Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Speech Enhancement Based on Estimating Expected Values of Speech Cepstra

Chengshi Zheng, Yi Zhou, Xiaohu Hu, Jing Tian, and Xiaodong Li

Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

PACS: 43.72

ABSTRACT

This paper proposes a novel speech enhancement (SE) algorithm based on estimating expected values of speech cepstra (EVSC), which will be herein referred as EVSC-SE. Unlike the conventional SE algorithms, where the *a priori* signal-to-noise-ratio (SNR) is estimated from expected values of speech spectra (EVSS) directly, the proposed EVSC-SE algorithm estimates the *a priori* SNR from the EVSC. Under the Gaussian assumption of speech signals, we propose two approaches to estimate the EVSC. One is a novel cepstral subtraction approach, which is the estimation-based approach. The other is a modified cepstrum thresholding approach, which is the detection-based approach. Compared with conventional EVSS-based SE (EVSS-SE) algorithms, the proposed EVSC-SE algorithm is capable of tracking the *a posteriori* SNR at word onsets and offsets rapidly, achieving less speech distortion. Moreover, the EVSC-SE algorithm could suppress non-stationary noise effectively. Simulation results show that the EVSC-SE algorithm outperforms the conventional EVSS-SE algorithms in terms of segmental SNR and log-spectral distance.

INTRODUCTION

The speech enhancement algorithms based on the *decision-directed* (DD-SE) approach of Ephraim and Malah and its improved methods [1-4] are efficient ways to eliminate the *musical noise* problem. Recent studies show that the speech enhancement algorithms based on cepstral smoothing (CS-SE) techniques [5,6] outperform the DD-SE approach in several aspects, such as the output segmental signal-to-noise-ratio (SNR), the noise reduction in non-stationary noise, and the speech distortion. The main reason is that the CS-SE could selectively smooth noise cepstra based on the relaxed model that the speech and the noise contribute to different parts of cepstra, while the DD-SE does not distinguish the noise spectral components from the speech spectral components, so the DD-SE is sensitive to high *a posteriori* SNR even in noise-only regions. The adaptive smoothing factors in the CS-SE rely on the voice activity detection (VAD)or the voice/unvoiced (V/UV) decision, so the CS-SE is somewhat empirical. Besides, there is still a lack of a theoretical explanation of the better performances of the CS-SE.

In [7-9], the speech presence probability (SPP) in each time-frequency point is used to improve the performances of the DD-SE. Most of the SPP estimators depend on the *a priori* SNR or/and the *a posteriori* SNR estimation, and often lead to overestimate the SPP when the noise power spectral density (NPSD) is underestimated for the non-stationary noise, such as babble noise. It is common to underestimate the NPSD for rapid increasing noise levels by the minimum statistics (MS) method of Martin [10] and the minima controlled recursive averaging (MCRA) method of Cohen [8], although several NPSD estimation algorithms for highly non-stationary noise environments have been proposed [11,12].

This paper gives a new insight into the *a priori* SNR estimators, including the maximum likelihood (ML) approach [1], the *decision-directed* (DD) approach [1], and the two cepstral smoothing techniques [5,6]. We reveal that the herein four estimators except the ML estimator could be approximately seen as trying to estimate expected values of speech spectra (EVSS), where the DD approach estimates in the frequency domain directly and the two cepstral smoothing techniques estimate in the cepstral domain indirectly. The relationship brings out a novel *a priori* SNR estimator, which is based on estimating expected values of speech cepstra (EVSC). Considering the property and the second-order statistics of speech cepstra under the Gaussian assumption, we propose two algorithms to estimate the EVSC. One is a novel cepstral subtraction method, of which the gain function is computed by two steps. After obtaining the gain function using oversubtraction method, it is adaptively smoothed over time according to the gain value. The other is the detection-based algorithm, which can be seen as a modified cepstrum thresholding method (MCT)[13,14]. Simulation results confirm validity and high performance. Our proposed approaches are comparable with the CS-SE at low input SNR and much better than the CS-SE at high input SNR. Besides, our approaches do not need the VAD or the V/UV decision, and have the least speech distortion at speech onsets due to its capability to track large values of cepstra fast, while the CS-SE smoothes the cepstra by constant factors even when they are extremely large at low quefrencies.

BACKGROUND AND THE RELATIONSHIP OF THE A *PRIROI* SNR ESTIMATORS

Background

Let $y(n)$, $s(n)$ and $d(n)$ denote noisy speech, clean speech and additive uncorrelated noise, respectively, where n is a discrete-time index. Noisy speech can be expressed by $y(n) = s(n) + d(n)$. In the cepstral domain, Sagayama *et al* show the relationship between cepstra of speech, noise, and speech plus noise as follows [15]

$$\mathbf{C}_Y(l) = \mathbf{F}^* \left[\log\left\{\exp\left(\mathbf{F}\mathbf{C}_S(l)\right) + \exp\left(\mathbf{F}\mathbf{C}_D(l)\right)\right\}\right] \quad (1)$$

where l is the frame index, $\mathbf{F}$ and $\mathbf{F}^*$ denote the $N \times N$ Fourier and inverse Fourier transform matrices, respectively, where N is the frame length. $\mathbf{C}_Y(l)$, $\mathbf{C}_S(l)$, and $\mathbf{C}_D(l)$ are noisy speech

cepstrum, clean speech cepstrum, and noise cepstrum, respectively. The short-time Fourier transform (STFT) domain of $y(n)$, $s(n)$ and $d(n)$ are given by

$$\mathbf{Y}(l)=\mathbf{Fy}(l),\mathbf{S}(l)=\mathbf{Fs}(l),\quad \mathbf{D}(l)=\mathbf{Fd}(l) \tag{2}$$

where $\mathbf{y}(l)$, $\mathbf{s}(l)$, and $\mathbf{d}(l)$ are the lth frame of consecutive samples of $y(n)$, $s(n)$, and $d(n)$ multiplied by a Hanning window, respectively. The periodograms can be computed by $\mathbf{Y}_p(l)=\mathbf{Y}(l)\circ\mathbf{Y}^*(l)$, $\mathbf{S}_p(l)=\mathbf{S}(l)\circ\mathbf{S}^*(l)$, $\mathbf{D}_p(l)=\mathbf{D}(l)\circ\mathbf{D}^*(l)$, respectively, where "$\circ$" is the Hadamard product, and "$*$" is the complex conjugate operator. Their cepstra are related to their periodograms by $\mathbf{F}[\mathbf{C}_Y(l)]=\log\left(\mathbf{Y}_p(l)\right)$, $\mathbf{F}[\mathbf{C}_S(l)]=\log\left(\mathbf{S}_p(l)\right)$, $\mathbf{F}[\mathbf{C}_D(l)]=\log\left(\mathbf{D}_p(l)\right)$, respectively.

