International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.6, No.5 (2013), pp.411-422
http://dx.doi.org/10.14257/ijsip.2013.6.5.36

A Two-Step Noise Estimation Algorithm for Noisy Speech Enhancement

Shifeng Ou, Chao Geng, Xianyun Wang and Ying Gao*

Institute of Science and Technology for Opto-electronic Information, Yantai University, Yantai, 264005, China
ousfeng@126.com, wang2005ji@126.com, 834349366@qq.com

Abstract

Noise estimation is an important part for noisy speech enhancement due to its momentous effect on the intelligibility and quality of the enhanced speech. In this paper, an effective noise estimation algorithm is presented by combining the minimum statistics estimation and Gaussian model assumption. In contrast to other methods, the proposed approach works in two steps. The noise power estimated by minimum statistics method in the first step suffers some bias which is removed by the second step of the proposed approach using Bayes theorem. As the noise estimation is refined in the second step, more accurate estimation can be obtained. The performance of the proposed approach is evaluated by objective and subjective tests under various noise environments and found to yield better results compared with conventional MS -based estimate.

Keywords: *Noise estimation, Speech enhancement, Minimum statistics, Gaussian model*

1. Introduction

The purpose of noisy speech enhancement technique is to improve the naturalness and perceptual quality for the noisy speech signal in order to reduce the fatigue of human listeners. It also aims at achieving a better intelligibility of the noisy speech for listeners or to increase the accuracy of a speech recognition system operating in a noisy environment. Speech enhancement is useful in many applications such as voice communication and automatic speech recognition where efficient noise reduction techniques are required. A crucial component of a practical noisy speech enhancement system is the estimation of the noise power spectrum density (PSD), which can be used to compute the *a priori* signal-to-noise ratio (SNR) and subsequently the spectral gain function. Thus, the performance of a speech enhancement system highly depends on the accurate PSD estimation of the unknown noise [1]. In a non-stationary noise environment, this task is very difficult to solve since updating the noise PSD estimate during speech pause only is not sufficient to obtain a fast tracking of the noisy PSD. In fact the noise power estimation is still a very challenging problem and there are many algorithms in the literature which shows lots of interests in this topic [2-4].

One of the successful noise PSD estimation approach is the minimum statistics (MS), which obtains the noise PSD estimate as the minimum values of a smoothed power estimate of the noisy speech signal [5]. The MS method was originally motivated by the observation that the speech and the disturbing noise are usually statistically independent and the power of a noisy speech frequently reduces to that of the noise. However, this noise PSD estimation is

*Corresponding author: claragaoying@126.com

ISSN: 2005-4254 IJSIP

sensitive to its outliers and its variance is about twice as large as the variance of conventional noise estimation. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [6]. Soft decision (SD), the other well known noise power estimation technique adapts the noise statistics based on the uncertainty of speech absence [7]. This method does not rely on voice activity detector (VAD). However, it updates the noise statistics even in the presence of speech, and it is difficult to accurately measure the mixture ratio between speech and noise. The inaccurate measurement of speech absence could seriously distort the enhanced speech. Another state of the art approach to estimate the noise PSD is subspace decomposition based technique called subspace noise tracking algorithm [8]. This method is based on the eigenvalue decomposition of the noisy speech correlation matrices in Fourier transform (DFT) domain and works well in many environments. But, the accurateness of this algorithm depends on the estimation of the eigenvalues numbers in the speech signal subspace.

In this paper, we focus on the MS approach for noise PSD estimation and present an improved method called two-step noise estimation technique in discrete cosine transform (DCT) domain. Our algorithm uses the Gaussian distribution models of speech and noise signals to refine the estimation of the MS approach, and thus, a better estimate can be obtained. The remained of this paper is organized as follows: Section 2 gives the MS algorithm. A novel noise PSD estimate is proposed and evaluated in Section 3. In Section 4, a number quality tests are conducted to evaluate the performance of the algorithms with the minimum mean square error (MMSE) based speech enhancement system, and finally in Section 5, some concluding remarks are drawn.

