancement (MSE) approach', the filter coefficient λ in Equation (12) determines the frequency response of the high-pass filter for the VAD process of MSE. The case $\lambda = 0$ corresponds to using the logarithmic magnitude spectrum (logMS) and the log-energy (log E) directly as the VAD features. On the other hand, increasing values of λ indicate that the lower/higher modulation frequency components are further reduced/emphasized in the logMS and log E streams, as shown in Figure 1. This parameter was preliminarily set to 0.7 in the previous experiments. Now, we vary its value from 0 to 0.9, spaced in 0.1 intervals, to perform the corresponding MSE. (Note that setting $\lambda = 1$ will result in an unstable filter.)

Figure 13 shows the recognition results averaged over five SNR conditions (0 ∼ 20 dB) and all ten noise types in the three Test Sets using different values of λ for MSE (the exponent α in Equation (26) is fixed as 0.5). We first find that applying MSE with a positive λ achieves better results than applying MSE with $\lambda = 0$ in most cases, indicating

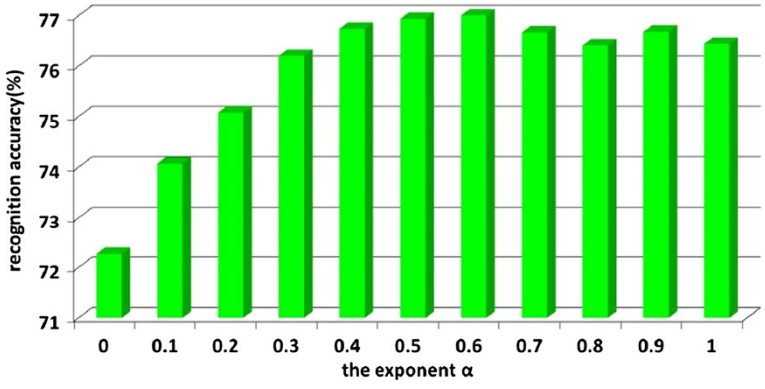

Figure 12 Recognition accuracy (%) (averaged over five SNR values and all the ten noise types in three Test Sets) versus different assignments of the exponent α in MSE.

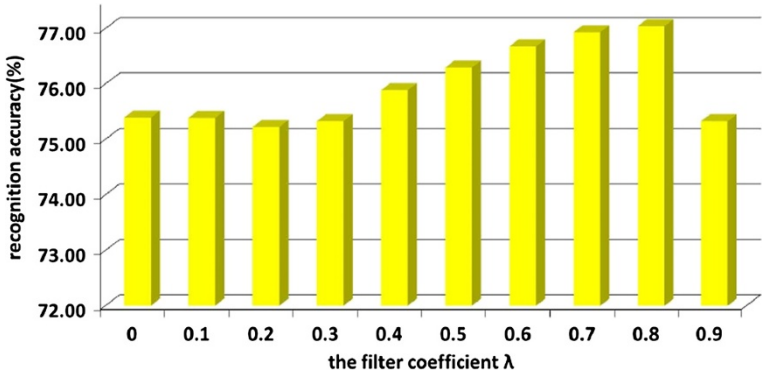

Figure 13 Recognition accuracy (%) (averaged over five SNR values and all ten noise types in three Test Sets) versus different assignments of the filter coefficient λ in MSE.

that emphasizing the higher modulation frequency components enhances the VAD of MSE. Next, setting λ to 0.8 yields the optimal accuracy rate (77.04%), 0.10% better than the accuracy obtained by setting $\lambda = 0.7$ (76.94%). Finally, when the value of λ is within the range [0.1, 0.9], the differences among the accuracy rates obtained with different values of λ are relatively small, and the decrease in maximum accuracy is just 1.82%. This result implies that nearly optimal performance can be obtained without meticulous adjustment of the parameter λ.

The effect of processing the short pauses within the utterance in MSE.