Generally, the inequation, $\mathbf{C}_Y(l)\neq\mathbf{C}_S(l)+\mathbf{C}_D(l)$, holds. Thus, the conventional cepstral subtraction method [16] conflicts with the model presented in (1). The method is effective in broadband noise, the reason may be that cepstra of broadband noise are very close to zero if ignoring their variances. However, the method may cause serious speech distortion due to the large values of the very narrowband noise cepstra. The CS-SE does not use any explicit model. The basic assumption in the CS-SE is that cepstra of noisy speech can be decomposed into cepstra related to the speech envelop, the excitation, and noise. Thus, strong smoothing is applied to noise cepstra, and little smoothing is applied to most likely speech cepstra. The assumption is somewhat empirical. Besides, the smoothing factors at low quefrencies are set to constant values, which may cause more speech distortion at speech onsets and lead to lower noise reduction for white noise. This is because cepstra of white noise are all close to zero except $C_D(0,l)$, if larger values of smoothing factors are chosen at the lower cepstral bins for white noise, better performances will be achieved. In the following part, the relationship between the DD approach and the two cepstral smoothing techniques is discussed in detail.

The relationship of the *a priori* SNR estimators

In the frequency domain, $y(n)=s(n)+d(n)$ is given by

$$Y(k,l)=S(k,l)+D(k,l) \tag{3}$$

where $k=0,\cdots,N-1$ is the frequency bin index; $\{Y(k,l)\}_{k=0}^{N-1}\in\mathbf{Y}(l)$, $\{S(k,l)\}_{k=0}^{N-1}\in\mathbf{S}(l)$, and $\{D(k,l)\}_{k=0}^{N-1}\in\mathbf{D}(l)$. The *a priori* SNR, $\xi(k,l)$, is defined as

$$\xi(k,l)=\frac{E\left\{|S(k,l)|^2\right\}}{E\left\{\lambda_d(k,l)\right\}}\approx\frac{E\left\{\left|\hat{S}(k,l)\right|^2\right\}}{\lambda_d(k,l)} \tag{4}$$

where $E\{\bullet\}$ is an expectation function, $\lambda_d(k,l)$ is an estimate of the NPSD, and $\left|\hat{S}(k,l)\right|^2=\max\left\{|Y(k,l)|^2-\beta\lambda_d(k,l),0\right\}$ is an estimate of $|S(k,l)|^2$, where β ($\beta\geq1$) is an oversubtraction factor [17]. We use the approximation in (4) due to the limited performances of most NPSD estimation algorithms in highly non-stationary environments and the estimation errors of the clean speech spectra.

The ML and the DD approaches

Define the *a posteriori* SNR, $\gamma(k,l)$, as follows:

$$\gamma(k,l)=\frac{|Y(k,l)|^2}{\lambda_d(k,l)}. \tag{5}$$

The ML approach and the DD approach are given by [1]

$$\xi_{ml}(k,l)=P[\gamma(k,l)-1] \tag{6}$$

$$\xi_{dd}(k,l)=\alpha_{dd}\frac{\left|S_{dd,s}(k,l-1)\right|^2}{\lambda_d(k,l-1)}+(1-\alpha_{dd})\,\xi_{ml}(k,l) \tag{7}$$

where $P[x]=x$ if $x\geq0$, and $P[x]=0$ otherwise; α_{dd} is a smoothing factor, and $S_{dd,s}(k,l-1)$ corresponds to an estimate of the clean speech in the previous frame. (7) can be rewritten as

$$\frac{|S_{dd}(k,l)|^2}{\lambda_d(k,l)}=\alpha_{dd}\frac{\left|S_{dd,s}(k,l-1)\right|^2}{\lambda_d(k,l-1)}+(1-\alpha_{dd})\frac{\left|\hat{S}(k,l)\right|^2}{\lambda_d(k,l)} \tag{8}$$

Assuming $\lambda_d(k,l-1)=\lambda_d(k,l)$, multiplying both sides of the above equation by $\lambda_d(k,l)$ yields

$$\begin{aligned}&E\left\{\left|\hat{S}(k,l)\right|^2\right\}\approx|S_{dd}(k,l)|^2\\&=\alpha_{dd}\left|S_{dd,s}(k,l-1)\right|^2+(1-\alpha_{dd})\left|\hat{S}(k,l)\right|^2.\end{aligned} \tag{9}$$

Eq. (9) indicates that the DD approach can be approximately seen as estimating the *a priori* SNR by using the EVSS. It needs to be pointed out that $\left|S_{dd,s}(k,l)\right|^2$ is not equal to $|S_{dd}(k,l)|^2$:

$$\left|S_{dd,s}(k,l)\right|^2=|G_{dd}(k,l)\,Y(k,l)|^2 \tag{10}$$

where the gain function, $G_{dd}(k,l)$, is determined by the *a priori* SNR and the *a posteriori* SNR.

Cepstral smoothing of the ML speech spectra estimate

In [6], the *a priori* SNR estimation based on cepstral smoothing of the ML speech spectra estimate, $\xi_{ceps,1}(k,l)$, is computed as

$$\xi_{ceps,1}(k,l)=\frac{\exp\left(0.5\kappa+\mathtt{FFT}\left\{\bar{C}_{\hat{S}}(q,l)\right\}\right)}{\lambda_d(k,l)} \tag{11}$$

where $\kappa=0.5772156966490$ is the Euler constant [6, 13, 14, 18], and $C_{\hat{S}}(q,l)$ is the cepstral representation of $\left|\hat{S}(k,l)\right|^2$, where $q=0,\cdots,N-1$ is the cepstral bin index. In [6], $\bar{C}_{\hat{S}}(q,l)$ is given by

$$\bar{C}_{\hat{S}}(q,l)=\alpha(q,l)\bar{C}_{\hat{S}}(q,l-1)+(1-\alpha(q,l))C_{\hat{S}}(q,l) \tag{12}$$

where the smoothing factor, $\alpha(q,l)$, is chosen by the V/UV decision.

Under the Gaussian assumption of speech signals, the relationship between speech cepstra and speech spectra is given by

$$E\left\{\left|\hat{S}(k,l)\right|^2\right\}=\exp\left(\kappa_k+\mathrm{FFT}\left\{E\left\{C_{\hat{S}}(q,l)\right\}\right\}\right) \tag{13}$$

where $\kappa_k=\kappa+\left(\delta_k+\delta_{k-N/2}\right)\log2$, and δ_k is the Dirac function. If ignoring the difference between κ_k and κ, $\xi_{ceps,1}(k,l)$ can be seen as estimating the EVSS by using the EVSC. Obviously, $\bar{C}_{\hat{S}}(q,l)$ is an estimate of the EVSC.