2. Minimum statistics in DCT domain

It is assumed that the noise signal $n(t)$ is additive, *i.e.*

$$y(t) = x(t) + v(t) \tag{1}$$

with $x(t)$, $y(t)$ are the clean speech and noisy speech at time *t*. Taking the DCT to the observed noisy signal gives us

$$Y(k,m) = X(k,m) + N(k,m), \quad k = 0,\cdots,K-1 \tag{2}$$

where $X(k,m)$, $Y(k,m)$ and $N(k,m)$ denote the DCT transformed components of the clean speech, noisy speech and noise signals respectively, *K* is the total number of frequency components, *k* and *m* represent the frequency and frame index. It is assumed that different DCT components along *k* and *m* are statistically independent. In this approach, the minimum values of a smoothed power spectral density estimate of the noisy signal are tracked, and multiplied by a constant that compensates the estimate for possible bias. To search the minimum values of the local energy, the smoothed noisy speech periodogram $P(k,m)$ is considered as following

$$P(k,m) = \partial P(k,m-1) + (1-\partial)\,|Y(k,m)|^2 \tag{3}$$

where ∂ is the smoothing parameter, and its value is chosen to be very close to 1. As this smoothing parameter widens the peaks of speech activity of the smoothed PSD estimate $P(k,m)$, a fixed ∂ will lead to inaccurate noise estimates. To overcome this drawback, a time varying smoothing parameter $\partial(k,m)$ need to be searched so that the tracking capabilities of the smoothed periodogram $P(k,m)$ and this variance can be

better balanced. In [5], a time varying smoothing parameter $\partial(k,m)$ is given as following

$$\partial(k,m) = \frac{1}{1+\left(P(k,m-1)/\lambda_N(k,m)-1\right)^2} \tag{4}$$

where $\lambda_N(k,m) = E\left[\left|N(k,m)\right|^2\right]$ is the noise PSD.

By searching the minimum value within a finite window length of the noisy speech periodogram $P(k,m)$, the estimated noise power can be obtained as following

$$P_{\min}(m,k) = \min\{P(k,m-D+1),\cdots,P(k,m-1),P(k,m)\} \tag{5}$$

where D is the window length, and min[.] represents the minimum value operator. As the minimum noise power estimate is smaller than the average value, this method requires a bias compensation as given below

$$\lambda_N^{MS}(k,m) = B \cdot P_{\min}(k,m) \tag{6}$$

where B is the bias compensation factor and $\lambda_N^{MS}(k,m)$ is the unbiased noise power estimation.

3. Two-step Noise PSD Estimation

The MS based noise PSD estimate in Section 2 is based on the assumption that within the observed time span, the speech signals are absent during at least a small fraction of the total time span. The noise PSD is then obtained from the minimum of the estimated power periodogram of noisy speech. However, if the noise PSD rises within the observed time span, it will be underestimated or can only be tracked with a large delay. In order to enhance the performance of the MS method, we propose to estimate the noise PSD in a two-step procedure. This method will be referred to as the Two-Step Noise Estimation (TSNE) algorithm in the following. In the first step, we compute the noise PSD $\lambda_N^{MS}(k,m) = B \cdot Q_{\min}(k,m)$ as described in Section 2. Then the result in the first step will be used to refine the noise PSD estimation using the following equation

$$\lambda_N^{TSNE}(k,m) = E\left\{N^2(k,m)\middle|Y(k,m)\right\} \tag{7}$$

The numerator of (7) gives a more accurate estimation of the noise PSD. Equation (7) can be rewritten using Bayes theorem, as (for simplicity of notation, the index k and m are dropped)

$$\lambda_N^{TSNE} = E\left\{N^2\middle|Y\right\} = \frac{\int_{-\infty}^{\infty} N^2 p(Y|N)p(N)dN}{\int_{-\infty}^{\infty} p(Y|N)p(N)dN} \tag{8}$$

where $p(\cdot)$ denotes the probability density function (PDF), Under the Gausssian distribution assumptions for speech and noise, $p(Y|N)$ and $p(N)$ are given by the following equations [9]

$$p\{Y|N\} = \frac{1}{\sqrt{2\pi\lambda_X}} \exp\left\{-\frac{(Y-N)^2}{2\lambda_X}\right\} \tag{9}$$

