Front. Electr. Electron. Eng. China 2011, 6(4): 542–546
DOI 10.1007/s11460-011-0181-8

RESEARCH ARTICLE

Yu FANG, Gang LIU, Jun GUO

Speech enhancement based on modified a priori SNR estimation

Abstract To solve the frame delay problem and match the previous frame, Plapous et al. [IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(6): 2098–2108] introduced a novel approach called two-step noise reduction (TSNR) technique to improve the performance of the speech enhancement system. However, TSNR approach results in spectral peaks of short duration and the broken spectral outlier, which degrade the spectral characteristics of the speech. To solve this problem, a cepstral smoothing step is added in order to remove these spectral peaks brought by TSNR approach. Theory analysis shows that the proposed approach can effectively smooth the spectral peaks and keep the spectral outlier so as to protect the speech characteristics. Experiment results also show that the proposed approach can bring significant improvement compared to decision-directed (DD) and TSNR approaches, especially in non-stationary noisy environments.

Keywords speech enhancement, decision-directed (DD), two-step noise reduction (TSNR), signal-to-noise ratio (SNR) estimation

1 Introduction

With the development of mobile communication, new challenges are introduced. To improve the communication quality in mobile environment and achieve communication "anywhere and anytime", noisy speech signal should be enhanced before transmitting to far end. Therefore, the suppression of background acoustic noise is becoming more and more important.

Many speech enhancement approaches have been investigated in the past few decades, such as minimum mean square error (MMSE) [1], spectral subtraction (SS) [2], etc. In the speech enhancement process, the estimation of the a priori signal-to-noise ratio (SNR) is one of the most important parts, especially in non-stationary environments. Decision-directed (DD) approach is a practical way of estimating the a priori SNR [1] and has been widely used. DD approach can reduce musical noise, which is a very common phenomenon in speech enhancement approaches. Some modifications based on DD approach are given in Refs. [3,4]. However, DD approach considers the previous frames rather than the current frame and introduces a simple delay of one short time, which degrades the performance of the noise reduction. To solve this problem, Ref. [5] proposes a new method, called two-step noise reduction (TSNR), to refine the estimation of the a priori SNR. However, musical noise phenomenon still exists, which can lead to speech distortion and have impact on personal perception.

Recently, applying temporal smoothing in cepstral domain was found to be a promising approach for speech enhancement, especially in non-stationary environments [6]. Noisy speech can be decomposed into several coefficients in cepstral domain, such as speech envelop, excitation, and noise [7]. Applying selective temporal smoothing to different kinds of coefficients can keep the speech characteristics compared to the approaches that apply spectral smoothing in short-time Fourier transform (STFT) domain. In this paper, a modified a priori SNR estimation using cepstral smoothing for noisy speech enhancement is proposed, which can suppress musical noise. Experimental results show that the proposed approach can improve performance over DD approach and TSNR approach.

The paper is organized as follows. In Sect. 2, DD approach, which is most widely used in state-of-the-art estimator, is reviewed. Section 3 describes TSNR approach and gives a modified a priori SNR estimation. Then, experiments and evaluation are given in Sect. 4. Finally, conclusions are given in Sect. 5.

Received May 6, 2011; accepted July 14, 2011

Yu FANG (✉), Gang LIU, Jun GUO
Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
E-mail: anniefangyu@gmail.com

2 Review of a priori SNR estimation

2.1 Speech enhancement in DFT domain

Consider an additive noise model where speech and noise are independent:

$$y(t) = s(t) + n(t), \tag{1}$$

where $y(t)$, $s(t)$, and $n(t)$ represent the observed noisy speech, clean speech, and noise, respectively, and are considered to be random processes. The noisy signal $y(t)$ is transformed into frequency domain by applying a Hanning window $w_{\text{hann}}(\tau)$, $\tau = 0,1,...,M$, to a frame of M consecutive samples of $y(t)$ and by computing the discrete Fourier transform (DFT) of size M on the windowed data. Before the next DFT computation, the window is shifted by $M/2$ samples. The sliding window DFT analysis that results in a set of frequency domain signals can be written as