In the VAD procedure of MSE, each frame in an utterance is always classified as either speech or non-speech. Therefore, no frame will be classified as a "transient frame", as is the case for some more delicate VAD processes. In fact, the transient frames that exist in the short region between two connected acoustic units (which are often called "short pauses") are quite often classified as non-speech in MSE, and thus their magnitude spectrums are assigned as very small. For this reason, the VAD in MSE is unlike some conventional end-point detectors, in which only the onset and offset frames of an utterance are decided, while the inter-word or inter-syllable frames that often possess lower energy are not processed. However, we find that further processing of these detected short pauses between the onset and offset times for utterances is quite helpful in speech recognition, especially when the SNR is low. To demonstrate this phenomenon, a simpler form of MSE is designed, in which we only process the first and the last detected non-speech segments (the corresponding frames are assigned to have very small magnitude spectra) and treat the remaining non-speech segments as speech (the magnitude spectra of the corresponding frames are weighted as in Equation (24)). This method is called "$\text{MSE}^{(s)}$" here for simplicity, and we compare it with the original MSE with respect to speech recognition performance.

Figure 14 shows the recognition accuracy rates for $\text{MSE}^{(s)}$ under different SNR conditions for the three Test Sets. In this figure, we see that almost no performance

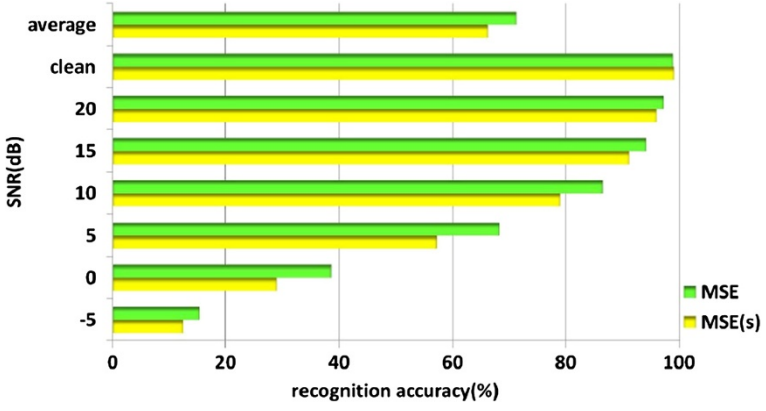

Figure 14 Recognition accuracy (%) (averaged over all the ten noise types in three Test Sets) achieved by MSE and $\text{MSE}^{(s)}$ for different SNR conditions.

Table 5 Recognition accuracy (%) achieved by various approaches for Aurora-2 multi-condition training task averaged across the SNRs between 0 and 20 dB, where AVG (%) and RR (%) are the averaged accuracy rate and the relative error rate reduction over the baseline

Method	Set A	Set B	Set C	AVG	RR
MFCC baseline	86.10	86.05	83.88	85.64	–
$SS_{Berouti}$	83.66	84.00	82.93	83.65	-13.86
WF_{PSNR}	83.96	84.50	83.03	83.99	-11.49
MMSE log-STSA	81.21	82.62	80.82	81.69	-27.48
MSE	87.91	87.41	82.21	86.57	6.49
MVN	90.38	90.41	89.82	90.28	32.31
$SS_{Berouti}$+MVN	86.89	87.91	85.72	87.06	9.91
WF_{PSNR}+MVN	84.78	85.24	84.57	84.92	-5.00
MMSE log-STSA+MVN	86.99	86.82	85.99	86.72	7.53
MSE+MVN	90.00	89.59	87.01	89.24	25.07
HEQ	89.98	90.05	89.59	89.93	29.87
$SS_{Berouti}$+HEQ	87.38	88.21	86.23	87.48	12.83
WF_{PSNR}+HEQ	84.77	85.16	84.16	84.80	-5.84
MMSE log-STSA+HEQ	86.42	86.53	85.27	86.24	4.15
MSE+HEQ	89.78	90.03	87.74	89.47	26.67
MVA	90.97	91.04	90.85	90.98	37.19
$SS_{Berouti}$+MVA	88.14	88.69	87.37	88.21	17.90
WF_{PSNR}+MVA	85.80	85.68	85.38	85.67	0.20
MMSE log-STSA+MVA	87.40	87.17	86.59	87.14	10.46
MSE+MVA	90.69	89.75	88.28	89.83	29.17