$$p\{N\} = \frac{1}{\sqrt{2\pi\lambda_N}} \exp\left\{-\frac{N^2}{2\lambda_N}\right\} \tag{10}$$

where $\lambda_X = E(|X|^2)$, Substituting (9) and (10) into (8), the estimation of λ_N^{TSNE} can be derived by the following equation

$$\begin{aligned} E\{N^2|Y\} &= \frac{\int_{-\infty}^{\infty} N^2 \exp\{-\frac{(Y-N)^2}{2\lambda_X} - \frac{N^2}{2\lambda_N}\} dN}{\int_{-\infty}^{\infty} \exp\{-\frac{(Y-N)^2}{2\lambda_X} - \frac{N^2}{2\lambda_N}\} dN} \\ &= \frac{\int_{-\infty}^{\infty} N^2 \exp\{-\frac{N^2(\lambda_X+\lambda_N)}{2\lambda_X\lambda_N}\} \exp\{\frac{NY}{\lambda_X}\} dN}{\int_{-\infty}^{\infty} \exp\{-\frac{N^2(\lambda_X+\lambda_N)}{2\lambda_X\lambda_N}\} \exp\{\frac{NY}{\lambda_X}\} dN} \\ &= \frac{\frac{-\lambda_X\lambda_N}{\lambda_X+\lambda_N} \int_{-\infty}^{\infty} N \exp\{\frac{NY}{\lambda_X}\} d\exp\{-\frac{N^2(\lambda_X+\lambda_N)}{2\lambda_X\lambda_N}\}}{\int_{-\infty}^{\infty} \exp\{-\frac{N^2(\lambda_X+\lambda_N)}{2\lambda_X\lambda_N}\} \exp\{\frac{NY}{\lambda_X}\} dN} \\ &= \frac{\lambda_X\lambda_N}{\lambda_X+\lambda_N}\{1 + \frac{Y}{\lambda_X}[1 - E(X|Y)]\} \\ &= \frac{\lambda_X\lambda_N}{\lambda_X+\lambda_N} + \frac{{\lambda_N}^2}{(\lambda_X+\lambda_N)^2} Y^2 \end{aligned} \tag{11}$$

Then the proposed TSNE approach for the noise PSD estimation is obtained which can be described as following two steps.

The first step: Compute $\lambda_N^{MS}(k,m)$ using the MS method.

The second step: Refine the result obtained in the first step by

$$\lambda_N^{TSNE}(k,m) = \frac{\lambda_X(k,m)\lambda_N^{MS}(k,m)}{\lambda_X(k,m) + \lambda_N^{MS}(k,m)} + \frac{\left(\lambda_N^{MS}(k,m)\right)^2}{\left(\lambda_X(k,m) + \lambda_N^{MS}(k,m)\right)^2} Y^2(k,m) \tag{12}$$

where the speech PSD $\lambda_X(k,m) = E\left[|X(k,m)|^2\right]$ can be estimated using the well known decision-directed (DD) approach as following

$$\lambda_X(k,m) = \beta\lambda_X(k,m-1) + (1-\beta)\max\left\{Y^2(k,m) - \lambda_N^{MS}(k,m), 0\right\} \tag{13}$$

where max(.) is the maximum function used to ensure that a nonnegative value is obtained as an estimate, $\lambda_X(k,m-1)$ is the estimated speech PSD at previous frame, while β is a

controller parameter which can be adjusted to achieve the best result, and in our experimental the parameter is chosen as 0.98.

Figure 1 shows an example of the estimated noise power contours by the MS and our proposed TSNE methods in conjunction with the noisy speech signals. The figure is comprised of the noisy speech $Y(k,m)$, actual noise PSD, the estimated noise power $\lambda_N^{MS}(k,m)$ and $\lambda_N^{TSNE}(k,m)$ for a signal frequency bin k=14. The noise in the noisy speech is a nonstationary buccaneer noise with an input SNR about 0 dB. The window length is D=256. From this figure, we can see that the proposed TSNE approach follows the noise level much better than the MS approach. This is due to the fact that the proposed approach can refine the noise estimation of the MS approach in its second step.