$$\begin{aligned} Y(k,l) &= \text{DFT}\left\{ w_{\text{hann}}(\tau) y\left(l\frac{M}{2} + \tau \right) \right\} \\ &= S(k,l) + N(k,l), \end{aligned} \tag{2}$$

where $l \in \mathbb{Z}$ denotes the frame index, $k = 0,1,...,M$ denotes the frequency bin index, and $S(k,l)$ and $N(k,l)$ denote the speech and noise in the DFT domain. In this paper, the sampling frequency of the signal is $f_s = 8$ kHz, and the DFT length is $M = 256$.

It is useful to consider the amplitude estimator $\hat{S}(k,l)$ as being obtained from $Y(k,l)$ by a multiplicative nonlinear gain function, which is defined by

$$\hat{S}(k,l) = G(k,l)Y(k,l), \tag{3}$$

where $G(k,l)$ is the gain function. Generally, the a posteriori SNR γ and a priori SNR ξ are defined as follows:

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_n(k,l)}, \tag{4}$$

$$\xi(k,l) = \frac{\text{E}[|S(k,l)|^2]}{\lambda_n(k,l)}, \tag{5}$$

where $Y(k,l)$ denotes the observed signal in DFT domain. In this paper, the gain function can be described as the Wiener filter similar to [1]

$$G(k,l) = \frac{\hat{\xi}(k,l)}{1 + \hat{\xi}(k,l)}, \tag{6}$$

where $\hat{\xi}(k,l)$ denotes the estimation of a priori SNR.

2.2 DD approach

We will describe the DD approach as follows [1], which is found to be very useful when it is combined with either the MMSE and Wiener amplitude estimator. The derivation of the a priori SNR estimator is based on the definition of $\xi(k,l)$ (see Eq. (5)) and its relation to the a posteriori SNR $\gamma(k,l)$,

$$\xi(k,l) = \text{E}\{\gamma(k,l) - 1\}. \tag{7}$$

Using Eqs. (5) and (7), we can write

$$\xi(k,l) = \text{E}\left\{ \frac{1}{2}\frac{|S(k,l)|^2}{\lambda_n(k)} + \frac{1}{2}[\gamma(k,l) - 1] \right\}. \tag{8}$$

The proposed estimator $\hat{\xi}(k,l)$ of $\xi(k,l)$ is deduced from Eq. (8) and is given by

$$\hat{\xi}(k,l) = \alpha\frac{\left|\hat{S}(k,l-1)\right|^2}{\lambda_n(k,l-1)} + (1-\alpha)P[\gamma(k,l) - 1] \quad (0 \leqslant \alpha < 1), \tag{9}$$

where $\hat{S}(k,l-1)$ is the amplitude estimator of the kth signal spectral component in the $(l-1)$th analysis frame; α is usually set to be 0.98, which can result in a great reduction of the noise, and provide enhanced speech with colorless residual noise. $P[\ ]$ is an operator, which is defined by

$$P[x] = \begin{cases} x, & \text{if } x \geqslant 0, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

$P[\ ]$ is used to ensure the positiveness of the proposed estimator in case $\gamma(k,l) - 1$ is negative. It is also possible to apply the operator P on the right-hand side of Eq. (9) rather than on $\gamma(k,l) - 1$ only. Equation (9) can also be written in the following way:

$$\hat{\xi}(k,l) = \max\left\{ \alpha\frac{\left|\hat{S}(k,l-1)\right|^2}{\lambda_n(k,l-1)} + (1-\alpha)[\gamma(k,l) - 1],\ \xi_{\min} \right\}, \tag{11}$$

where $\xi_{\min}$ is a minimum value that is used to control the distortion of speech transients.