difference exists between $MSE^{(s)}$ and MSE for the clean condition. However, when noise is present, $MSE^{(s)}$ always performs worse than MSE, and the performance difference becomes more significant as the SNR decreases. On average, $MSE^{(s)}$ is around 4% less effective in recognition accuracy than MSE. In general, in acoustic model training, a short pause model is trained to aid in word or syllable boundary determination and thus to improve the recognition accuracy. However, under noise-corrupted conditions, the short pause model becomes less helpful, as shown in the $MSE^{(s)}$ results. Furthermore, in MSE, we see that to further classify the transient frames as non-speech (the corresponding magnitude spectrum then becomes very small) within the voice-activated region of an utterance significantly improves the recognition accuracy for noise-corrupted environments.

Results for the task of multi-condition training and multi-condition testing

We perform the MSE method, $SS_{Berouti}$ (which performs the best among the three SS methods in Table 2), WF_{PSNR} (which performs the best among the three WF methods in Table 2) and three cepstral-domain methods aforementioned with the Aurora-2 database under the mode of "multi-condition training". As stated earlier, here the training data have five SNR conditions (clean, 20, 15, 10, and 5 dB) and four types of noise the same as those in Test Set A. In addition to the individual method, here we also investigate the effect of the pairing of the spectral-domain method and the cepstral-domain method to see if further accuracy improvement can be achieved. Table 5 presents the individual set recognition accuracy rates averaged over five SNR conditions for Test Sets A, B, and C, achieved by the various methods. We have the following findings from Table 5:

1. For the spectral-domain methods, $SS_{Berouti}$ and WF_{PSNR} degrade the accuracy of the MFCC. The proposed MSE provides the MFCC with around 1% accuracy improvement for Test Sets A and B (additive-noise environments), but it still worsens the recognition accuracy for Test Set C (with both additive noise and channel distortion). A possible explanation is that, these spectral-domain methods introduce distortions and mitigate the discriminative components in the speech features when they alleviate the noise effect in the multi-condition training data.

2. In contrast with the spectral-domain methods, the three cepstral-domain methods can give significant performance improvement over the MFCC baseline. MVA behaves the best, followed by MVN and then HEQ. We find that MVN outperforms HEQ slightly, which is not the case for the mode of clean-condition training as shown in Table 2. This phenomenon is probably because the mismatch between the training data and the testing data is relatively small in the mode of multi-condition training, and the over-normalization problem may occur in HEQ, which results in worse accuracy relative to MVN.
3. None of the three spectral-domain methods, $SS_{Berouti}$, WF_{PSNR} and MSE, can help the subsequent cepstral-domain method to provide better recognition accuracy rates in comparison with the single cepstral-domain method. These results again imply these spectral-domain methods very probably diminish the helpful speech components in the noisy training data and are inappropriate for the task of multi-condition training.