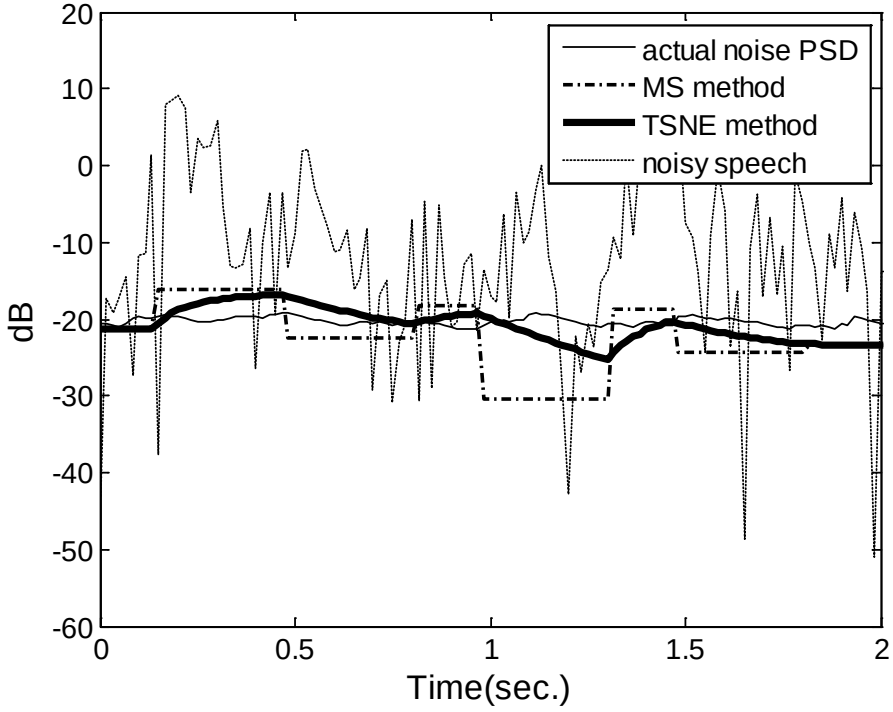


Figure 1. Comparison of Noise Estimation Obtained by MS and TSNE Methods

The normalized relative estimation error is also used to evaluated the noise estimation accuracy of each approach in various background noise environments, and the relative estimation error ε_N is defined by [10]

$$\varepsilon_N = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_k [\lambda_N(k,m) - \hat{\lambda}_N(k,m)]^2}{\sum_k [\lambda_N(k,m)]^2} \tag{14}$$

where $\lambda_N(k,m)$ is the actual noise power, $\hat{\lambda}_N(k,m)$ is the noise power estimated by the two tested methods and M is the number of the frames in the noisy speech signal. The noise signals used in our evaluation are taken from http://spib.ece.rice.edu/, which include the white noise (White), the Factory room noise (Factory), the Pink noise (Pink), the F16 noise (F16),

and the Buccaneer jet cockpit noise (Buccaneer). The speech signal is sampled at 8 kHz and degraded by these noises at the SNR of 0dB, 5dB, 10dB and 15dB. Table 1 gives the results of the relative estimation error for the evaluated noise estimation methods under the given noise conditions. From this table, we see that in general for different noise environments the proposed TSNE method achieves a consistent improvement in the relative estimation error over the MS method. Especially for the nonstationary noises (Factory, F16 and Buccaneer), we see that the proposed TSNE method distinctly outperforms the MS method.

Table 1. Relative Estimation Error Obtained from the MS and TSNE Methods

Noise type	Input SNR	MS method	TSNE method
White	0 dB	0.7418	0.6355
	5 dB	0.7421	0.6315
	10 dB	0.7686	0.6801
	15 dB	0.9662	0.8520
Factory	0 dB	0.6452	0.2952
	5 dB	0.6528	0.3098
	10 dB	0.6012	0.2208
	15 dB	0.6110	0.2361
Pink	0 dB	0.6276	0.2281
	5 dB	0.6825	0.3360
	10 dB	0.8006	0.6074
	15 dB	0.6062	0.4512
F16	0 dB	0.6315	0.3639
	5 dB	0.6784	0.3602
	10 dB	0.6130	0.3154
	15 dB	0.6004	0.2681
Buccaneer	0 dB	0.6145	0.3928
	5 dB	0.5923	0.3192
	10 dB	0.5873	0.2704
	15 dB	0.6082	0.3804

