HHS Public Access

Author manuscript

Health Innov Point Care Conf. Author manuscript; available in PMC 2020 July 23.

Published in final edited form as:

Health Innov Point Care Conf. 2017 November ; 2017: 32–35. doi:10.1109/hic.2017.8227577.

Formant Frequency-based Speech Enhancement Technique to improve Intelligibility for hearing aid users with smartphone as an assistive device

Gautam S Bhat, **Nikhil Shankar**, **Chandan K A Reddy [Student Members, IEEE]**, **Issa M.S Panahi [Senior Member, IEEE]**

Statistical Signal Processing Research Laboratory (SSPRL), Department of Electrical and Computer Engineering, The University of Texas at Dallas.

Abstract

In this paper, we present a Speech Enhancement (SE) method implemented on a smartphone, and this arrangement functions as an assistive device to hearing aids (HA). Many benchmark single channel SE algorithms implemented on HAs provide considerable improvement in speech quality, while speech intelligibility improvement still remains a prime challenge. The proposed SE method based on Log spectral amplitude estimator improves speech intelligibility in the noisy real world acoustic environment using the priori information of formant frequency locations. The formant frequency information avails us to control the amount of speech distortion in these frequency bands, thereby controlling speech distortion. We introduce a 'scaling' parameter for the SE gain function, which controls the gains over the non-formant frequency band, allowing the HA users to customize the playback speech using a smartphone application to their listening preference. Objective intelligibility measures show the effectiveness of the proposed SE method. Subjective results reflect the suitability of the developed Speech Enhancement application in real-world noisy conditions at SNR levels of −5 dB, 0 dB and 5 dB.

Keywords

Formant frequencies; Speech Enhancement; Hearing Aid; Smartphone; Speech intelligibility

I. Introduction

According to World Health Organization (WHO), 360 million people across the globe have disabling hearing loss. Statistics obtained by National Institute on Deafness and Other Communication Disorders (NIDCD) show that approximately 15% of American adults (37.5 million) aged 18 and over report some concern in hearing. About 2 to 3 out of every 1,000 children in the United States are born with a detectable level of hearing loss. Researchers in academia and industry are developing viable solutions in the form of Hearing Aid Devices (HADs). However, a recent study [1] shows that the efficiency of these devices degrades in the presence of background noises. But due to the limitations concerning size, power and its processors, HADs lack processing capability of complex yet useful signal processing algorithms [1-3]. One practicable solution to overcome this problem is to use smartphones that are widely available and can perform complex computations with sophisticated

processors, as assistive tools for HADs. The noisy speech can be captured using the smartphone microphone; enhanced using the algorithms running on the processor of smartphone and wirelessly transmit the enhanced speech to HADs.

For example, recently, Apple has come up with new HA feature on iPhone called Live Listen [4] to enhance the overall quality of speech perceived by hearing impaired using smartphone.

Speech Enhancement (SE) is a key component in the HAD signal processing pipeline. Several studies have revealed that SE algorithms would improve the listening comfort for the HAD users. The most challenging task in single channel SE is to reduce the background noise without introducing speech distortion. Traditional SE methods like Spectral Subtraction [5] and statistical model based methods proposed by Ephraim and Malah [6-7] can be implemented in real-time. However, these algorithms do not improve speech intelligibility adequately. Recent developments include SE based on deep neural networks (DNN) [8], which is not suitable for real-time applications, as it requires rigorous training data and extensive training period. Studies on speech intelligibility [9] show that speech intelligibility improvement is inadequate in many widely used algorithms. Researchers have shown that ideal binary mask in SE could improve intelligibility [10], but accurate estimation of the binary mask is challenging, especially in lower SNR conditions. In [11], Loizou explained two types of distortions that play a significant role in speech intelligibility. Formant frequency trajectories are one of the major acoustical cues for identification of vowels, nasal consonants, and Diphthongs [12] which represent typical characteristics of a speech signal. Several studies have been carried out to increase the intelligibility in speech coders [13] using formant frequencies. Recently in [14], formant shaping method is used to improve speech intelligibility.