3 Modified a priori SNR estimation

3.1 TSNR approach

Some analysis of DD approach has been given in Refs. [5,8], and the results indicated that DD approach can limit the level of musical noise, but the estimated a priori SNR is biased since it depends on the speech spectrum estimation in the previous frame. Therefore, the gain function matches the previous frame rather than the current one, which degrades the noise reduction performance. Reference [5] proposed a method called TSNR technique, which can solve this problem while maintaining the benefits of DD approach. TSNR can be divided into the following steps.

In the first step, the gain function $G_{\mathrm{DD}}(k,l)$ is computed using DD approach, which is shown in Sect. 2, that is,

$$G_{\mathrm{DD}}(k,l)=\frac{\hat{\xi}_{\mathrm{DD}}(k,l)}{\hat{\xi}_{\mathrm{DD}}(k,l)+1}. \tag{12}$$

In the second step, the estimation of the a priori SNR is refined to remove the bias of DD approach, thus removing the reverberation effect. The estimation of the a priori SNR using TSNR is obtained as follows:

$$\hat{\xi}_{\mathrm{TSNR}}(k,l)=\frac{|G_{\mathrm{DD}}(k,l)Y(k,l)|^2}{\lambda_n(k,l)}. \tag{13}$$

Without loss of generality, in the following, the chosen spectral gain is also the Wiener filter:

$$G_{\mathrm{TSNR}}(k,l)=\frac{\hat{\xi}_{\mathrm{TSNR}}(k,l)}{\hat{\xi}_{\mathrm{TSNR}}(k,l)+1}. \tag{14}$$

3.2 Modified approach

DD approach and TSNR approach are all short-time spectrum estimation approaches. In short-time spectrum estimation approaches, the estimation of the a priori SNR estimation usually results in spectrum peaks in short-time duration, which will destroy the spectral outliers and then degrade the performance of the speech enhance system and the speech characteristics. In cepstral domain, coefficients can be decomposed into different kinds with relation to speech envelop, excitation, and the noise. The speech envelope is always represented by the same small set of cepstral coefficients. The coefficients that represent the excitation can be found by searching for a cepstral peak in a defined range [7]. The remaining coefficients are dominated by noise. The speech characteristics and the noise can also be represented by the a priori SNR function in cepstral domain. Applying selective cepstral smoothing to the coefficient dominated by noise and speech respectively for the estimated a priori SNR function obtained by DD and TSNR approaches can degrade the artificial noise obviously. The block diagram of the modified noise reduction system is depicted in Fig. 1. It is described in detail as follows.

At the end of the first step of TSNR approach, a cepstral representation of $\hat{\xi}_{\mathrm{DD}}(k,l)$ is calculated for each frame l as

$$\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps}}(q,l)=\mathrm{IDFT}\left\{\log\hat{\xi}_{\mathrm{DD}}(k,l)\right\}, \tag{15}$$

where IDFT{ } is the inverse DFT of length M resulting in cepstral bins q. Note that the symmetry condition $\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps}}(M-q,l)=\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps}}(q,l)$ holds.

Define q_{pitch} as the cepstral index that most likely represents f_0. It is found via a maximum search for a given frame l [6],

$$q_{\mathrm{pitch}}=\arg\max\left\{\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps}}(q,l)|q_{\mathrm{low}}\leqslant q\leqslant q_{\mathrm{high}}\right\}, \tag{16}$$

where the search is limited to possible fundamental frequencies between $f_{0,\mathrm{low}}$ and $f_{0,\mathrm{high}}$, resulting in the range of $q_{\mathrm{low}}=\lfloor f_{\mathrm{s}}/f_{0,\mathrm{high}}\rfloor$ to $q_{\mathrm{high}}=\lfloor f_{\mathrm{s}}/f_{0,\mathrm{low}}\rfloor$, with f_{s} the sampling frequency and $\lfloor\ \rfloor$ flooring operator toward the nearest integer number.