4. Performance Evaluation with Speech Enhancement System

As an application of the noise estimation technique, we consider a speech enhancement system based on MMSE as follows

$$\hat{X}(k,m) = E\left\{ X(k,m) \middle| Y(k,m) \right\} \tag{15}$$

where $\hat{X}(k,m)$ is estimated clean speech signal which can computed using Bayes theorem as

$$\hat{X}(k,m)=\frac{\int_{-\infty}^{\infty} X(k,m)\,p\big(Y(k,m)\,|\,X(k,m)\big)\,p\big(X(k,m)\big)\,dX(k,m)}{\int_{-\infty}^{\infty} p\big(Y(k,m)\,|\,X(k,m)\big)\,p\big(X(k,m)\big)\,dX(k,m)} \tag{16}$$

with the Gaussian distribution assumption, $p\big(Y(k,m)\,|\,X(k,m)\big)$ and $p\big(X(k,m)\big)$ are given as

$$p\big(Y(k,m)\,|\,X(k,m)\big)=\frac{1}{\sqrt{2\pi\lambda_N(k,m)}}\exp\left\{-\frac{\big(Y(k,m)-X(k,m)\big)^2}{2\lambda_N(k,m)}\right\} \tag{17}$$

$$p\big\{X(k,m)\big\}=\frac{\sqrt{1}}{\sqrt{2\pi\lambda_X(k,m)}}\exp\left\{-\frac{X^2(k,m)}{2\lambda_X(k,m)}\right\} \tag{18}$$

Combining equation (16), (17) and (18) gives us

$$\hat{X}(k,m)=\frac{\xi(k,m)}{1+\xi(k,m)}Y(k,m) \tag{19}$$

where

$$\xi(k,m)=\frac{\lambda_X(k,m)}{\lambda_N(k,m)} \tag{20}$$

which is known as the a priori SNR.

With the MMSE theorem in equation (19), the block scheme of the noised estimation based speech enhancement system is shown in Figure 2. This system works in DCT domain where the clean speech signals are estimated on a frame by frame basis.

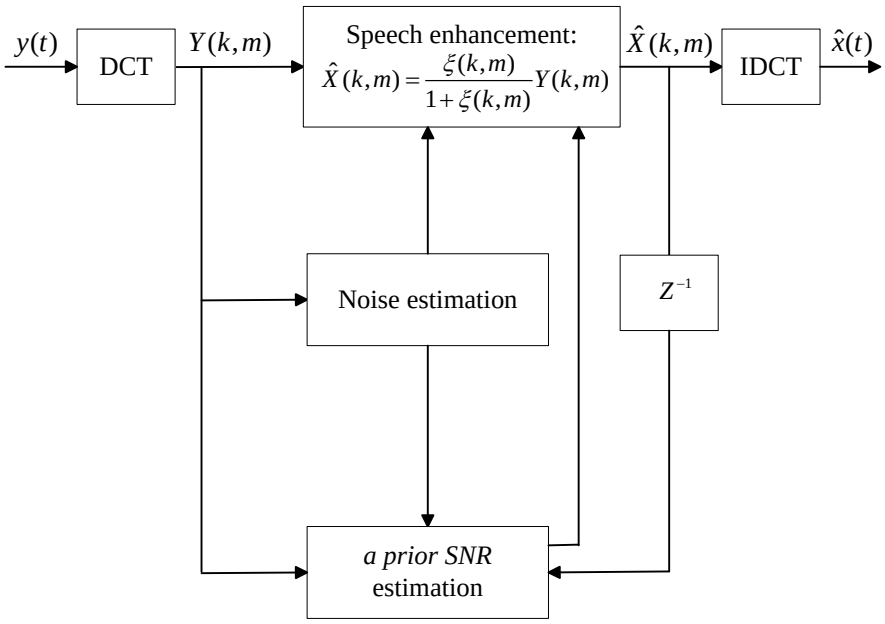


Figure 2. Block Scheme of Noise Estimation Speech Enhancement System

To evaluate the performance of the proposed method in speech enhancement system, four type noise signals are used which include White, F16, Destroyer engine room noise (Destroyerengine), and Babble noise. The speech signal is sampled at 8 kHz and degraded by these noises at the SNR of 0dB, 5dB, and 10dB. The number of samples per frame is K=256 with an overlap of 128 samples.