The formant frequency based SE method presented in this work makes use of the *priori* information of formants to apply scaled value of the Minimum mean square error Log spectral amplitude estimator (Log-MMSE) gain function on the acoustically unimportant bands which in turn suppresses the background noise without inducing speech distortion and any residual musical noise. The ‘scaling’ factor for the gain in the non-formant locations can be varied in real time allowing the hearing impaired smartphone user to control the amount of noise suppression and speech distortion. The proposed SE method is computationally efficient, and inexpensive. Objective and subjective evaluations of the proposed method show good improvement in quality and intelligibility reveals the overall usability of the developed algorithm.

II. Proposed Formant based Real time Log-Spectral Amplitude Estimator

In this section we describe the developed speech processing method implemented on Smartphone. Fig. 1 shows the block diagram of the proposed method that runs on the smartphone in real time (Timing is discussed in Section III).

A. Log-Spectral Amplitude Estimator

In the Log-MMSE method, speech and noise models are considered to be statistically independent Gaussian Random Variables. The goal is to minimize the mean squared error of log magnitude spectra between estimated and true speech signals. Considering the additive mixture model for noisy speech $y(n)$, with clean speech $s(n)$ and noise $w(n)$, as

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

The noisy k^{th} Discrete Fourier Transform (DFT) coefficient of $y(n)$ for frame λ is given by,

$$Y_k(\lambda) = S_k(\lambda) + W_k(\lambda) \tag{2}$$

Where S and W are the clean speech, and noise DFT coefficients respectively. In polar coordinates, (2) can be written as,

$$R_k(\lambda)e^{j\theta Yk(\lambda)} = A_k(\lambda)e^{j\theta S_k(\lambda)} + B_k(\lambda)e^{j\theta W_k(\lambda)} \tag{3}$$

Where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectra of noisy speech, clean speech, and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{S_k}(\lambda)$, $\theta_{W_k}(\lambda)$ are the phase spectra of noisy speech, clean speech and noise respectively. Looking at the estimator $\hat{A}_k$, which minimizes the distortion measure as explained in [6], the mean-square error of the log-magnitude spectra is given by,

$$E\left\{(log\, A_k - log\, \hat{A}_k)^2\right\} \tag{4}$$

Where, A_k is the k^{th} bin of magnitude spectrum, and $\hat{A}_k$ is the k^{th} bin of estimated clean speech magnitude spectrum. The optimal log-MMSE estimator can be obtained by evaluating the conditional mean of the $log\, A_k$, that is,

$$log\, \hat{A}_k = E\{log\, A_k \mid Y_k(\lambda)\} \tag{5}$$

Hence, the estimate of the speech magnitude is given by,

$$\hat{A}_k = exp(E\{log\, A_k \mid Y_k(\lambda)\}) \tag{6}$$

Solving the above expectation, the final estimate of speech magnitude spectrum according to [6] is given by,

$$\begin{aligned}\hat{A}_k &= \frac{\xi_k}{\xi_k + 1}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k \\ &\triangleq G_{LSA}(\xi_k, v_k)R_k\end{aligned} \tag{7}$$

Where $v_k = \frac{\xi_k}{1+\xi}\gamma_k$ here $\xi_k = \frac{\sigma^2_{S_k}}{\sigma^2_{W_k}}$ is the *a priori* SNR and $\gamma_k = \frac{R^2_k}{\sigma^2_{W_k}}$ is the *a posteriori* SNR. $\sigma^2_{W_k}$ is estimated using a voice activity detector (VAD) [15]. σ_{S_k} is the estimated instantaneous clean speech power spectral density. The optimal phase spectrum is the noisy phase itself $\theta_{S_k} = \theta_{Y_k}$.

B. Formant Frequency Band Estimation

In this paper, we approximate the formant frequency bands by calculating the exact formant locations to enhance the speech intelligibility. The first four formant frequency trajectories ($f_0 - f_3$) of the clean speech or speech degraded with noise at high SNR can be calculated by the method explained in [12] which employs adaptive voice detector, and gender detector for formant extraction from the voice segments of continuous speech. We require a frequency range to approximate the presence of speech and apply considerably less noise suppression over that band. Therefore, the mean of formants $f_0 - f_3$ are calculated for large data sets and mean absolute error for each formant is determined over the data sets to find the frequency band of probable formant location. The frequency band is given by (8)

$$F_X = \left[\left(f_X - \frac{f_a}{2}\right), \left(f_X + \frac{f_a}{2}\right)\right] \tag{8}$$

Where F_X is the frequency band for a particular formant. f_X represents mean formant frequency computed over entire database for $X = 0, 1, 2$ and 3. f_a is the mean absolute error determined for each formant. Thus, we estimate four frequency bands (F_0 *to* F_3) from the respective mean formant locations. The FFT bins corresponding to the four frequency bands are thus calculated.