Experiments for the Num-100A database

Besides the Aurora-2 database, here we adopt another database, called NUM-100A [40], to test the performance of the presented MSE. The NUM-100A database consists of 8,000 Mandarin digit strings produced by 50 male and 50 female speakers, recorded in a normal laboratory environment at an 8 kHz sampling rate. These 8000 digit strings include 1000 each of two-, three-, four-, five-, six-, and seven-digit strings, respectively, plus 2000 single digit utterances. Among the 8000 Mandarin digital strings, 7520 with different lengths are selected for training, while the other 480 are for testing. In particular, the 480 clean testing strings are added with four types of noise (white, babble, pink and f16) taken from the NOISEX-92 database [44] at four different SNRs (20, 15, 10, and 5 dB) to produce the noise-corrupted testing data. The speech features used here are the same as those in the Aurora-2 task, which contain 13 MFCCs ($c1$–$c12$, $c0$) and their delta and delta-delta. With the feature vectors in the training set, the HMMs for each of the 10 digits and silence were trained with the HTK toolkit [45]. Each digit HMM contains five states and eight mixtures per state, and the silence HMM has three states and eight mixtures per state.

For simplicity, we use the MSE with the same parameter settings in Aurora-2 task to process the training and testing signals and to create the corresponding MFCC features. In addition, since we just intend to investigate if MSE is also helpful to improve the noisy speech recognition for another database besides Aurora-2, we do not perform the other spectral-domain methods like SS and WF, and simply choose one cepstral-domain method, MVN, for processing the MFCC features.

Figures 15 and 16a–d show the recognition accuracy rates for the four methods, MFCC baseline, MSE, MVN and the pairing of MSE and MVN, under the clean and four noise-corrupted situations with different SNRs. From these figures, we have the following findings:

1. Under the clean and matched condition, both MSE and MVN degrades the recognition rate of the MFCC slightly, and the combination of MSE and MVN gets the worst results. These results imply that the robustness methods can probably reduce the discriminability of the original features when the environment is noise-free.
2. The recognition accuracy of the original MFCC gets apparently worse at mismatched noisy situations. However, the presented MSE can enhance the MFCC and bring about significant accuracy improvement irrespective of the type of noise. For example, at the SNR of 10 dB, MSE provides the MFCC with the accuracy rate improvements of 20.09%, 53.42%, 26.77%, and 41.74% for the noise being white, babble, pink and f16, respectively. Therefore, we show that MSE works well as a noise-robustness approach for this Mandarin digit database in addition to Aurora-2.
3. MVN promotes the recognition accuracy very well relative to the MFCC baseline when the environment is noisy, and it outperforms MSE in most cases. However, the cascade of MSE and MVN performs better than MVN alone (except for the babble and f16 noises at the SNR of 20 dB), showing again that MSE is well additive to the cepstral-domain method, MVN.

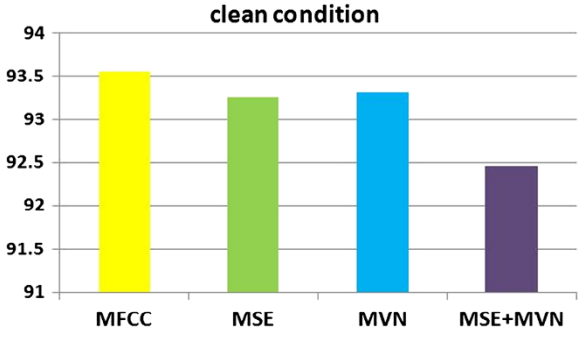

Figure 15 Recognition accuracy (%) achieved by various approaches for the NUM-100A database under the noise-free environment.