Firstly, the results of the two algorithms for speech enhancement are compared in the frequency domain by means of the spectrogram. Figure 3 shows the spectrogram of clean speech signal, Figure 4 are the spectrograms of noisy speech signals corrupted by the Destroyer noise with SNR=5 dB and F16 noise with SNR=10dB, respectively. Figure 5 and Figure 6 give the results of the enhanced speech signals estimated by MS and TSNE methods. From the obtained results, it is apparent that our proposed approach has a better noise reduction capability, while keeping more of the speech signals energy unchanged than the MS approach.

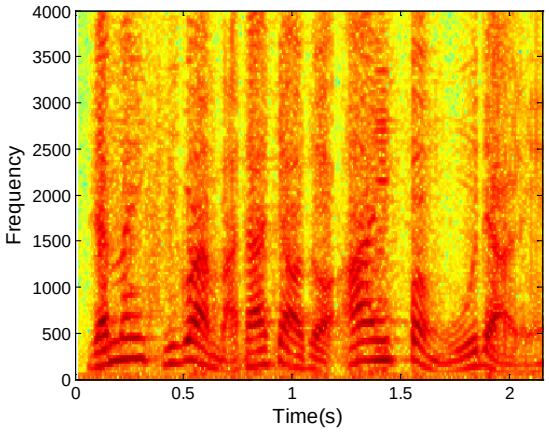
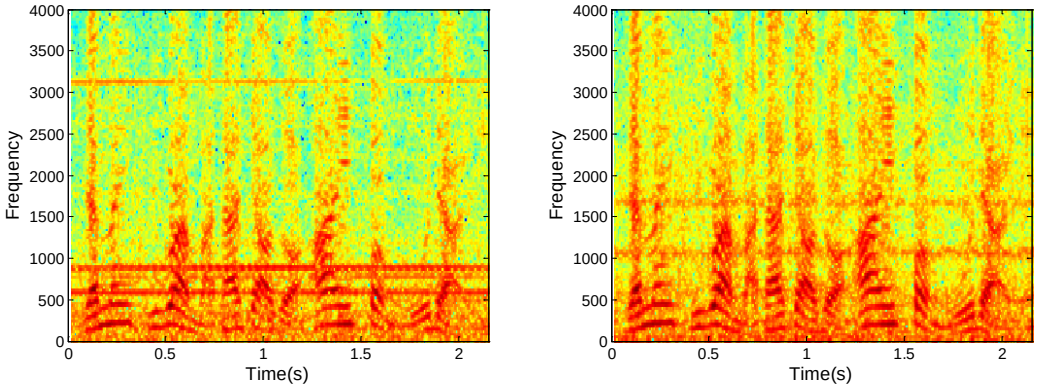
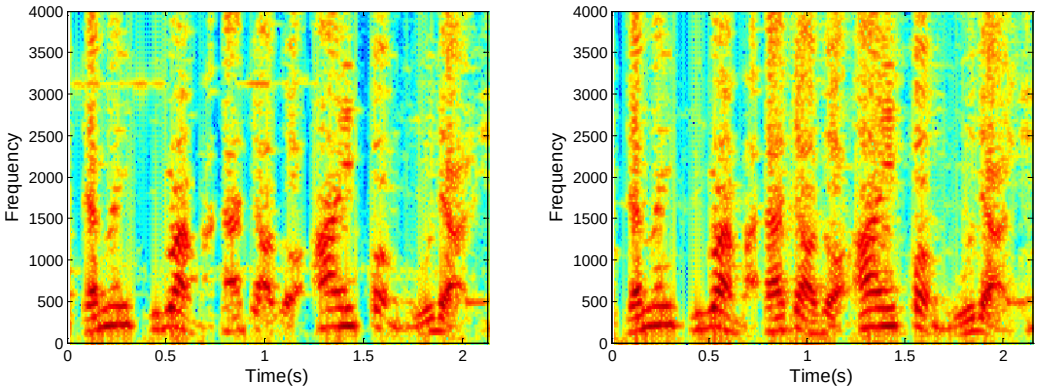