The detection of speech sounds is based on the following criteria [6]. First, the voiced speech sound has a comparatively high energy, so the cepstral peak value should be above a certain threshold. Second, speech has more energy at low frequencies since it is spectrally tilted. Therefore, the range of cepstral bins that are most likely to represent the fundamental frequency is

$$\mathbb{Q}_{\mathrm{pitch}}=\{q_{\mathrm{pitch}}-\Delta q_{\mathrm{pitch}},\ q_{\mathrm{pitch}},\ q_{\mathrm{pitch}}+\Delta q_{\mathrm{pitch}}\},$$

where $\Delta q_{\mathrm{pitch}}$ is a small margin.

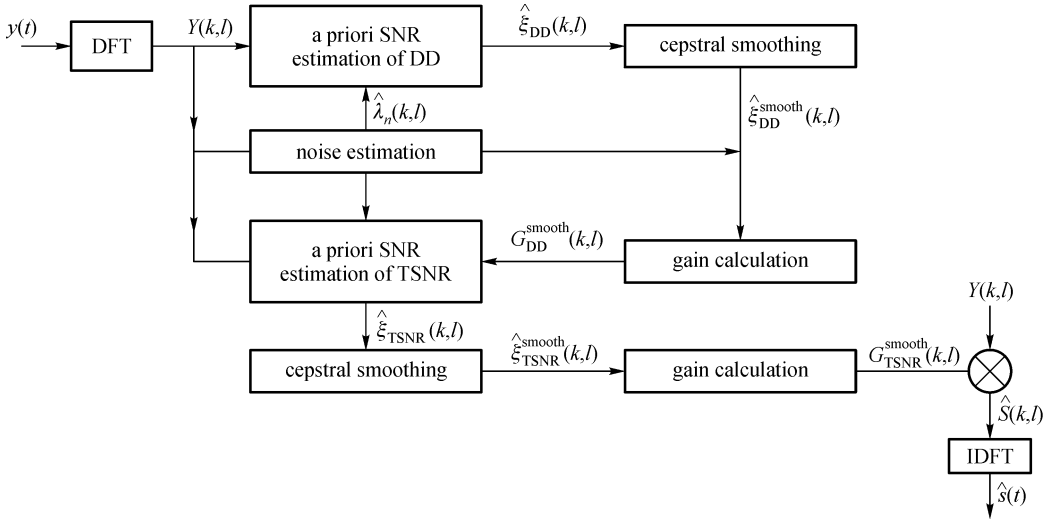

Fig. 1 Block diagram of modified noise reduction system

Therefore, all future modifications applied to the cepstral bins $q = 0,1,...,M/2$ are applied accordingly to the symmetric counterpart $q = M/2+1,...,M-1$. Then, smoothing is applied in the cepstral domain:

$$\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps\_smooth}}(q,l) = \beta\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps\_smooth}}(q,l-1) + (1-\beta)\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps}}(q,l), \tag{17}$$

where β is the smoothing factor. In addition, the cepstral bins that are likely to represent the fundamental frequency f_0 should not be smoothed, which means that the smoothing is applied to cepstral bins $q \in \{q_{\mathrm{low}},..., M/2\}\backslash\mathbb{Q}_{\mathrm{pitch}}$, and for $q \in \mathbb{Q}_{\mathrm{pitch}}$, no smoothing is applied at all.

A smoothed estimation $\hat{\xi}_{\mathrm{DD}}^{\mathrm{smooth}}(k,l)$ of the DD approach in the spectral domain is finally obtained by transforming back to the spectral domain, that is,

$$\hat{\xi}_{\mathrm{DD}}^{\mathrm{smooth}}(k,l) = \exp\left(\mathrm{DFT}\left\{\hat{\xi}_{\mathrm{DD}}^{\mathrm{ceps\_smooth}}(q,l)\right\}\right), \tag{18}$$

where $\hat{\xi}_{\mathrm{DD}}^{\mathrm{smooth}}(k,l)$ is additionally constrained to values below or equal to one.