Cepstral Smoothing of the Gain Function

In [5], the cepstral smoothing technique is applied to the gain function, $G_{dd}(k,l)$. The time-domain representation of $G_{dd}(k,l)$ is $g_{dd}(n)$, then the estimated clean speech in the time domain is given by

$$s_{dd}(n)=y(n)\otimes g_{dd}(n) \tag{14}$$

where $s_{dd}(n)=\mathtt{IFFT}\{S_{dd}(k,l)\}$, and $\otimes$ is the convolution operator. In the cepstral domain, (14) can be rewritten as

$$C_{s_{dd}}(q,l)=C_y(q,l)+C_{g_{dd}}(q,l). \tag{15}$$

where $C_{s_{dd}}(q,l)$, $C_y(q,l)$, and $C_{g_{dd}}(q,l)$ are the cepstral representations of $|S_{dd}(k,l)|^2$, $|Y(k,l)|^2$, and $G_{dd}(k,l)$, respectively. The final gain function is obtained by

$$G_{ceps,2}(k,l)=\exp\left(\mathtt{FFT}\left\{\bar{C}_{g_{dd}}(q,l)\right\}\right. \tag{16}$$

where $\bar{C}_{g_{dd}}(q,l)$ is calculated as

$$\bar{C}_{g_{dd}}(q',l)=\beta_{ceps}\bar{C}_{g_{dd}}(q',l-1)+\left(1-\beta_{ceps}\right)C_{g_{dd}}(q',l) \tag{17}$$

where β_{ceps} is a constant smoothing factor, and q' is determined by the VAD and the index of the maximum value of gain cepstrum, $C_{g_{dd}}(q,l)$, in the range that corresponds to all the possible pitch periods. $\bar{C}_{g_{dd}}(q,l)$ can be seen as an estimate of the expected values of gain cepstra, which is

$$E\left\{C_{g_{dd}}(q,l)\right\} \approx \bar{C}_{g_{dd}}(q,l) \tag{18}$$

Substituting (18) and (15) into (16), and assuming $G_{ceps,2}(k,l)$ is obtained by Wiener filter, (16) is rewritten as

$$\frac{\xi_{ceps,2}(k,l)}{\xi_{ceps,2}(k,l)+1} = \exp\left(\mathtt{FFT}\left\{E\left\{C_{s_{dd}}(q,l)\right\} - E\left\{C_y(q,l)\right\}\right\}\right) \tag{19}$$

If we assume that $y(n)$ and $s_{dd}(n)$ are independent Gaussian distributions, then

$$\frac{\exp\left(\mathtt{FFT}\left\{E\left\{C_{s_{dd}}(q,l)\right\}\right\}\right)}{\exp\left(\mathtt{FFT}\left\{E\left\{C_y(q,l)\right\}\right\}\right)} = \frac{E\left\{|S_{dd}(k,l)|^2\right\}}{E\left\{|Y(k,l)|^2\right\}} \tag{20}$$

Substituting (20) into (19), the *a priori* SNR, $\xi_{ceps,2}(k,l)$, could be obtained

$$\xi_{ceps,2}(k,l) = \frac{E\left\{|S_{dd}(k,l)|^2\right\}}{E\left\{|Y(k,l)|^2\right\} - E\left\{|S_{dd}(k,l)|^2\right\}}. \tag{21}$$

For $\lambda_d(k,l) \approx E\left\{|Y(k,l)|^2\right\} - E\left\{|S_{dd}(k,l)|^2\right\}$, (21) becomes

$$\xi_{ceps,2}(k,l) = \frac{E\left\{|S_{dd}(k,l)|^2\right\}}{\lambda_d(k,l)} = \xi(k,l). \tag{22}$$

Eq. (22) demonstrates that $\xi_{ceps,2}(k,l)$ is obtained by estimating the expected values of $|S_{dd}(k,l)|^2$, which can be computed by the EVSC as shown in (20).

By now, we could summarize that both the DD and the cepstral smoothing techniques estimate the *a priori* SNR by estimating the EVSS. The only difference is that the DD approach estimates the EVSS in the frequency domain, while the two cepstral smoothing techniques estimate it in the cepstral domain.

PROPOSED ALGORITHMS

Eq. (13) indicates that the EVSS can be obtained by estimating the EVSC. At the beginning of this section, the property of speech cepstra is briefly studied by the speech generation model. Then, two algorithms will be proposed to estimate the EVSC, which is based on the property and the second-order statistics under the Gaussian assumption of speech signals.

The property of speech cepstra

For voiced speech, $s(n)$ could be modeled as a convolution of a periodic impulse train $p(n)$ with pitch period n_p and the combined impulse response $v(n)$ of vocal cords, vocal tract,and radiation characteristics [19,20]

$$s(n) = p(n) * v(n). \tag{23}$$

It is well-known that $v(n)$ only contributes to low quefrencies and $p(n)$ only contributes to the high quefrencies corresponding to the integral multiple of n_p. Thus, the speech generation model gives a theoretical explanation why a large proportion of voiced speech cepstra have small values.

For unvoiced speech, $s(n)$ is given by

$$s(n) = r(n) * u(n) \tag{24}$$

where $r(n)$ is random noise, and $u(n)$ is the combined impulse response of vocal tract and radiation characteristics. Both $r(n)$ and $u(n)$ only contribute to expected values of low quefrencies, so most of unvoiced speech cepstra also have small values.

The property of speech cepstra, of which a large proportion have small values, has already been used in the harmonics-to-noise ratio (HNR) estimation [21]. The property is also the basis of our proposed approaches. The TIMIT database [22] is used to experimentally study the property of speech cepstra. To further show the differences between speech cepstra and noise cepstra, the Noisex92 database [23] is also used to study the property of noise cepstra.

Under Gaussian assumptions of $s(n)$ and $d(n)$, $\{C_S(q,l)\}_{q=0}^{N-1} \in \mathbf{C}_S(l)$ and $\{C_D(q,l)\}_{q=0}^{N-1} \in \mathbf{C}_D(l)$, have the different unknown mean values but the same known variances [13,14]

$$\begin{aligned}\chi_q^2 &= \mathrm{var}\left(C_{S|D}(q,l)\right) \\ &= E\left\{C_{S|D}^2(q,l)\right\} - \left[E\left\{C_{S|D}(q,l)\right\}\right]^2 \\ &= \left(1+\delta_q+\delta_{(q-N/2)}\right)\frac{\pi^2}{6N}; \quad q=0,1,\cdots,N/2\end{aligned} \tag{25}$$

where δ_q and $\delta_{(q-N/2)}$ are the Dirac functions, and $\mathrm{var}(\bullet)$ is the variance function; $C_{S|D}(q,l)$ means $C_S(q,l)$ or $C_D(q,l)$ for compact notation. When the frame length is $N=512$ for the sample rate $f_s = 16$kHz. The percentages of $\left|\tilde{C}_{S|D}(q,l)\right|^2 \le \hbar^2\chi_q^2$,with $\hbar^2 = 1,4,9$, for voiced speech signals and two types of noise signals(including the white noise and the babble noise), are shown in Table 1. $\tilde{C}_{S|D}(q,l)$ is a smoothed version of $C_{S|D}(q,l)$

$$\tilde{C}_{S|D}(q,l) = \sum_{i=-w}^{w} b(i)\, C_{S|D}(q-i,l) \tag{26}$$

where b is a normalized Hanning window function of length $2w+1$, i.e., $\sum_{i=-w}^{w} b(i) = 1$; and $\tilde{C}_{S|D}(q,l)$ means $\tilde{C}_S(q,l)$ or $\tilde{C}_D(q,l)$ for compact notation.

