A Comparative Study on Single-Channel Noise Estimation Methods for Speech Enhancement

Hadi Veisi
Department of Computer Engineering
Sharif University of Technology
Tehran-Iran
veisi@ce.sharif.edu

Hossein Sameti
Department of Computer Engineering
Sharif University of Technology
Tehran-Iran
sameti@sharif.edu

Abstract— **This paper studies a number of well-known noise estimation techniques and provides a comparative performance analysis of them in speech enhancement platform. Two types of evaluation data that simulate consistent and inconsistent noisy conditions are prepared in the presence of six noise types at different SNR levels. The performance of speech enhancement systems and the spectrum distance of the estimated and original noise spectrums are used as evaluation criteria. The evaluations indicate that a simple VAD method outperforms noise estimation methods in most of the consistent noisy conditions.**

Keywords- noise estimation; speech enhancement; minimum controlled recursive average; minimal tracking;VAD

I. INTRODUCTION

Noise estimation is an essential part of the majority of speech processing applications such as speech enhancement, hearing aids, communication systems and noise robustness techniques for speech recognition. The performance of these systems are highly dependent on the noise estimation ability such that, for example, if we have an accurate noise estimation technique for speech enhancement it is possible to realize a high performance speech enhancement only by using a simple noise reduction approach such as spectral subtraction [1]. Regarding this fact, many researchers have proposed various techniques for noise estimation [2]-[11].

Voice activity detection (VAD) can be considered as a simple and old approach for noise estimation in which the frames of signal are labeled as speech presence or speech absence (i.e., silence or noise). To estimate the noise, the speech absence segments of signal are considered as noise and the noise characteristics are updated during these periods. Various VAD algorithms based on energy, zero crossing rate (ZCR), cepstral features, linear prediction coding (LPC) parameters [12] and statistical modeling [13],[14] have been proposed. Also there are other VAD methods such as ITU-T G.729 Annex B [15] and the European Telecommunications Standards Institute Adaptive Multi-Rate (AMR) VAD option 2 [16] that are using a combination of different parameters.

VAD can be considered as a *discrete* noise estimation approach that labels signal frames in binary format (1 for speech presence and 0 for speech absence). This approach fails to provide a robust estimation of noise for low signal-to-noise ratio (SNR) levels and in the presence of non-stationary noises [2]. Therefore, another class of noise estimation that can be considered as *continuous* noise estimation is commonly used in speech enhancement. In this approach, noise characteristics are continuously approximated even during speech presence frames. There are many proposed continuous noise estimation methods such as time-recursive averaging methods [3]-[6], minimal tracking algorithms [7]-[10], quantile-based [11] and histogram-based techniques [2]. It is shown that these methods estimate noise characteristics effectively [2]. In this paper, we have compared the performance of a discrete noise estimation technique (i.e., a VAD) and seven continuous noise estimation methods. An objective comparison of a number of continuous noise estimation methods is done in [2], however in contrast with that study; in this paper, the performance of these techniques is evaluated in the speech enhancement platform using two designed data sets in the presence of various noise types at different SNR levels. The noise estimation methods are integrated in three speech enhancement methods, spectral subtraction [1], minimum mean square error (MMSE)-based short-time spectral amplitude (STSA) estimator [17] and MMSE-based log-spectral amplitude (LSA) estimator [18] and their performances are evaluated. In addition, a spectral distance measure is also used as a criterion to determine the difference of the original noise spectrum and the estimated noise spectrum.

In the continuant of the paper, the noise estimation methods used in our experiments are briefly reviewed in Section 2. In Section 3, the evaluation platform including the speech enhancement methods and the spectral distance method are given. Section 4 provides the experimental results and the analyses. Finally, summary and conclusions of the paper are given in Section 5.

II. NOISE ESTIMATION METHODS

Let $y(t)=s(t)+n(t)$ denotes the t^{th} frame of a noisy speech signal where $s(t)$ and $n(t)$ indicate clean speech and noise frames, respectively. Assuming the independence of clean speech and noise, the power spectral density (psd) of noisy speech frame is achieved as $|Y(t,f)|^2=|S(t,f)|^2+|N(t,f)|^2$ where f indicates the frequency bin index. According to this equation, the final goal of noise estimation methods is to approximate $|N(t,f)|^2$. In this paper, following eight noise estimation methods are considered.