Then, $\hat{\xi}_{\mathrm{DD}}^{\mathrm{smooth}}(k,l)$ is used to calculate the a priori SNR of TSNR approach by Eqs. (12) and (13), and the a priori SNR estimation of TSNR can be obtained, representing by $\hat{\xi}_{\mathrm{TSNR}}(k,l)$. Applying cepstral smoothing to $\hat{\xi}_{\mathrm{TSNR}}(k,l)$ again by Eqs. (15)–(18), $\hat{\xi}_{\mathrm{TSNR}}^{\mathrm{smooth}}(k,l)$ can be obtained. In the end, we can get the gain of TSNR through Eq. (14).

4 Experiments and evaluation

In this section, the performance of the proposed approach is tested for speech enhancement, compared to that of DD approach and TSNR approach. The test consisted of ten speech utterances, five male and five female, from the TIMIT database [9], resampled at 8 kHz. The noise types considered were white Gaussian noise, speech babble noise, and HF channel noise (from NOISEX-92) [10].

The parameters used in the proposed approach are listed as follows: $\xi_{\min} = -25$ dB, $\beta = 0.9$, $\Delta q_{\mathrm{pitch}} = 1$, $f_{0,\mathrm{low}} = 70$ Hz, and $f_{0,\mathrm{high}} = 300$ Hz.

Table 1 shows the segmental SNR improvements of various noise types, in dB, defined by

$$\mathrm{SegSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10\lg\frac{\sum_{n=0}^{N-1} s^2\left(n+\frac{lN}{2}\right)}{\sum_{n=0}^{N-1}\left[s\left(n+\frac{lN}{2}\right)-\hat{s}\left(n+\frac{lN}{2}\right)\right]^2}, \tag{19}$$

where M is the number of frames that contain active speech, N is the number of samples per frame (corresponding to 32 ms frames), and l is a discrete-time index. The SNR at each frame is limited to perceptually meaning range between 35 dB and -10 dB. This prevents the segmental SNR measure from being biased in either a positive or negative direction due to a few silence or unusually high SNR frames that do not contribute significantly to the overall speech quality. For each noise type and SNR value, the average segmental SNR is the result of the averaging of the segmental SNRs obtained from ten sentences [11].

As can be seen in Table 1, the modified approach can get higher output averaged segmental SNR and achieve better performance. Improvement for noise reduction is obtained both in stationary environment and non-stationary environment compared to DD approach and TSNR approach.

Figure 2 shows the magnitude of clean speech, noisy speech signal corrupted by HF channel noise at 5 dB, noisy speech enhanced by DD, TSNR, and the proposed approach. Figure 2 indicates that in the proposed approach, the speech onsets and low-energy speech components can be preserved compared to DD and TSNR approach. Informal subjective listening also reveals that signals using the proposed approach sound clearer, as analyzed above that the spectral outlier was kept.

Table 1 Output segmental SNRs improvements using the proposed approach, DD approach and TSNR approach in various noise and SNR condition

noise type	input SNR/dB	averaged segmental SNR improvements/dB		
		DD	TSNR	the proposed approach
white	0	5.31	5.74	6.42
	5	3.96	4.50	5.43
	10	2.69	3.49	4.20
babble	0	4.45	4.63	5.23
	5	3.32	3.98	4.83
	10	2.61	3.57	3.92
HF channel	0	4.18	4.47	4.97
	5	3.93	3.94	4.62
	10	2.85	3.63	4.16