We have to explain the reason for using $\tilde{C}_S(q,l)$ and $\tilde{C}_D(q,l)$. If we assume the speech/noise cepstra over the $2w+1$ bins are independent and identically distributed (i.i.d.) random variables, both $\tilde{C}_S(q,l)$ and $\tilde{C}_D(q,l)$ are still Gaussian distributions and their variances are reduced by a constant factor $\bar{r}$, where $\bar{r}$ is determined by the window function b. The variances of $\tilde{C}_S(q,l)$ and $\tilde{C}_D(q,l)$ are give by

$$\begin{aligned}\tilde{\chi}_q^2 &= \mathrm{var}\left(\tilde{C}_{S|D}(q,l)\right) \\ &= E\left\{\tilde{C}_{S|D}^2(q,l)\right\} - \left[E\left\{\tilde{C}_{S|D}(q,l)\right\}\right]^2 \\ &= \bar{r}\cdot\chi_q^2\end{aligned} \tag{27}$$

where $\tilde{C}_{S|D}(q,l)$ means $\tilde{C}_S(q,l)$ or $\tilde{C}_D(q,l)$ for compact notation, and $\bar{r}$ is determined by the window function b

$$\bar{r} = \sum_{i=-w}^{w} [b(i)]^2 \tag{28}$$

Obviously, $\bar{r} \le 1$, so the variances are reduced by (26). To make the i.i.d. assumption valid, the window length should not be too large. In this paper, $w=2$ is used. Note that since the variances of $C_{S|D}(q,l)$ are reduced by (26), it is more meaningful to give the percentages of $\left|\tilde{C}_{S|D}(q,l)\right|^2 \le \tilde{\hbar}^2\tilde{\chi}_q^2$ than those of $\left|\tilde{C}_{S|D}(q,l)\right|^2 \le \hbar^2\chi_q^2$. Whereas, if $\hbar^2 = \bar{r}\tilde{\hbar}^2$ is used, the same results, as shown in Table 1, can be obtained.

Table 1 reveals that nearly 80% of voiced speech cepstra and over 88% of noise cepstra have small values, which satisfy

Table 1: Percentages of $\left|\tilde{C}_S(q,l)\right|^2 \le \hbar^2\chi_q^2$, and $\left|\tilde{C}_D(q,l)\right|^2 \le \hbar^2\chi_q^2$ for three different types of signals.

Signal Types	$\hbar^2=1$	$\hbar^2=4$	$\hbar^2=9$
Voiced Speech	79.81%	93.94%	96.83%
babble Noise	88.20%	96.95%	97.97%
White Noise	93.92%	99.46%	99.59%

$\left|\tilde{C}_S(q,l)\right| \le \chi_q$ and $\left|\tilde{C}_D(q,l)\right| \le \chi_q$. The results confirm that a large proportion of speech/noise cepstra are close to zero. Table 1 also indicates that speech cepstra contain more large values than the other two types of noise signals. There are three reasons for this phenomenon. First, the noise signals are not harmonic, their cepstra at high quefrencies are often small; While voiced speech cepstra often have large values at high quefrencies due to the periodic impulse train $p(n)$. Second, the smoothing operator in (26) broadens the rahmonic peaks resulting in more large values of voiced speech cepstra. Third, there are more large values of speech cepstra at low quefrencies than those of noise cepstra, especially when the noise is broadband with small dynamic range.

We further show a comparison of $|\tilde{C}_S(n_p,l)|$ and $|\tilde{C}_D(q_{\max,l},l)|$. The pitch period n_p could be achieved by the pitch detection algorithms, such as the fast adaptive representation-spectrum-based scheme and the FFT-spectrum-based scheme. Since the pitch period n_p is estimated from the clean speeches, we use the FFT-spectrum-based scheme for its moderate performance with high computational efficiency [24]. $q_{\max,l}$ is defined as

$$q_{\max,l} = \arg\max_{q\in\tilde{n}_p} \left(\left|\tilde{C}_D(q,l)\right|\right). \tag{29}$$

where $\tilde{n}_p = \left[\begin{array}{cc} f_s/f_{0,high} & f_s/f_{0,low} \end{array}\right]$ represents all the possible pitch periods n_p, and $f_{0,low} = 50$Hz and $f_{0,high} = 500$Hz are the minimum and the maximum fundamental frequencies, respectively. We define

$$\tilde{\eta}_S(n_p,l) = \left|\tilde{C}_S(n_p,l)\right| / \chi_{n_p}, \tag{30}$$

and

$$\tilde{\eta}_D(q_{\max,l},l) = \left|\tilde{C}_D(q_{\max,l},l)\right| / \chi_{q_{\max,l}}. \tag{31}$$

The distributions of $\tilde{\eta}_S(n_p,l)$ for voiced speech, and those of $\tilde{\eta}_D(q_{\max,l},l)$ for white noise and babble are shown in Figure 1, which demonstrates that most of $\tilde{\eta}_S(n_p,l)$ are much larger than $\tilde{\eta}_D(q_{\max,l},l)$. Thus, by selecting a proper thresholding at high quefrencies, it could distinguish voiced speech signals from unvoiced speech and noise signals. The characteristic has been widely used in several fields, such as the V/UV decision [6], the pitch determination [26,27], and the HNR estimation [21].

By now, we have summarized the two characteristics of speech cepstra or noise cepstra. First, a majority of speech cepstra or noise cepstra have small values. Second, $|\tilde{C}_S(n_p,l)|$ is much larger than $|\tilde{C}_D(q_{\max,l},l)|$ in most cases. In the following parts, we propose two algorithms to estimate the EVSC, which are based on the properties and the second-order statistics of speech cepstra under the Gaussian assumption.

To reduce the influences of the noise, it is better to remove the noise before estimating the EVSC. The spectral subtraction algorithm is used to estimate the clean speech spectra $|\hat{S}(k,l)|^2$ at the first step, then the EVSC are estimated by $|\hat{S}(k,l)|^2$.

A novel cepstral subtraction method

Substitute $C_{\hat{S}}(q,l)$ into (25), we obtain

$$\chi_q^2 = \mathrm{var}\left\{C_{\hat{S}}(q,l)\right\} = E\left\{C_{\hat{S}}^2(q,l)\right\} - \left[E\left\{C_{\hat{S}}(q,l)\right\}\right]^2. \tag{32}$$

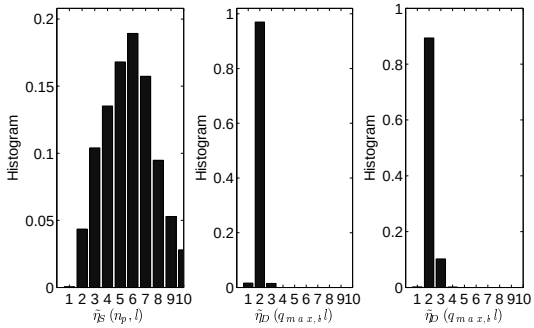

Figure 1: Histogram of $\tilde{\eta}_S(n_p,l)$ for voiced speech (left), and Histograms of $\tilde{\eta}_D(q_{max,l},l)$ for white noise (middle), and for babble noise (right). The pitch period n_p is estimated from the clean speeches by the FFT-spectrum-based scheme.