1. Minimum tracking (MINT) [7], [8]

978-1-4673-5119-5/12/$31.00 ©2012 IEEE

2. Continuous minimum tracking (MINTC) [9]
3. Connected frequency regions (CONFR) [10]
4. Minimum controlled recursive average (MCRA) [3]
5. Improved MCRA (IMCA) [4]
6. A variant of MCRA (MCRA2) [5]
7. Weighted spectral average (WSA) [6]
8. An energy-based VAD method

The first two methods are algorithms of minimal tracking approach. The basic idea of these methods is that by tracking the minimum of noisy speech power in each frequency band, the noise level can be estimated in that frequency band. This idea has come from this assumption that the power of noisy speech signal decay to the noise level at each frequency band [8]. The MINT method was first proposed by Martin in [7] and then improved in [8]. In this paper the refined version of [8] is used. The MINTC method [9] is an extension of MINT to resolve the inability of MINT to respond to fast changes of noise spectrum. The CONFR algorithm [10] is as also an extension of MINT in which a technique for speech presence detection is used to find connected time-frequency regions of speech presence. Then it is applied to find bias compensation factors for minimal tracking noise estimation.

The MCRA, IMCRA, MCRA2 and WSA methods are all from time-recursive averaging family [2]. The basic idea behind this approach is that noise signal has non-uniform effects on speech spectrum in different frequency bands. Therefore, each frequency band of the estimated noise can be updated whenever the probability of speech absence is high at that frequency band. All of the time-recursive averaging algorithms employ the following equation to approximate the noise, where $\hat{N}(t,f)$ denote the estimated noise psd and $\alpha(t,f)$ is a smoothing factor.

$$\hat{N}(t,f) = \alpha(t,f)\hat{N}(t-1,f) + (1-\alpha(t,f))\,|Y(t,f)|^2 \qquad (1)$$

The various time-recursive averaging algorithms [3]-[6] are only different in the method that they use to estimate $\alpha(t,f)$. To calculate the smoothing factor in MCRA, the ratio of noisy speech psd to its local minimum is estimated and it is compared to a threshold value. In IMCRA, $\alpha(t,f)$ is calculated using likelihood ratio approach and a priori probability of speech absence is also estimated for this purpose. The value of this probability assumed to be fixed in MCRA. MCRA2 is the same as MCRA but it uses a different method to calculate the minimum of noisy psd. In WSA, $\alpha(t,f)$ is fixed and a simple method is used to control the update of noise psd. To update the noise psd in this method, estimated a posteriori SNR is compared to a threshold value.

The VAD that is used in this paper is a simple energy based method that assumes first five frames of the noisy signal is speech absence and the average of their psd are considered as the initial estimate of noise spectrum. For each frame of noisy speech, by comparing the estimated segmental SNR of that frame against a threshold value it is decided to update the noise psd or not. To update the noise psd, the same equation as (1) with the fixed value for $\alpha(t,f)$ is used.

III. Evaluation Platform

To perform a comparative performance evaluation of the noise estimation methods, two types of experiments are used.

1. The mentioned noise estimation methods in Section 2 are integrated in speech enhancement systems and the performance of the speech enhancement systems are studied to compare the performance of noise estimation methods. For this purpose, three renowned speech enhancement methods including spectral subtraction [1], MMSE-STSA [17] and MMSE-LSA [18] are used.
2. The estimated noise psd by noise estimation methods are compared to the original noise psd. To this end, the distance measure of Eq. (2) is used that calculates root mean square log spectral distance (RMS-LSD) between the estimated noise psd, $\hat{N}(t,f)$, and the corresponding original noise psd, $|N(t,f)|^2$. In this equation, T and F denote number of noise frames and number of frequency bands, respectively.

$$D = \mathrm{sqrt}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\left(10\log_{10}\frac{\hat{N}(t,f)}{|N(t,f)|^2}\right)^2 \qquad (2)$$

IV. Experimental Results

In this section, the experimental setup and the results of evaluations are presented. In the experiments, the values of the parameters for the noise estimation methods and for the speech enhancement techniques are the same as their values in the original references. The value of smoothing factor for the VAD method is $\alpha(t,f) = 0.98$ in all experiments.