Figure 3. Spectrogram of Clean Speech Signal



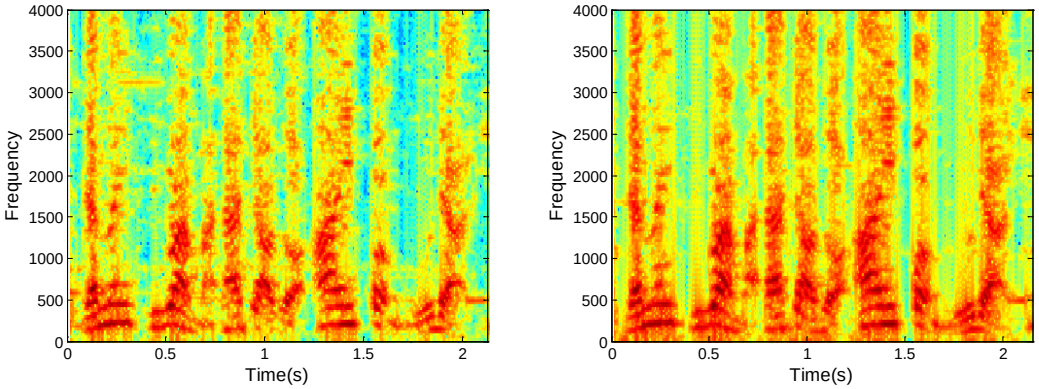
(a) Destroyer Noise with SNR=5 dB **(b) F16 Noise with SNR=10 dB**

Figure 4. Spectrogram of Noisy Speech Signal



(a) Destroyer noise with SNR=5 dB **(b) F16 noise with SNR=10 dB**

Figure 5. Spectrograms of Enhanced Speech Signals by MS Method



(a) Destroyer noise with SNR=5 dB **(b) F16 noise with SNR=10 dB**

Figure 6. Spectrograms of Enhanced Speech Signals by TSNE Method

Table 2. Comparison of SEGSNR of Enhanced Signals

Noise type	Input SNR	SEGSNR (dB)	
		MS method	TSNE method
White	0 dB	4.0190	4.1863
	5 dB	5.6831	5.9468
	10 dB	8.0545	8.3186
F16	0 dB	3.6365	3.8856
	5 dB	5.5044	5.8674
	10 dB	8.1322	8.4338
Babble	0 dB	3.0458	3.3048
	5 dB	5.3848	5.5828
	10 dB	8.3560	8.6901
Destroyerengine	0 dB	3.8741	4.2271
	5 dB	5.5152	6.3219
	10 dB	8.3186	8.7567

The segmental SNR (SEGSNR) and log-spectral distortion (LSD) measures are also adopted for the objective evaluation. For the segmental SNR, only frames with segmental SNR values greater than -10 dB and less than 35 dB are considered. Table 2 gives the output SEGSNR results of the enhanced speech signals obtained using MS and the proposed TSNE algorithms in various noise conditions and levels. The results of the log spectral distance are showed in Table 3. From the two tables, we can observe that the proposed algorithm always has a higher SEGSNR and lower LSD as compared to the MS algorithm in all tested environmental conditions. In order to evaluate the subjective quality of the proposed method, a set of informal listening tests are carried out. Opinion scores were recorded by fifteen listeners and averaged to yield mean opinion score (MOS). The results indicate that the MOS of the MS and the proposed methods at the input SNR=18dB are 3.28 and 3.41, respectively. These results confirm that the TSNE approach is an effective approach for noisy speech enhancement.