In a similar way of spectral subtraction in the frequency domain, the EVSC can be obtained by a "cepstral subtraction" method

$$\left[E\left\{C_{\hat{S}}(q,l)\right\}\right]^2 = G_S^2(q,l)\,E\left\{C_{\hat{S}}^2(q,l)\right\} \tag{33}$$

where $G_S^2(q,l)$ is a Wiener filter

$$G_S^2(q,l) = \frac{E\left\{\hat{\gamma}_{ceps}(q,l)\right\}-1}{E\left\{\hat{\gamma}_{ceps}(q,l)\right\}} = \frac{\xi_{ceps}(q,l)}{1+\xi_{ceps}(q,l)} \tag{34}$$

where $\hat{\gamma}_{ceps}(q,l)$ and $\xi_{ceps}(q,l)$ are the *a posteriori* SNR and the *a priori* SNR of speech cepstrum, respectively

$$\hat{\gamma}_{ceps}(q,l) = \frac{C_{\hat{S}}^2(q,l)}{\chi_q^2}, \quad \xi_{ceps}(q,l) = \frac{\left[E\left\{C_{\hat{S}}(q,l)\right\}\right]^2}{\chi_q^2}. \tag{35}$$

In practice, $E\left\{\hat{\gamma}_{ceps}(q,l)\right\}$ is unknown, an estimate of $\hat{\gamma}_{ceps}(q,l)$ is given by

$$\tilde{\gamma}_{ceps}(q,l) = \frac{\left[\sum_{i=-w}^{w} b(i)\,C_{\hat{S}}(q-i,l)\right]^2}{\chi_q^2} \tag{36}$$

The EVSC is proposed to estimate in the following way

$$E\left\{C_{\hat{S}}(q,l)\right\} = \tilde{G}_S(q,l)\,C_{\hat{S}}(q,l) \tag{37}$$

where the gain function $\tilde{G}_S(q,l)$ is obtained by the following two steps. The oversubtraction is used at the first step

$$\tilde{\xi}_{ceps}(q,l) = \max\left\{\tilde{\gamma}_{ceps}(q,l) - v, 0\right\} \tag{38}$$

$$\tilde{G}_{ceps}(q,l) = \sqrt{\max\left\{\frac{\tilde{\xi}_{ceps}(q,l)}{\tilde{\xi}_{ceps}(q,l)+1}, 0.01\right\}} \tag{39}$$

where $\upsilon > 1$ is an oversubtraction factor. Small values of cepstra are reduced by (38) and (39), to preserve more speech components at the end of words, we propose to adaptively smooth $\tilde{G}_{ceps}(q,l)$ at the second step as follows

$$\tilde{G}_S(q,l) = \left\{\begin{array}{ll} \tilde{G}_{ceps}(q,l) & if \quad \tilde{G}_S(q,l-1) \le \tilde{G}_{ceps}(q,l) \\ \alpha_1\tilde{G}_S(q,l-1) + (1-\alpha_1)\,\tilde{G}_{ceps}(q,l) & else \end{array}\right. \tag{40}$$

where α_1 is a constant smoothing factor.

Compared with the conventional cepstral subtraction method, the proposed approach only uses the estimated speech spectra to obtain the EVSC, and is totally based on the probability theory. There are two benefits by doing so. First, our approach does not need to calculate noisy speech cepstrum and noise cepstrum. Second, our approach is based on the estimation, so it does not conflict with the nonlinear model presented in (1).

A modified cepstrum thresholding method

Given the following two hypotheses, let $H_0(q,l)$ indicate a hypothesis $E\left\{C_{\hat{S}}(q,l)\right\}=0$, and let the alternative hypothesis $H_1(q,l)$ indicate $E\left\{C_{\hat{S}}(q,l)\right\}\neq 0$. Based on the property of speech cepstra, a large proportion of them are close to zero, that is to say, the probability of $H_0(q,l)$ is much larger than that of $H_1(q,l)$. The false alarm rate (FAR) P_F, which defines the probability of inferring that $H_1(q,l)$ is true when in fact $H_0(q,l)$ is true, can be derived as

$$P_F = P\left(\left|C_{\hat{S}}(q,l)\right| > \hbar\chi_q \middle| E\left\{C_{\hat{S}}(q,l)\right\}=0\right) \quad (41)$$

where $\hbar$ is the threshold. We propose to use $\tilde{C}_{\hat{S}}(q,l)$ instead of $C_{\hat{S}}(q,l)$ in (41), where $\tilde{C}_{\hat{S}}(q,l)$ is obtained by substituting $C_{\hat{S}}(q,l)$ into (26), then the FAR P_F is

$$P_F = P\left(\left|\tilde{C}_{\hat{S}}(q,l)\right| > \tilde{\hbar}\tilde{\chi}_q \middle| E\left\{\tilde{C}_{\hat{S}}(q,l)\right\}=0\right) \quad (42)$$

If the threshold $\hbar=\tilde{\hbar}\sqrt{\bar{r}}$ is used, (42) can be written as

$$P_F = P\left(\left|\tilde{C}_{\hat{S}}(q,l)\right| > \hbar\chi_q \middle| E\left\{\tilde{C}_{\hat{S}}(q,l)\right\}=0\right) \quad (43)$$

When the FAR P_F is given, the threshold $\hbar$ can be derived theoretically. The EVSC can be given by

$$E\left\{C_{\hat{S}}(k,l)\right\} = G_{CT}(q,l)\,C_{\hat{S}}(q,l) \quad (44)$$

where

$$G_{CT}(q,l)=\begin{cases} 1 & \texttt{if} \quad \left|\tilde{C}_{\hat{S}}(q,l)\right| > \hbar\chi_q \\ 0.1 & \texttt{else} \end{cases} \quad (45)$$

Using (44) and (45) directly may cause audible speech distortion at the end of words, so (40) is applied to smooth $G_{CT}(q,l)$:

$$\tilde{G}_{CT}(q,l)=\begin{cases} G_{CT}(q,l) & \texttt{if} \quad \tilde{G}_{CT}(q,l-1)\le G_{CT}(q,l) \\ \alpha_2\tilde{G}_{CT}(q,l-1)+(1-\alpha_2)\,G_{CT}(q,l) & \texttt{else} \end{cases} \quad (46)$$

where α_2 is a smoothing factor. Substituting $\tilde{G}_{CT}(q,l)$ into (44), $E\left\{C_{\hat{S}}(k,l)\right\}$ is given by

$$E\left\{C_{\hat{S}}(q,l)\right\} = \tilde{G}_{CT}(q,l)\,C_{\hat{S}}(q,l). \quad (47)$$

Obviously, the proposed method is based on the detection, which is similar to the cepstrum thresholding method [13,14]. The only difference between the modified cepstrum thresholding method and the original method is that, in this paper, $H_0(q,l)$ indicates the hypothesis $E\left\{C_{\hat{S}}(q,l)\right\}=0$ instead of $\left|E\left\{C_{\hat{S}}(q,l)\right\}\right|\le\chi_q$. This is based on the fact that the expected values of most of speech cepstra at high quefencies are zero. The assumption is not true for speech cepstra at low quefrencies, but using $E\left\{C_{\hat{S}}(q,l)\right\}=0$ can detect more low values of speech cepstra at the same FAR compared with using $\left|E\left\{C_{\hat{S}}(q,l)\right\}\right|\le\chi_q$.

Application to speech enhancement

After obtaining the EVSC by the two algorithms, the *a priori* SNR, $\xi_{ceps}(k,l)$, can be computed by (13) and (4). The enhanced speech signal is calculated by the Wiener filter

$$\tilde{s}(n)=\texttt{IFFT}\left\{\max\left\{\frac{\xi_{ceps}(k,l)}{1+\xi_{ceps}(k,l)},G_{\min}\right\}\bullet Y(k,l)\right\} \quad (48)$$

where $G_{\min}$ is a small lower bound on the gain function.