Two data sets are prepared to perform the evaluations that are described below. All noisy speech signals are framed by 20-ms length with 50% of overlap, and then they are windowed using Hamming window. The standard overlap-and-add method is used to reconstruct the enhanced speech signals in which the phase information of noisy signals is used as the phase of enhanced speech signals. The results of the enhancement systems are evaluated using objective criteria, overall SNR (in dB) [2] and perceptual evaluation of speech quality (PESQ) [19]. In all evaluations, for both mentioned criteria, the *improvement* values (i.e., the difference to the original values) are reported. The values that are reported for RMS-LSD are the original values.

A. Evaluation data sets

To evaluate the performance of the noise estimation methods two data sets are prepared and used. In the first data set that it is called as *consistent environment* in this paper, six speech sentences from six speakers (four males and two females) are selected from TIMIT corpus. These clean speech sentences are down-sampled to 8 kHz and six noise types are added to them at five SNR levels. The noise types are white, office, babble, Volvo, F16 and machinegun, and the SNR levels are -10, 5, 0, 5 and 10 dB. Therefore, each sentence of this data set only includes one noise type at a specific SNR level. This data set simulates 30 different environments each of them containing six sentences.

For the second data set that we identify it as *inconsistent environment*, two sentences from two different environments

(e.g., having different noise types and/or different SNR levels) are concatenated and formed a new sentence. This data set is prepared to evaluate the ability of noise estimation methods in responding to changes in the environment including changing in the background noise and SNR level. To prepare this data set, 18 different environments of the previous data set (i.e., *consistent environment*) including six noise types at three SNR levels, 0, 5 and 10 dB are selected. For each environment, two sentences are chosen and each sentence is concatenated to the sentences of all other environments. It results in 36*34=1224 sentences for this data set, each of the sentences contains two different noise types at the same/different SNR level.

B. *Evaluation results on consistent environment*

In this section, the performance results of the three speech enhancement methods for the *consistent environment* data set are presented. In each speech enhancement method, eight mentioned noise estimation methods are used to estimate the noise psd (e.g., in spectral subtraction) or other related parameters such as a priori SNR and posteriori SNR (e.g., in MMSE-STSA and MMSE-LSA). In this section, the results of the speech enhancement systems based on overall SNR improvement (in dB) and PESQ improvement (in MOS) are reported.The resultsofRMS-LSD arenot given for the brevity.

The evaluation results of spectral subtraction using the various noise estimation methods and in the presence of different noise types at different SNR levels are given in Figure 1. As the results indicate, various noise estimation methods have resulted in different performances for different noise types and different evaluation criteria. For white, Volvo and F16 noise types, MCRA has generally given the higher performance. Interestingly, the simple VAD has achieved higher SNR improvement in the presence of office, babble and high SNR levels of machinegun noises.

In Figure 2. the results of MMSE-STSA are given. Inconsistently with the results of spectral subtraction, CONFR method has resulted higher improvement for white and babble noises and simple VAD has provided higher performance for Volvo, F16 and machinegun noises. The VAD has had the higher PESQ improvement for office noise and high SNR levels of babble noise. Similarly, the experimental results of MMSE-LSA method are shown in Figure 3. As it could be observed, the VAD has resulted in higher performance almost for all noise types and for both evaluation criteria.