Table 3. Comparison of LSD of Enhanced Signals

Noise type	Input SNR	LSD (dB)	
		MS method	TSNE method
White	0 dB	14.8135	14.5863
	5 dB	14.1536	12.4167
	10 dB	10.7585	9.1429
F16	0 dB	11.9937	11.7548
	5 dB	11.1857	9.7249
	10 dB	7.4923	7.3795
Babble	0 dB	14.8089	13.5368

	5 dB	8.6098	7.0511
	10 dB	5.6557	5.1967
Destroyerengine	0 dB	11.5508	10.3809
	5 dB	10.8543	8.5170
	10 dB	7.4419	7.3938

5. Conclusions

In this paper, we have analyzed the conventional approaches applied to noise estimation for noisy speech enhancement in various noise environments. Then we have presented the TSNE algorithm, a novel noise estimation algorithm that refines the result of MS method using two steps under Gaussian distribution assumption. On the basis of the relative estimation error and a number of objective and subjective evaluation tests, the performance of the proposed algorithm was found to be superior to that of the conventional MS approach.

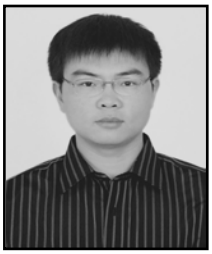
Acknowledgements

This work was supported in part by NSFC under Grant Nos. 61005021, 61201457 and A Project of Shandong Province Higher Educational Science and Technology Program under contract J12LN27.

References

[1] X. Hu, S. Wang, C. Zheng and X. Li, "A Cepstrum-based Preprocessing and Oostprocessing for Speech Enhancement in Adverse Environments", Applied Acoustics, vol. 74, no. 12, **(2013)**, pp. 1458-1462.
[2] Y. D. Cho, K. Al-Naimi and A. Kondoz, "Mixed Decision-based Noise Adaptation for Speech Enhancement", Electron. Lett., vol. 37, no. 8, **(2001)**, pp. 540-542.
[3] J. Choi, J. Chang, D. K. Kim and S. Kim, "Speech Enhancement Based on Adaptive Noise Power Estimation Using Spectral Difference", IEICE Trans. Fundamentals, vol. 94-A, no. 10, **(2011)**, pp. 2031-2034.
[4] J. Chang, "Noisy Speech Enhancement Based on Improved Minimum Statistics Incorporating Acoustic Environment-awareness", Digital Signal Processing, vol. 23, no. 4, **(2013)**, pp. 1233-1238.
[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Trans. Speech Audio Process., vol. 9, no. 5, **(2001)**, pp. 504–512.
[6] C. R. Hendriks, J. Jensen and R. Heusdens, "Noise Tracking Using DFT Domain Subspace Decompositions", IEEE Trans. Speech and Audio Process., vol. 16, no. 3, **(2008)**, pp. 541–553.
[7] S. Rangachari and P. C. Loizou, "A Noise-estimation Algorithm for Highly Non-stationary Environments", Speech Commun., vol. 48, no. 2, **(2006)**, pp. 220–231.
[8] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", IEEE Trans. Speech and Audio Process., vol. 11, no. 5, **(2003)**, pp. 466–475.
[9] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based Noise Power Estimation with Low Complexity and Low Tracking Delay", IEEE Trans. Audio, Speech and Language Process., vol. 20, no. 4, **(2012)**, pp. 1383–1393.
[10] Y. Park and J. Chang, "A Probabilistic Combination Method of Minimum Statistics and Soft Decision for Robust Noise Power Estimation in Speech Enhancement", IEEE Signal Process. Lett., vol. 15, no. 1, **(2008)**, pp. 95-98.

Authors



Shifeng Ou received the Ph.D. degree in communication and information system from Jilin University, China in 2008. Currently, he is an associate professor at Yantai University, China. His research interest covers blind and speech signal processing.

Chao Geng received the B.S. degree in electrical engineering from Yantai University, China in 2011. Currently, he is a M.S. candidate at Yantai University, China. His research interests include speech and blind information processing.

Xianyun Wang received the M.S. degree in signal and information processing from Yantai University, China in 2012. His research interests include speech and blind signal processing.

Ying Gao received the Ph.D. degree in geodetection and information technology from Jilin University, China in 2008. Currently, she is an associate professor at Yantai University, China. Her research interest is signal and information processing.