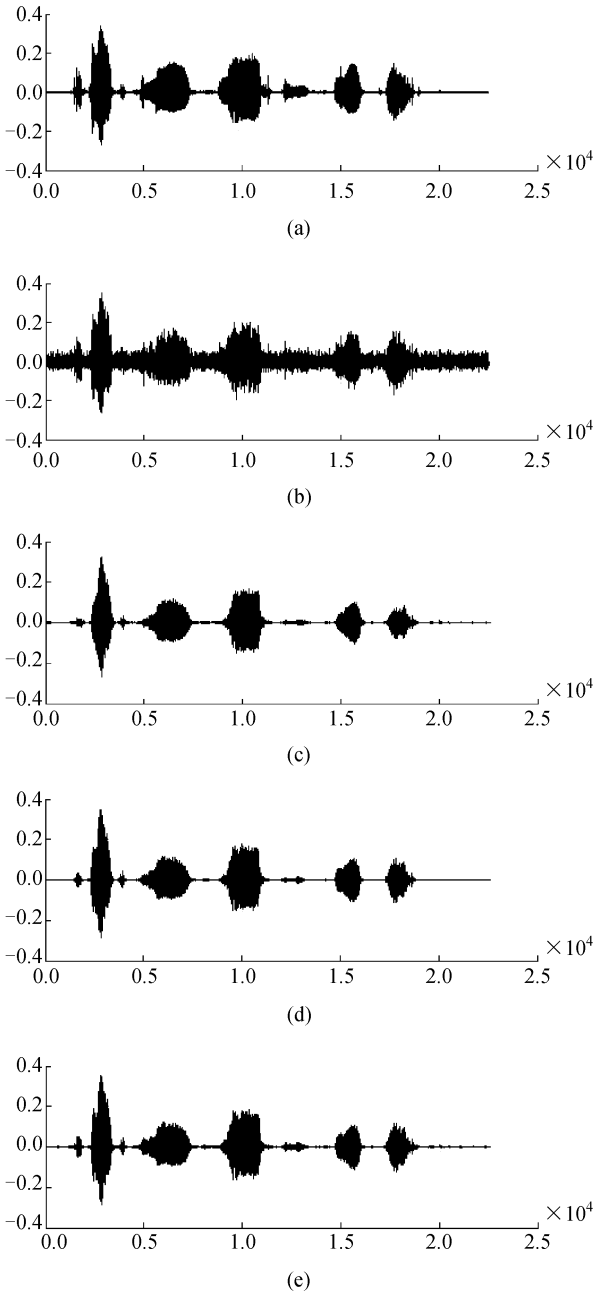

Fig. 2 Magnitude of speech. (a) Clean speech; (b) noisy speech signal corrupted by HF channel noise at 5 dB; (c) enhanced speech using DD approach; (d) enhanced speech using TSNR approach; (e) enhanced speech using the proposed approach

5 Conclusions

A modified the a priori SNR estimation approach for speech enhancement is proposed in this paper, which applies cepstral smoothing to TSNR approach, thus avoiding the disadvantages of TSNR by degrading the artificial musical noise. Simulation results show that the proposed algorithm has good performances for speech enhancement in different kinds of noisy environment, especially in non-stationary environments.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 61005004, 61175011, and 61171193), the Next-Generation Broadband Wireless Mobile Communications Network Technology Key Project (No. 2011ZX03002-005-01), the 111 project (No. B08004), and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

1. Ephraim Y, Malah D. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(6): 1109–1121
2. Boll S F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113–120
3. Cohen I. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 870–881
4. Cohen I. Speech enhancement using a noncausal a priori SNR estimator. IEEE Signal Processing Letters, 2004, 11(9): 725–728
5. Plapous C, Marro C, Scalart P. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(6): 2098–2108
6. Mauler D, Gerkmann T, Martin R. An analysis of quefrency selective temporal smoothing of the cepstrum in speech enhancement. In: Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control. 2008, 1–4
7. Noll A M. Cepstrum pitch estimation. Journal of the Acoustical Society of America, 1967, 41(2): 293–309
8. Cappe O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Transactions on Speech and Audio Processing, 1994, 2(2): 345–349
9. Garofolo J S, Lamel L F, Fisher W M, Fiscus J G, Pallett D S, Dahlgren N L, Zue V. DARPA TIMIT Acoustic-phonetic continuous speech corpus. NIST Speech Disc1–1.1, 1993
10. Varga A, Steeneken H J M, Tomlinson M, Jones D. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. The NOISEX-92 CD-ROMs, 1992
11. Deller J R Jr, Hansen J H L, Proakis J G. Discrete-Time Processing of Speech Signals. 2nd ed. New York: IEEE Press, 2000