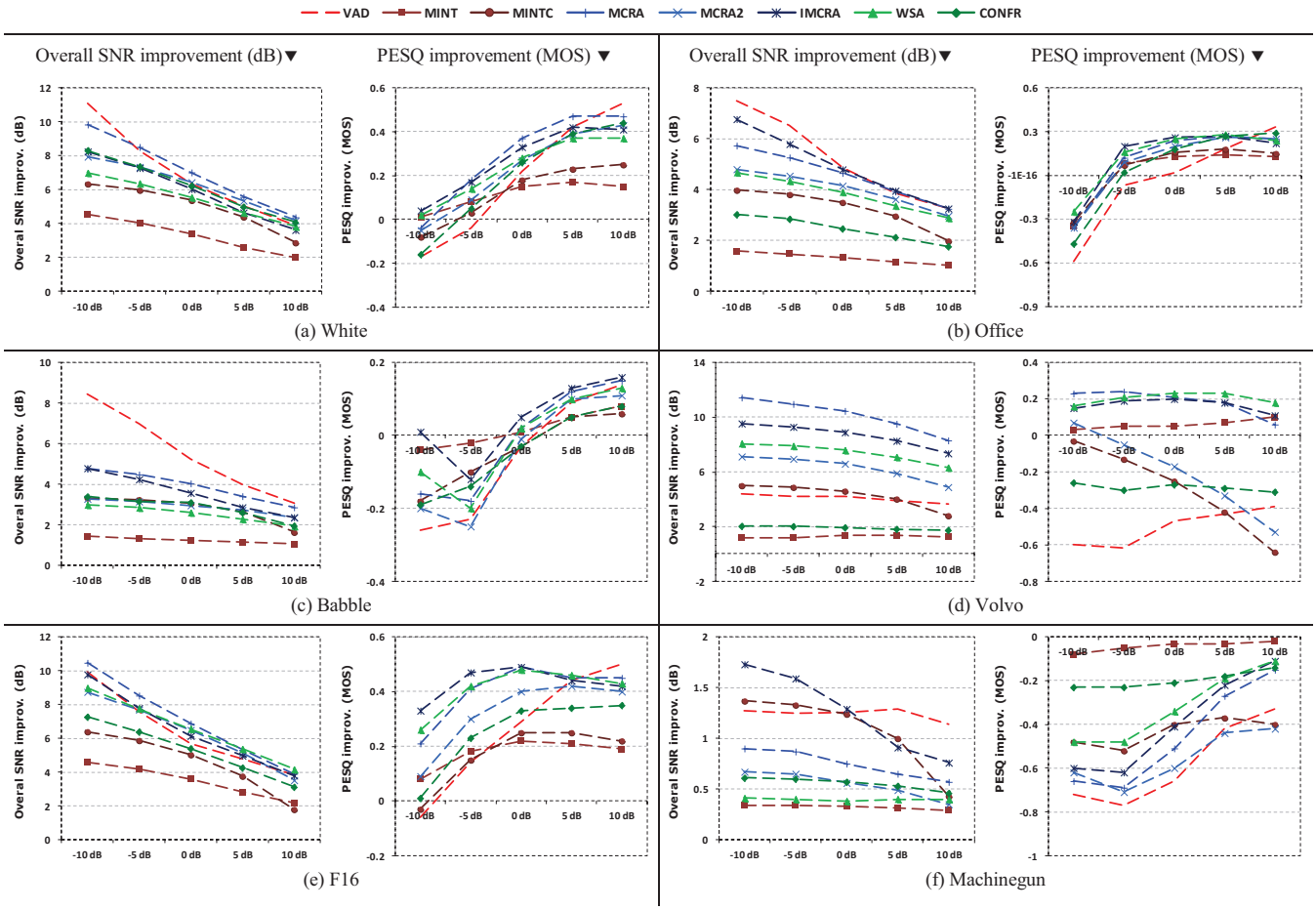


Figure 1. Performance of spectral subtraction based on overal SNR improvement (left) and PESQ improvement (right) in the presence of six noise types at different SNR levels using various noise estimation methods