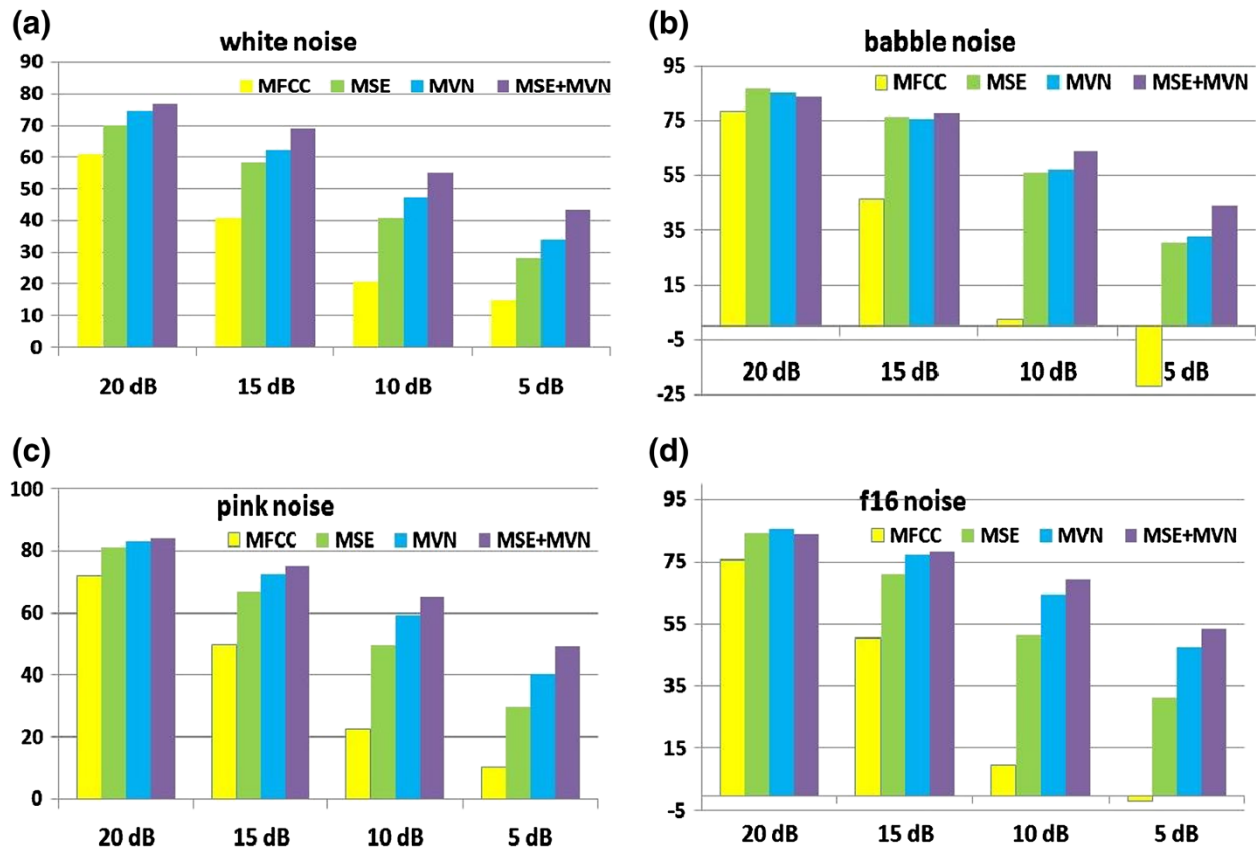

Figure 16 Recognition accuracy (%) achieved by various approaches for the NUM-100A database under the environments with additive noise being (a) white, (b) babble, (c) pink, (d) f16, respectively, at four SNR levels.

Conclusions

In this article, we investigate the effect of additive noise on the linear and logarithmic spectra of noise-corrupted utterances and provide a compensation scheme, called magnitude spectrum enhancement (MSE), to enhance the noise robustness of speech features. MSE aims to shrink the magnitude spectra in the silence portion of an utterance and to strengthen them in the speech portion. Experimental results show that MSE is very effective in promoting recognition performance under various noise conditions for the Aurora-2 clean-condition training task, and its performance is greater than that of spectral subtraction and Wiener filtering. Furthermore, MSE can successfully be implemented additively to cepstral-domain methods to deliver even better recognition rates.