In the following sections, estimating the EVSC by the novel cepstral subtraction will be referred as the proposed algorithm 1 (ALG1), and estimating them by the modified cepstrum thresholding method is denoted by the proposed algorithm 2 (ALG2).

PERFORMANCE EVALUATION

This section presents the performance evaluation of the ALG1 and the ALG2, as well as a comparison with two state-of-the-art speech enhancement(SE) algorithms, including the DD-SE, and the CS-SE. The noise signals (white Gaussian noise (WGN), factory noise, and babble) used in our evaluation are taken from the Noisex92 database [23], where the NPSD is estimated by the MS method [10]. The clean speech signals, which are sampled at 16kHz, are taken from the TIMIT database [22]. More than 400 clean speech samples, which are summed up to about 20 minutes without intervening pauses, are degraded by the various noise types with segmental SNRs in the range [-5 15]dB. The parameters used in the ALG1 and the ALG2 are chosen as follows: $\beta=2$, $w=2$, $\alpha_1=\alpha_2=0.85$, $\upsilon=8$, $G_{\min}=-15$dB, and the thresholding $\hbar=3.30$ corresponding with the FAR P_F close to zero. To be mentioned, when the FAR $P_F=0.1\%$ is selected, the threshold $\tilde{\hbar}=3.30$ could be derived theoretically. In the paper, $\hbar=3.30$ instead of $\tilde{\hbar}=3.30$ is used, then the FAR P_F will be much less than 0.1%. Moreover, better performances could be achieved if $\hbar=3.30$ is used in practice. The frame length is $N=512$ with the frame shift $N/2=256$.

Comparison of the *a priori* SNR estimators

Since the *a priori* SNR is the only different parameter for the four SE algorithms, we believe it is due to the better estimate of the *a priori* if one algorithm is better than the others [1-4,27]. In this part, the word "Jane" corrupted by the WGN and/or the babble at 10dB is used to evaluate the performance of the *a priori* SNR for the four SE algorithms.

The *a priori* SNR estimation at frequency 562.5Hz for the WGN case, as illustrated in Figure 2, reveals that the better performances of the ALG1 and the ALG2 are particularly obvious at word onsets and offsets. As can be seen, the *a priori* SNR estimation is smoothed at the noise-only segments by the four estimators, which could eliminate the *musical noise* problem. Clearly, the DD-SE SNR, as well as the CS-SE SNR could not response fast enough to an abrupt increase in the *a posteriori* SNR. The main reason is that the constant smoothing factor α_{dd} for the DD-SE SNR estimator and the constant smoothing factor $\alpha(q,l)$ at the low quefrencies for the CS-SE SNR estimator cause noticeable delay. The ALG1 and the ALG2 estimators are capable of tracking the *a posteriori* SNR at word onsets and offsets rapidly. Figure 3 shows the results of the *a priori* SNR estimation at frequency 562.5Hz for the babble noise case. Obviously, the DD-SE SNR could not discriminate between speech onsets and non-stationary noise components from the fifth frame to the tenth frame. Whereas, the CS-SE SNR and our proposed SNR estimators could suppress the non-staionary nose components effectively. Therefore, our proposed algorithms are expected to suppress more *musical noise* and non-stationary noise components, while preserving more speech components.

Objective measures

In this part, we measure the improvement of the segmental SNR and the log-spectral distance. Table 2 and Table 3 present a comparison of the averaged segmental SNR improvement and the averaged log-spectral distance for the three different types of noise signals. The CS-SE and the two proposed algorithms consistently yield a higher improvement of the segmental SNR and a lower log-spectral distortion than the DD-SE. The main reason is that most of the residual noise is neither periodic, nor predominantly slowly varying, and absolute values of cepstra of the residual noise are so small that the cepstral smoothing technique and the two proposed algorithms can suppress them effectively. Cepstra of the very non-stationary noise are also small due to the fact that most of noise is not harmonic. Thus,

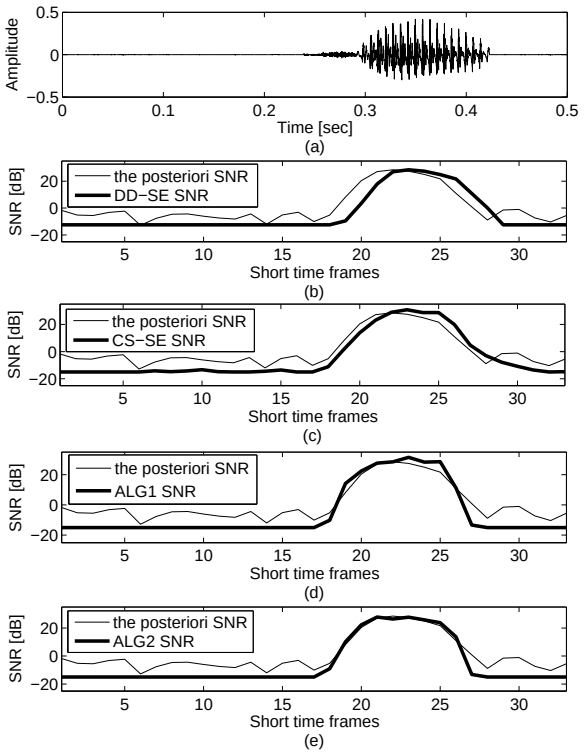

Figure 2: Estimation of the *a priori* SNR at frequency 562.5Hz for the word "Jane" corrupted by the WGN noise at 10dB. (a) Original speech waveform; (b) the DD-SE SNR estimator; (c) the CS-SE SNR estimator; (d) the ALG1 SNR estimator; (e) the ALG2 SNR estimator.

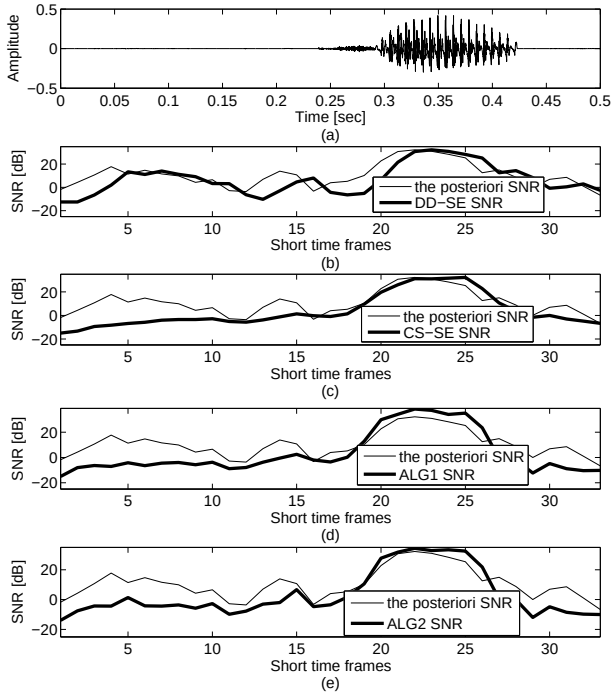

Figure 3: Estimation of the *a priori* SNR at frequency 562.5Hz for the word "Jane" corrupted by the babble noise at 10dB. (a) Original speech waveform; (b) the DD-SE SNR estimator; (c) the CS-SE SNR estimator; (d) the ALG1 SNR estimator; (e) the ALG2 SNR estimator.

even the very non-stationary noise could be effectively reduced when its cepstra are suppressed. At low input SNR, like -5dB the CS-SE is comparable with our approaches. Whereas, the ALG1 and the ALG2 are much better than the CS-SE at high input SNR for all types of noise. This is because the CS-SE uses the constant smoothing factors at low quefrencies, which may cause audible speech distortion at speech onsets. Our approaches are based on the characteristics and the second-order statistics of speech cepstra, which can adaptively suppress/preserve the cepstra according to their values. So, our approaches produce less speech distortion at speech onsets even when the frame shift is $N = 512$.