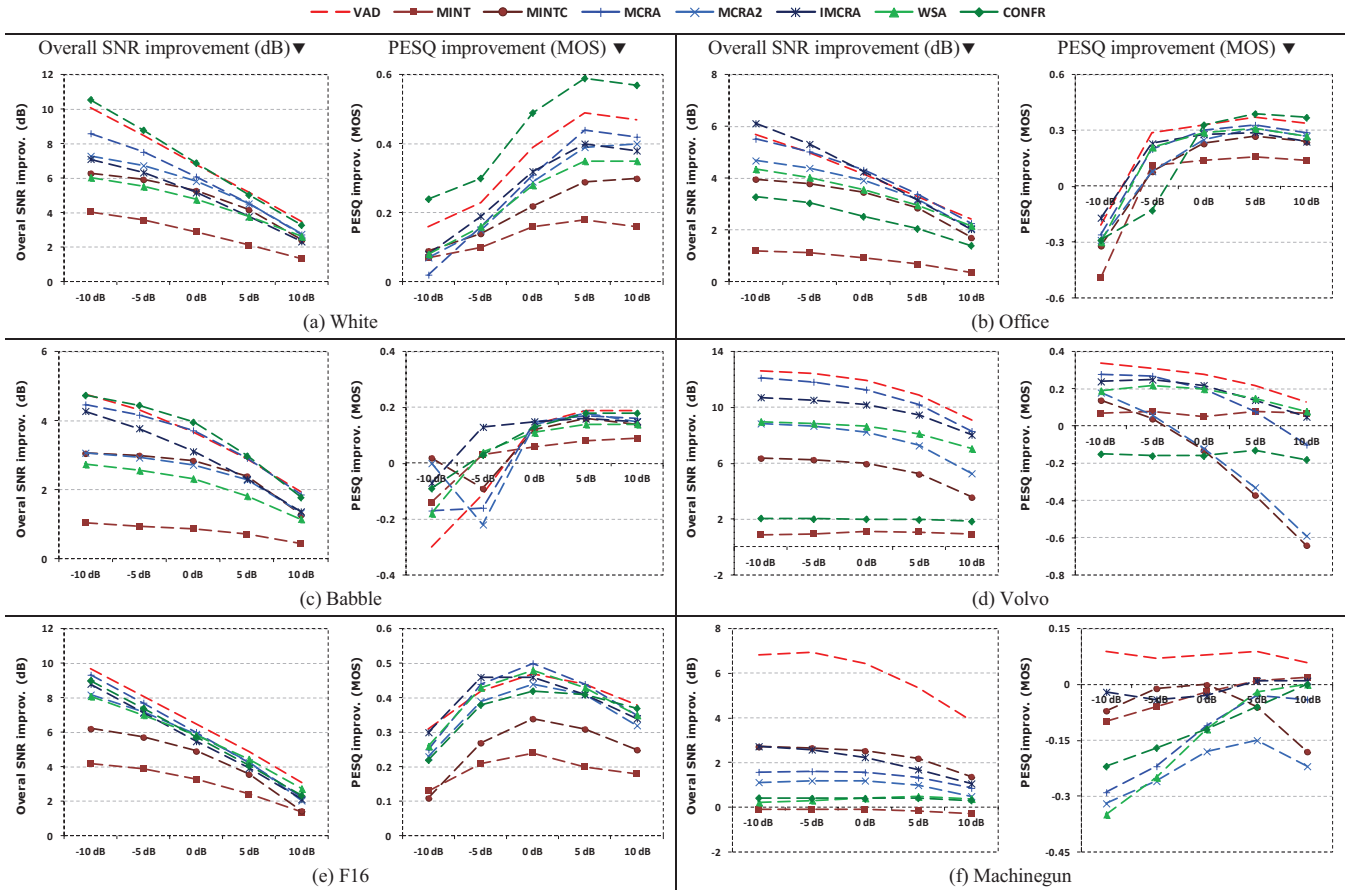
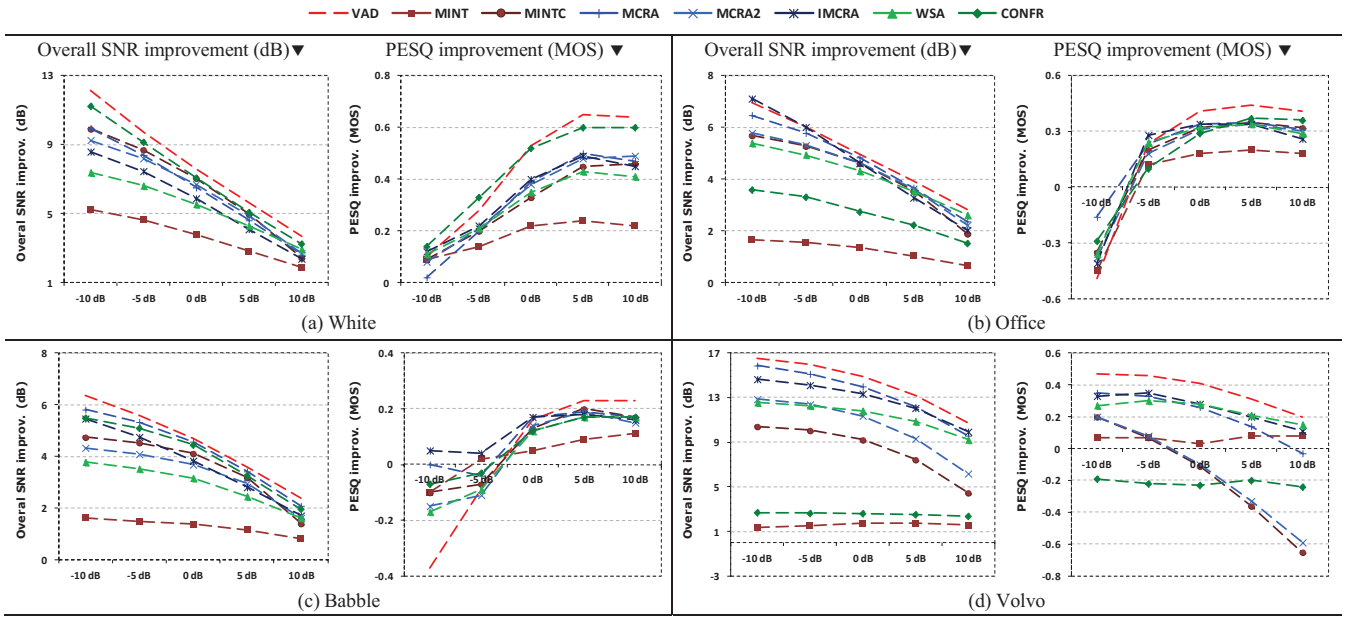
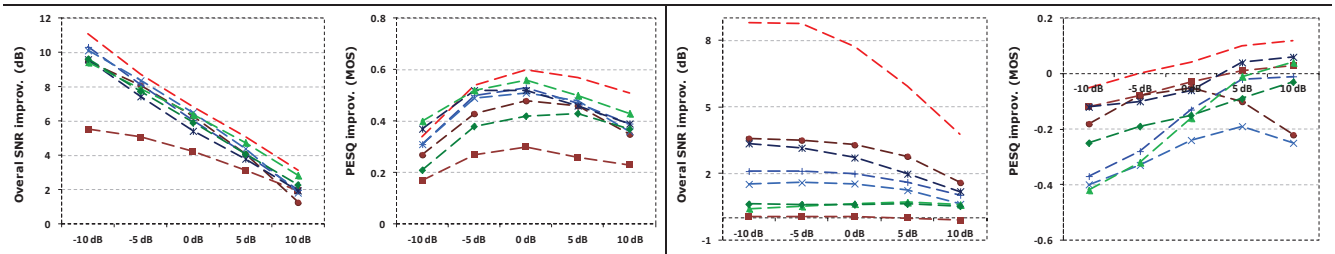