C. Gain function Customization based on Frequency Bands

In conventional methods, the gain function shown in (7) is usually applied over the entire frequency range. This induces speech distortion if the estimate of gain function is inaccurate. The proposed method allows suppressing more noise on acoustically unimportant bands and far lowering noise suppression on formant frequency bands to retain the integrity of clean speech. Thus, we obtain two different gain functions based on the frequency bands.

$$\widehat{G}_k = \begin{cases} G_{LSA}(\xi_k, v_k), & if\ k \in F_X\ for\ X = 0,\ 1, 2, 3 \\ \delta\, G_{LSA}(\xi_k, v_k), & otherwise \end{cases} \tag{9}$$

Where k represents the k^{th} frequency bin, F_X represents the bins associated with formant frequency bands. δ represents the scaling factor which allows the smartphone user to obtain more noise suppression without speech distortion. The δ ranges from 0 to 1 which the HAD user can adjust in real time based on his/her listening preference under continuously varying acoustical environment.

We know from the literature that the phase is perceptually unimportant. Therefore, we consider the phase of the noisy speech signal for reconstruction. The final clean speech estimate is,

$$\hat{S}_k = \hat{G}_k Y_k \tag{10}$$

The time domain reconstruction signal $\hat{s}(n)$ is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\hat{S}_k$. At lower values of δ there is more noise suppression. When $\delta = 1$ the algorithm acts as basic Log-MMSE method. The inaccuracies in calculating the exact formants does not affect the proposed method to a greater extent as it considers the band approximation. Near approximation of the formant frequency bands can improve the speech intelligibility substantially. Another advantage of the proposed is it does not induce any residual or musical noise so there is no requirement of any post filter after the enhancement. This reduces computational complexity and latency in real time.

III. Real-time Implementation on Smartphone to Function as an Assistive Device to HA

In this work, iPhone 7 running iOS 10.3 operating system is considered as a HA assistive device. Though smartphones come with 2 or 3 mics, manufacturers only allow default microphone (Fig. 2) on iPhone 7 to capture the audio signal, process the data and wirelessly transmit the enhanced signal to the HAD. Xcode [16] is used for coding and debugging of the SE algorithm. The data is acquired at a sampling rate of 48 kHz. Core Audio [17], an open source library from Apple Inc. was used to carry out input/output handling. After input callback, the short data is converted to float, and a frame size of 256 is used for the input buffer. Fig. 2 shows a snapshot of the configuration screen of the algorithm implemented on iPhone 7. When the switch button shown is in ‘OFF’ mode, the application merely plays back the audio through the smartphone without processing it. Switching ‘ON’ the button enables SE module to process the incoming audio stream by applying the proposed noise suppression algorithm on the magnitude spectrum of noisy speech. The enhanced signal is then played back through the HAD. Initially, when the switch is turned on, the algorithm uses a couple of seconds (1-2 sec) to estimate the noise power. Therefore, we assume that there is no speech activity at least for 2 seconds when the switch is turned on. Once the noise suppression is on, we have provided a parameter, which allows more noise suppression without inducing any speech distortion and musical noise. In (10), the gain function depends on δ which needs to be determined empirically. Through our experiments, it is known that the optimal value of δ depend on the type of noisy signal, background noise and acoustic environment. Hence, it is not advisable to fix the values of δ irrespective of changing conditions. In our smartphone application, the user can control this parameter by adjusting its value on the touch screen panel of smartphone to attain more noise suppression based on their level of hearing comfort. The processing time for a frame of 10 ms (480 samples) is 1.4 ms. Computation efficiency of the developed algorithm allows the smartphone application to consume very less power. Through our experiments, we found that a fully charged smartphone can run the application seamlessly for 6.3 hours on iPhone 7 with 1960 mAh

battery. We use Starkey live listen [18] to stream the data from iPhone to the HAD. The audio streaming is encoded for Bluetooth Low Energy consumption.