In this part, we further analyze the amount of spectral outliers, since listening test showed that reducing it yields a higher signal quality [6,9]. In Figure 4, log-histrograms of the normalized filtered spectrum $\left|\hat{D}(k,l)\right| \Big/ \sqrt{\lambda_d(k,l)}$ for the WGN and the babble noise cases are given. Note that in this experiment, $\left|\hat{D}(k,l)\right| = |G(k,l)D(k,l)|$ represents the residual noise component, where $G(k,l)$ is the Wiener filter obtained by the four *a priori* SNR estimators. Figure 4 does not give the result of the ALG2 due to its similar performance to that of the ALG1. From Figure 4, it is clear that, compared with the DD-SE algorithm, the CS-SE and the ALG1 algorithms could dramatically reduce the number of outliers. Moreover, the ALG1 produces less outliers than the CS-SE, which further indicates that the proposed algorithms do not introduce more *musical noise* components than the CS-SE.

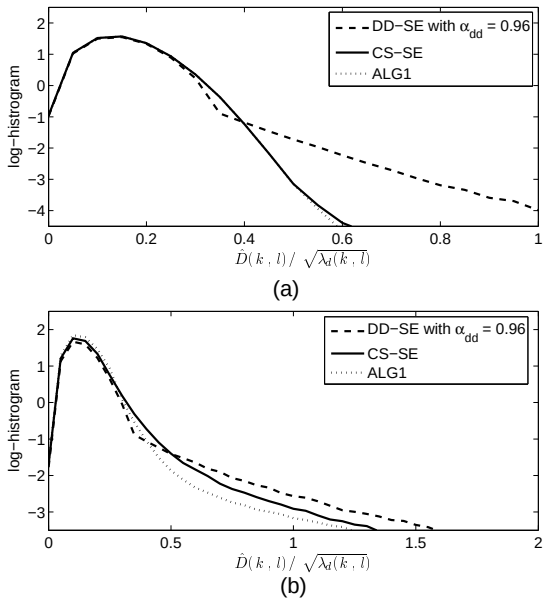

Figure 4: Log-histrograms of the normalized filtered spectrum $|\hat{D}(k,l)|/\sqrt{(\lambda_d(k,l))}$ for (a) the WGN; (b) the babble noise. The outliers of relatively large amplitude are perceived as *musical noise*.

Speech spectrogram

The results of objective measures are further confirmed by a subjective study of speech spectrograms. Figure 5 shows a comparison of speech spectrograms, where the clean speech is corrupted by the babble noise and enhanced speech signals are obtained by the DD-SE, the CS-SE, and the two proposed approaches. Obviously, more speech components are preserved by the CS-SE and the two proposed algorithms, and the preserved speech components by our proposed approaches are much stronger than those by the CS-SE. Moreover, the CS-SE and the two proposed approaches suppress more non-stationary

Table 2: Segmental SNR Improvement for Three Types of Noise, Obtained by The Decision-Directed Approach (DD), The Cepstral Smoothing Technique (CS), and The Proposed Two Approaches (ALG1 and ALG2).

Input SegSNR	WGN noise				Factory noise				babble			
[dB]	DD	CS	ALG1	ALG2	DD	CS	ALG1	ALG2	DD	CS	ALG1	ALG2
-5	8.30	8.44	8.66	8.69	5.90	6.02	6.04	6.16	5.06	5.89	6.04	5.97
0	6.35	6.67	7.25	7.28	3.75	4.34	4.64	4.77	3.29	4.40	4.73	4.70
5	4.28	4.76	5.69	5.72	2.07	2.90	3.49	3.58	1.69	3.09	3.46	3.45
10	2.36	3.00	4.19	4.20	0.96	1.69	2.39	2.45	0.88	1.89	2.29	2.31
15	0.68	1.54	2.86	2.86	-0.10	0.53	1.21	1.23	-0.09	0.65	1.20	1.18

Table 3: Log-Spectral Distance for Three Types of Noise, Obtained by The Decision-Directed Approach (DD), The Cepstral Smoothing Technique (CS), and The Proposed Two Approaches (ALG1 and ALG2).

Input SegSNR	WGN noise				Factory noise				babble			
[dB]	DD	CS	ALG1	ALG2	DD	CS	ALG1	ALG2	DD	CS	ALG1	ALG2
-5	3.61	3.53	3.50	3.51	3.45	3.34	3.34	3.32	2.99	2.73	2.70	2.69
0	2.70	2.53	2.47	2.47	2.81	2.57	2.51	2.50	2.37	2.05	1.99	1.99
5	2.22	1.98	1.88	1.88	2.20	1.90	1.83	1.82	1.80	1.49	1.45	1.45
10	1.85	1.58	1.46	1.46	1.62	1.37	1.29	1.29	1.29	1.08	1.05	1.05
15	1.41	1.18	1.03	1.03	1.16	0.98	0.92	0.92	0.93	0.80	0.78	0.77

noise components than the DD-SE. It is mostly due to that it is easier to distinct the speech components from the noise components in the cepstral domain. Therefore, the non-stationary noise can be reduced when its cepstra are suppressed by the cepstral smoothing techniques and the two proposed approaches. Informal listening tests also indicate that the two proposed approaches preserve more speech components at speech onsets and do not cause any reverberation effect at the end of words, where the results are consistent with the comparison of the *a priori* SNR estimators.

CONCLUSIONS

This paper presents a novel *a priori* SNR estimation method, which is based on expected values of speech cepstra. By a brief study of speech/noise cepstra, we reveal two underlying characteristics of them. One is that a significant proportion of speech cepstra are zero based on the speech generation model, the characteristic is also suitable for the broadband noise. The other is that the first rahmonic peak values of voiced speech cepstra are mostly larger than the largest absolute values of noise cepstra at high quefrencies. Based on the two characteristics and the second-order statistics of speech cepstra under the Gaussian assumption, two algorithms are proposed to estimate the EVSC. Both the novel cepstral smoothing method and the modified cepstrum thresholding method are capable of preserving large values of cepstra while suppressing small values of cepstra. Most of noise cepstra that often have small values are reduced, and the important large values of speech cepstra are preserved without distortion. Therefore, obtaining the *a priori* SNR by the EVSC in the cepstral domain is better than obtaining it by the *a posteriori* SNR in the frequency domain. Simulation results confirm that our approaches are much better than the DD-SE and comparable or somewhat better than the CS-SE in terms of output segmental SNR and log-spectral distance. A subjective study of speech periodograms further verify that the CS-SE and our approaches could suppress more non-stationary noise components and preserve more speech components than the DD-SE, and our approaches have the least speech distortion at speech onsets among them.