Figure 2. Performance of MMSE-STSA based on overal SNR improvement (left) and PESQ improvement (right) in the presence of six noise types at different SNR levels using various noise estimation methods





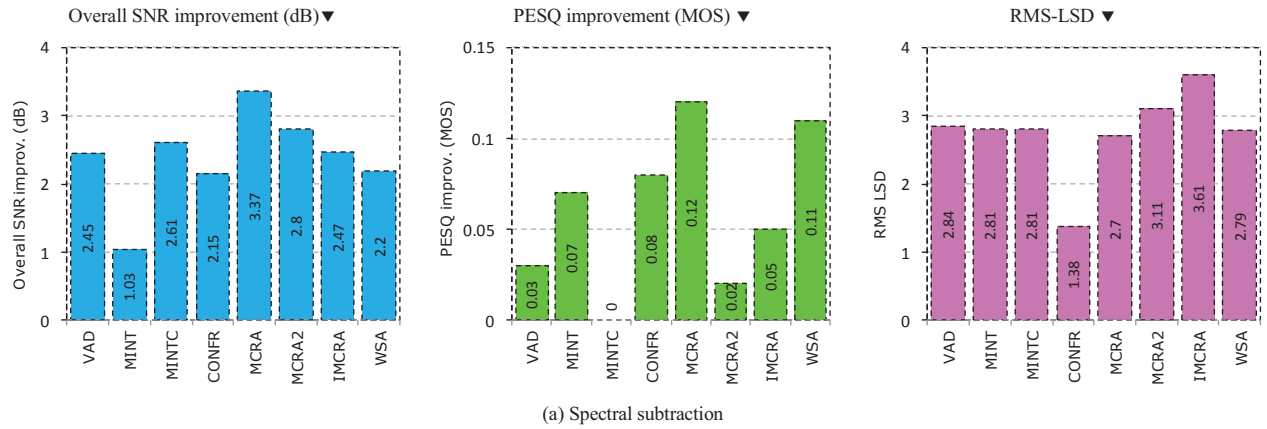
(e) F16 (f) Machinegun

Figure 3. Performance of MMSE-LSA based on overal SNR improvement (left) and PESQ improvement (right) in the presence of six noise types at different SNR levels using various noise estimation methods

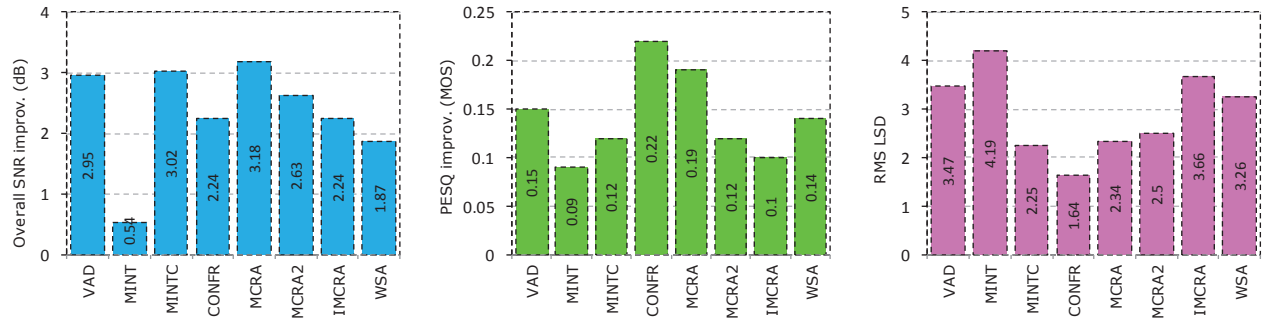
C. *Evaluation results on inconsistent environment*

To evaluate the ability of noise estimation methods in tracking the changing of noise type and noise power the *inconsistent environment* data set is used. In these evaluations, all sentences of the set are used and the average values of overall SNR improvement, PESQ improvement and RMS-LSD value are computed. The values of these criteria are shown in Figure 4. for (a) spectral subtraction, (b) MMSE-STSA and (c) MMSE-LSA. The results of spectral subtraction indicate that MCRA has resulted in higher improvement for SNR and PESQ but CONFR has produced lower RMS-LSD. Based on the RMS-LSD value in Figure 4. (a) MCRA has also resulted in the lowest spectrum distortion after CONFR. For MMSE-STSA method, CONFR is the superior noise estimation method according to PESQ and RMS-LSD measures. However, MCRA has resulted in the higher SNR improvement. The results of MMSE-LSA method show that MINTC, MCRA and CONFR are better than the other methods based on SNR improvement, PESQ improvement and RMS-LSD, respectively.