Appendix 1

Given that $A = |A|e^{j\phi}$ is a complex-valued constant, and $N = N_R + jN_I$ is a complex-valued random variable which real and imaginary parts, N_R and N_I, are independent Gaussian distributed with zero mean and a common variance σ^2, then it can be shown that [38]:

1. The random variable $|A + N|$ is Rician distributed, and its probability density function (pdf) is

$$f_{|A+N|}(x) = \frac{x}{\sigma^2}\exp\left(\frac{-(x^2+|A|^2)}{2\sigma^2}\right) I_0\left(\frac{|A|^2}{\sigma^2}x\right) u(x), \tag{27}$$

   where $I_0(.)$ is the modified Bessel function of the first kind with order zero, and $u(.)$ is the unit-step function.
2. The random variable $|N|$ is Rayleigh distributed, and its probability density function (pdf) is

$$f_{|N|}(x) = \frac{x}{\sigma^2}\exp\left(\frac{-x^2}{2\sigma^2}\right) u(x). \tag{28}$$

Therefore, the items $|S_p[k] + D_p[k]|$ and $|D_q[k]|$ in Equation (3) are Rician and Rayleigh distributed,

respectively. Furthermore, assuming $D_p[k]$ and $D_q[k]$ are statistically independent (since they correspond to different frames) and identically distributed, we have Equation (3) as

$$\begin{aligned}\gamma[k] &= E\left(\frac{|S_p[k]+D_p[k]|}{D_q[k]}\right)\\ &= E\left(\frac{1}{|D_q[k]|}\right)E(|S_p[k]+D_p[k]|)\\ &= \left(\int_0^\infty \frac{1}{x}\left(\frac{x}{\sigma^2}\exp\left(\frac{-x^2}{2\sigma^2}\right)\right)dx\right)\\ &\quad\times\left(\sigma\sqrt{\frac{\pi}{2}}{}_1F_1\left(-\frac{1}{2};1;-\frac{|S_p[k]|^2}{2\sigma^2}\right)\right)\\ &= \left(\sqrt{\frac{\pi}{2}}\frac{1}{\sigma}\right)\left(\sigma\sqrt{\frac{\pi}{2}}{}_1F_1(-\frac{1}{2};1;-\frac{|S_p[k]|^2}{2\sigma^2})\right)\\ &= \frac{\pi}{2}{}_1F_1\left(-\frac{1}{2};1;-\frac{|S_p[k]|^2}{2\sigma^2}\right),\end{aligned} \tag{29}$$

where ${}_1F_1(.,.,.)$ is the confluent hypergeometric function [38]:

$${}_1F_1\left(-\frac{1}{2};1;-x\right) = \exp\left(-\frac{x}{2}\right)\left((1+x)I_0\left(\frac{x}{2}\right)+xI_1\left(\frac{x}{2}\right)\right), \tag{30}$$

in which $I_1(.)$ is the modified Bessel function of the first kind with order one.

It can be shown that

$${}_1F_1\left(-\frac{1}{2};1;-x\right) > 0,\ \text{for} x > 0 \tag{31}$$

and since $\frac{d}{dx}I_0(x) = I_1(x)$ and $\frac{d}{dx}(xI_1(x)) = xI_0(x)$, we have

$$\begin{aligned}&\frac{d}{dx}\left({}_1F_1\left(-\frac{1}{2};1;-x\right)\right)\\ &\quad= \exp\left(-\frac{x}{2}\right)\left(\frac{1}{2}I_0\left(\frac{x}{2}\right)+\frac{3}{2}I_1\left(\frac{x}{2}\right)\right)>0,\ \text{ for } x>0.\end{aligned} \tag{32}$$

Therefore, ${}_1F_1(-\frac{1}{2};1;-x)$ is a positive and monotonically increasing function for $x > 0$, and we conclude that the parameter $\gamma[k] = \frac{\pi}{2}{}_1F_1(-\frac{1}{2};1;-\frac{|S_p[k]|^2}{2\sigma^2})$ in Equation (29) decreases as the noise variance σ^2 increases (with decreasing $\frac{1}{\sigma^2}$) , with two limiting cases $\lim_{\sigma^2\to 0}\gamma[k] = \infty$ and $\lim_{\sigma^2\to\infty}\gamma[k] = \frac{\pi}{2}$.