If the NPSD could be estimated more accurately in highly non-stationary noise environments, we believe that the performances of our approaches could be improved. This conclusion could be deduced by the performances of our approaches in the white noise case. The NPSD estimation errors have different influences on the performances of the speech enhancement algorithms. Further work should concentrate on analyzing the effects in detail.

Recently, we propose an adaptive-bandwidth and low-variance spectral estimator based on the structure of the noise power spectral density for spectral subtraction (NPSD-SS) [28]. Compared with the raw periodogram-based conventional spectral subtraction algorithm [17], the NPSD-SS could suppress the *musical noise* effectively, but it still could not reduce the influence of the non-stationary noise. If the EVSC-SE is applied to the NPSD-SS, the performance may be further improved.

REFERENCES

[1] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean square error short-time spectral amplitude estimator, *IEEE Trans. Acoustics, Speech, and Signal Process.*, **32**, 1109-1121(1984)

[2] C. Plapous, C. Marro and P. Scalart, Improved signal-to-noise ratio estimation for speech enhancement, *IEEE Trans. Audio, Speech, and Lang. process.*, **14**, 2098-2108(2006)

[3] I. Cohen, Relaxed statistical model for speech enhancement and *a priori* SNR estimation, *IEEE Trans. Speech and Audio Process.*, **13**, 870-881(2005)

[4] C. Zheng, Y. Zhou, and X. Li, A modified *a priori* SNR estimator based on the united speech presence probabilities, *Journal of Electronics and Information Technology*, **30**, 1680-1683(2008). (*in Chinese*)

[5] C. Breithaupt, T. Gerkmann and R. Martin, Cepstral smoothing of spectral filter gains for speech enhancement without musical noise, *IEEE Signal Process. Letter*, **14**, 1036-1039(2007)

[6] C. Breithaupt, T. Gerkmann and R. Martin, A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing, in: *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Las Vegas, Nevada, USA, March 2008.

[7] D. Malah, R. V. Cox, and A. J. Accardi, Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments, in: *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Phoenix, AZ, March 1999.

[8] I. Cohen and B. Berdugo, Speech enhancement for non-stationary noise environments, *Signal Processing*, **81**, 2043-2418(2001)

[9] T. Gerkmann, C. Breithaupt, and R. Martin, Improved *A Posteriori* Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors, *IEEE Trans. on Audio,*

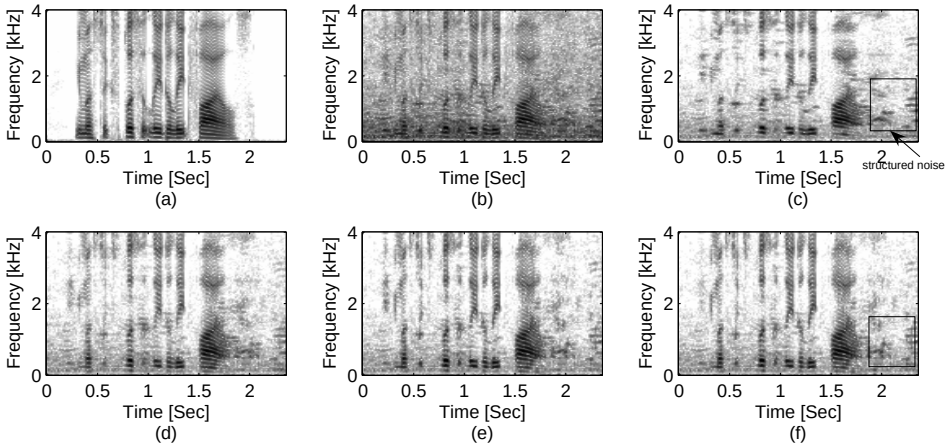

Figure 5: Comparison of spectrograms for the clean speech corrupted by the babble noise at a SNR=0dB. (a) Clean speech; (b) noisy speech (additive babble noise at a SNR=0dB); (c) enhanced speech signals using the DD-SE; (d) enhanced speech signals using the CS-SE; (e) enhanced speech signals using the ALG1; (f) enhanced speech signals using the ALG2.

Speech, and Lang. Process., **16**, 910-919(2008)

[10] R. Martin, Bias compensation methods for minimum statistics noise power spectral density estimation, *Signal Processing*, **86**, 1215-1229(2006)

[11] S. Rangachari and P. C. Loizou, A noise-estimation algorithm for highly non-stationary enviornments, *Speech Communication*, **48**, 220-231(2006)

[12] J. S. Erkelens and R. Heusdens, Tracking of Nonstationary Noise Based on Data-Driven Recursive Noise Power Estimation, *IEEE Trans. on Audio, Speech, and Lang. Process.*, **16**, 1112-1123(2008)

[13] P. Stoica and N. Sandgren, Smoothed nonparametric spectral estimation via cepstrum thresholding, *IEEE Signal Process. Mag.*, **23**, 34-45(2006)

[14] P. Stoica and N. Sandgren, Total-variance reduction via thresholding: application to cepstral analysis, *IEEE Trans. Signal Process.*, **55**, 66-72(2007)

[15] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, Jacobian approach to fast acoustic model adaptation, in: *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Munich, Germany, April 1997.

[16] C. S. Wu, V. V. Nguyen, H. Sabrin, W. Kushner, and J. Damoulakis, Fast self-adapting broadband noise removal in the cepstral domain, in: *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Toronto, Ont., Canada, April 1991.

[17] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on Acoustics, Speech, and Signal Process.*, **27**, 113-120(1979)

[18] Y. Ephraim and M. Rahim, On second-order statistics and linear estimation of cepstral coefficients, *IEEE Trans. on Speech and Audio Process.*, 7, 162-176(1999)

[19] A. V. Oppenheim and R. W. Schafer, Hamonorphic analysis of speech, *IEEE Trans. on Audio and Electroacoustics*, **AU-16**, 221-226(1968)

[20] W. Verhelst and O. Steenhaut, A new model for the short-time complex cepstrum of voiced speech, *IEEE Trans. on Acoust. Speech, and Signal Process.*, **ASSP-24**, 43-51(1986)

[21] G. d. Krom, A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals, *Journal of Speech and Hearing Research*, **36**, 254-266(1993)

[22] J. S. Garofolo, DARPA TIMIT acoustic-phonetic speech database, Nat. Inst. Standards Technol. (NIST), 1988

[23] A. Varga and H. J. M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, *Speech Commun.*, **12**, 247-251(1993)

[24] D.-J. Liu and C.-T. Lin, Fundamental frequency estimation based on the joint time-frquency analysis of harmonics spectral structure, *IEEE Trans. on Speech and Audio Process.*, **9**, 609-621(2001)

[25] A. M. Noll, Cepstrum pitch determination, *J. Acoust. Soc. Am.*, **41**, 293-309(1967)

[26] A. V. Oppenheim and R. W. Schafer, from frequency to quefrency: A history of the cepstrum, *IEEE Signal Process. Mag.*, **21**, 95-106(2004)

[27] Y. Hu and P. C. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum, *IEEE Trans. on Speech and Audio Process.*, **12**, 59-67(2004)

[28] C. Zheng, H. Hu, Y. Zhou, and X. Li, Spectral subtraction based on the structure of noise power spectral density, *ACTA ACUSTICA*, **35**, 215-222(2010) (*in Chinese*)