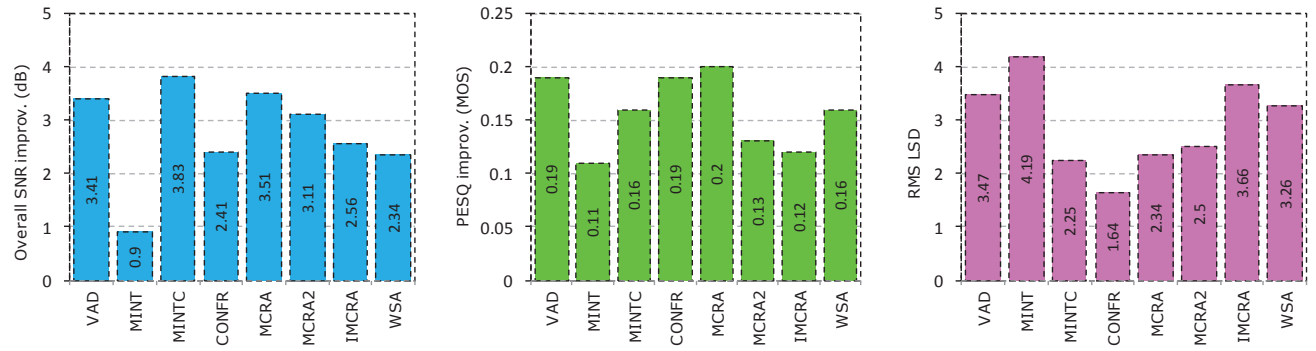
According to the results of the previous section and the results of Figure 4. , it can be observed that the VAD has achieved acceptable performance in spite of its simplicity. However, the continuous noise estimation methods have generally done the noise estimation better than the VAD in the inconsistent environments.



(a) Spectral subtraction



(b) MMSE-STSA



(c) MMSE-SLA

Figure 4. Evaluation results of noise estimation methods using (a) spectral subtraction, (b) MMSE-STSA and (c) MMSE-LSA.

V. SUMMARY AND CONCLUSIONS

In this paper, the performances of eight noise estimation methods were studied in the platform of speech enhancement using three known methods including spectral subtraction, MMSE-STSA and MMSE-LSA. For evaluations, two types of data were prepared to simulate consistent and inconsistent environments using six noise types at different SNR levels. The results showed that there was not a noise estimation method that outperformed the others in all conditions. Generally, different methods resulted in different performances in the presence of various noise types and by using different speech enhancement methods. The simple VAD has interestingly outperformed the other methods in the majority of consistent noisy environments. However, the continuous noise estimation methods achieved higher performance in inconsistent noisy conditions.

ACKNOWLEDGMENT

This work was supported by Iran's National Elite Foundation.

REFERENCES

[1] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'79), 1979, pp. 208-211.

[2] P.C. Loizou, Speech Enhancement: Theory and Practice, Boca Raton, FL: CRC Press, 2007.

[3] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Processing Letters, vol. 9, no. 1, 2002, pp. 12-15.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging." IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, 2003, pp. 466-475.

[5] S. Rangachari, and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," Speech Communication, vol. 28, 2006, pp. 220-231.

[6] H. Hirsch, and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95), 1995, pp.153-156.

[7] R. Martin, "Spectral subtraction based on minimum statistics," In Proceding of 7th European Signal Processing Conference (EUSIPCO'94), Sep 1994, pp. 1182-1185.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5, 2001, pp. 504-512.

[9] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," In Proceeding of Eurospeech, vol. 2, 1995, pp. 1513-1516.

[10] K. Sorensen, and S. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," EURASIP Journal on Applied Signal Processing, vol. 18, 2005, pp. 2954-2964.

[11] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00), vol. 3, June 2000, pp. 1875–1878.

[12] S. G. Tanyer and H. Özer, "Voice Activity Detection in Nonstationary Noise," IEEE Transaction on Speech and Audio Processing, vol. 8, July 2000, pp. 478–482.

[13] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection", IEEE Signal Processing Letter, vol. 6, no. 1, Jan 1999, pp. 1-3.

[14] S. Gazor and W. Zhang, "A Soft Voice Activity Detector Based on a Laplacian-Gaussian Model," IEEE Transaction on Speech and Audio Processing, 2003, 11, pp. 498505.

[15] ITU-T Recommendation, G.729, Annex B, "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70", November 1996.

[16] ETSI EN 301 708 V7.1.1, Digital cellular telecommunications system (Phase 2+); "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels: General Description", 1999.

[17] Y. Ephraim, and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator," In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84), 1984, pp. 1109-1121.

[18] Y. Ephraim, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'85), 1985, pp. 443-445.

[19] "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", ITU-T Recommendation P.862, Feb. 2001.

[20] H. Veisi, and H. Sameti, "An HMM-based Voice Activity Detector with High Speech Detection Rate for Speech Enhancement," IET Signal Processing, vol. 6, no. 1, 2012, pp. 54-63.