Competing interests

The authors declare that they have no competing interests.

Received: 15 February 2012 Accepted: 13 August 2012
Published: 30 August 2012

References

1. JN Holmes, NC Sedgwick, Noise compensation for speech recognition using probabilistic models. in *1986 International Conference on Acoustics, Speech and Signal Processing (ICASSP'86)*, vol. 11 (Tokyo, Japan, 1986), pp. 741–744
2. DH Klatt, A digital filter bank for spectral matching. in *1979 International Conference on Acoustics, Speech and Signal Processing (ICASSP'76)*, vol. 1 (Philadelphia, USA, 1976), pp. 573–576
3. A Nadas, D Nahamoo, M Picheny, Speech recognition using noise-adaptive prototypes. in *1988 International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)*, vol. 1 (New York, USA, 1988), pp. 517–520
4. AP Varga, RK Moore, Hidden Markov model decomposition of speech and noise. in *1990 International Conference on Acoustics, Speech and Signal Processing (ICASSP'90)*, vol. 2 (Albuquerque, USA, 1990), pp. 845–848
5. A Acero, L Deng, T Kristjansson, J Zhang, HMM adaptation using vector Taylor series for noisy speech recognition. in *2000 International Conference on Spoken Language Processing (ICSLP'00)*, vol. 3 (Beijing, China, 2000), pp. 869–872
6. CJ Leggester, PC Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput. Speech Lang. **9**, 171–186 (1995)
7. A Sankar, C-H Lee, A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. **4**, 190–202 (1996)
8. C-H Lee, On stochastic feature and model compensation approaches to robust speech recognition. Speech Commun. **25**, 29–47 (1998)
9. G-X Ning, G Wei, K-K Chu, Model compensation approach based on nonuniform spectral compression features for noisy speech recognition. EURASIP J. Adv. Signal Process. **2007** (2007)
10. PJ Moreno, B Raj, RM Stern, Data-driven environmental compensation for speech recognition: a unified approach. Speech Commun. **24**, 267–285 (1998)
11. MJF Gales, SJ Young, Cepstral parameter compensation for HMM recognition in noise. Speech Commun. **12**, 231–239 (1993)
12. MJF Gales, SJ Young, Robust speech recognition in additive and convolutional noise using parallel model combination. Comput. Speech Lang. **9**, 289–307 (1995)
13. MJF Gales, SJ Young, A fast and flexible implementation of parallel model combination. in *1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP'95)*, vol. 1 (Detroit, USA, 1995), pp. 133–136
14. SF Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**, 113–120 (1979)
15. M Berouti, R Schwartz, J Makhoul, Enhancement of speech corrupted by acoustic noise. in *1979 International Conference on Acoustics, Speech and Signal Processing (ICASSP'79)*, vol. 4 (Washington, USA, 1979), pp. 208–211
16. S Kamath, P Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. in *2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, vol. 4 (Orlando, USA, 2002), pp. IV–4164
17. B BabaAli, H Sameti, M Safayani, Likelihood maximizing based multi-band spectral subtraction for robust speech recognition. EURASIP J. Adv. Signal Process. **2009** (2009)
18. P Scalart, JV Filho, Speech enhancement based on a priori signal to noise estimation. in *1996 International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, vol. 2 (Atlanta, USA, 1996), pp. 629–632
19. C Plapous, C Marro, P Scalart, Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **14**, 2098–2108 (2006)
20. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**, 1109–1121 (1984)

21. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**, 443–445 (1985)
22. A Acero, Acoustical and environmental robustness in automatic speech recognition, *Ph.D. dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University*, Pittsburg, PA (1990)
23. KK Chu, SH Leung, SNR-dependent non-uniform spectral compression for noisy speech recognition. in *2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1 (Montreal, Canada, 2004), pp. 973–976
24. L Deng, A Acero, L Jiang, J Droppo, X Huang, High-performance robust speech recognition using stereo training data. in *2001 International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, vol. 1 (Salt Lake City, USA, 2001), pp. 301–304
25. J Droppo, L Deng, A Acero, Evaluation of the SPLICE algorithm on the Aurora2 database. in *2001 Eurospeech Conference on Speech Communications and Technology (Eurospeech'01)*, vol. 1 (Aalborg, Denmark, 2001), pp. 185–188
26. BS Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. **55**, 1304–1312 (1974)
27. S Tibrewala, H Hermansky, Multiband and adaptation approaches to robust speech recognition. in *1997 Eurospeech Conference on Speech Communications and Technology (Eurospeech'97)*, vol. 1 (Rhodes, Greece, 1997), pp. 2619–2622
28. C-P Chen, JA Bilmes, MVA processing of speech features. IEEE Trans. Audio Speech Lang. Process. **15**, 257–270 (2007)
29. S Yoshizawa, N Hayasaka, N Wada, Y Miyanaga, Cepstral gain normalization for noise robust speech recognition. in *2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1 (Montreal, Canada, 2004), pp. 209–212
30. F Hilger, H Ney, Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **14**, 845–854 (2006)
31. Y Suh, H Kim, Histogram equalization to model adaptation for robust speech recognition. EURASIP J. Adv. Signal Process. **2010** (2010)
32. J Du, R-H Wang, Cepstral shape normalization (CSN) for robust speech recognition. in *2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, vol. 1 (Las Vegas, USA, 2008), pp. 4389–4392
33. W Zhu, D O'Shaughnessy, Log-energy dynamic range normalization for robust speech recognition. in *2005 International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, vol. 1 (Philadelphia, USA , 2005), pp. 245–248
34. T-H Hwang, S-C Chang, Energy contour enhancement for noisy speech recognition. in *2004 International Symposium on Chinese Spoken Language Processing (ISCSLP'04)*, vol. 1 (Hong Kong, China, 2004), pp. 249–252
35. C-C Wang, C-A Pan, J-W Hung, Silence feature normalization for robust speech recognition in additive noise environments. in *2008 International Conference on Spoken Language Processing (Interspeech 2008-ICSLP)*, vol. 1 (Brisbane, Australia, 2008), pp. 1028–1031
36. W-H Tu, J-W Hung, Magnitude spectrum enhancement for robust speech recognition. in *2010 International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*, vol. 1 (Dallas, USA, 2010), pp. 4586–4589
37. HG Hirsch, D Pearce, The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, vol. 1 (Paris, France, 2000), pp. 181–188
38. S Haykin, *Communication Systems*, 4th edn. (John Wiley & Sons, Inc., New York, 2000)
39. GL Turin, An introduction to matched filters. IRE Trans. Inf. Theory. **6**, 311–329 (1960)
40. Available from: http://www.aclclp.org.tw, the Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
41. ITU recommendation G.712, Transmission performance characteristics of pulse code modulation channels (1996)
42. Available from: http://www.elda.org/article52.html, Evaluations and Language resources Distribution Agency (ELDA)
43. RV Hogg, EA Tanis, *Probability and Statistical Inference*, 7th edn. (Prentice Hall, Upper Saddle River, NJ, 2006)
44. AP Varga, HJM Steeneken, M Tomlinson, D Jones, The NOISEX-92 study on the effect of additive noise on automatic speech recognition, *Tech. Rep. DRA Speech Research Unit*, 1992
45. Available from: http://htk.eng.cam.ac.uk/, the Hidden Markov Model Toolkit, Cambridge University Engineering Dept. (CUED)

doi:10.1186/1687-6180-2012-189
Cite this article as: Hung *et al.*: **Enhancing the magnitude spectrum of speech features for robust speech recognition.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:189.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com