IV. Experimental Results

A. Objective Evaluation

To the best of our knowledge, there are no SE algorithms published that use formant frequencies and a user adjustable parameter to attain more noise suppression in particular bands and thereby personalizing and retaining the speech quality and intelligibility simultaneously in real time varying noisy conditions. We therefore fix the values of few parameters and evaluate the performance of the proposed method by comparing with Log-MMSE [7] method which has been known to show promising results. We note that the proposed SE method is an improved extension of the Log-MMSE method. In our experiment, the formant frequency bands were calculated by determining the mean of the formant locations and mean absolute error for over 200 clean speech files from TIMIT database. The experimental evaluations are performed for 3 different noise types: machinery, multi-talker babble and traffic noise. The reported results are the average over 30 sentences from TIMIT database. For objective evaluation, all the files are sampled at 16 kHz, and 20 ms frames with 50% overlap are considered. As objective evaluation criteria, we choose the perceptual evaluation of speech quality (PESQ) [19] as it had better correlation with subjective tests than the other objective measure. Another objective measure is Segmental SNR (SegSNR) as the amount of noise reduction, residual noise and speech distortion is generally measured by SegSNR. Fig. 3 shows the plots of PESQ and SegSNR versus SNR for the 3 noise types. The δ was adjusted empirically to give the best values for both PESQ and SegSNR and for each noise type. PESQ and SegSNR values show statistically significant improvements over Log-MMSE method for all three noise types considered. Objective measures reemphasize the fact that the proposed method achieves comparatively more noise suppression without distorting speech.

B. Subjective test setup and results

Although objective measures give useful evaluation results during the development phase of our method, they give very little information about the practical usability of our application. We performed Mean Opinion Score (MOS) tests [20] on 11 normal hearing subjects including male and female adults. Subjects were presented with noisy speech and enhanced speech using the proposed, and Log-MMSE methods at SNR levels of −5 dB, 0 dB and 5 dB. Subjects were asked to rate between 1 and 5 for each audio file based on how pleasant it is and how many words they can identify. This test provided a good comparison between proposed method and Log-MMSE method. The key contribution of this paper is in providing the user the ability to customize the parameter to suppress more noise without compromising much on intelligibility. Before starting the actual tests, the subjects were instructed to set δ for each noise type as per their preference. One key observation was, the preferred value of δ varied across subjects. This supports our claim that the proposed SE method and its developed application is user adaptive. We also conducted field test of our application in real world noisy conditions, which change dynamically. Subjective test results in Fig. 4 illustrate

the effectiveness of the proposed method in reducing the background noise and musical noise, simultaneously preserving the quality and intelligibility of the speech.

V. Conclusion

We developed a formant frequency based single microphone SE technique by introducing two gain functions depending on acoustical importance. The resulting gain allows smartphone user to suppress more noise on non-formant bands by retaining speech intelligibility and strike a balance between the amount of noise suppression and speech distortion. The proposed algorithm was implemented on a smartphone device, which works as an assistive device for HA. The objective and subjective results demonstrate the usability of the method in real world noisy conditions.

Acknowledgments

This work was supported by the National Institute of the Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) under the grant number 5R01DC015430-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1]. Klasen TJ, Bogaert den TV, Moonen M, Wouters J, “Binaural Noise Reduction algorithms for hearing aids that preserve interaural time delay cues,” IEEE Trans. Signal Process, vol. 55, pp. 1579–1585, 4 2007.

[2]. Kuo Y-T, Lin T-J, Chang W-H, Li Y-T, Liu C-W and Young S-T, “Complexity-effective auditory compensation for digital hearing aids,” IEEE Int. Symp on Circuits ad Systems (ISCAS), 5 2008.

[3]. Reddy CKA, Hao Y, Panahi I, “Two microphones spectral-coherence based speech enhancement for hearing aids using smartphone as an assistive device,” IEEE Int. Conf. on Eng. In Medicne and Biology soc., 10 2016.

[4]. https://support.apple.com/en-us/HT203990

[5]. Boll S, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoustic, Speech and Signal Process, vol. 27,pp. 113–120, 4 1979.

[6]. Ephraim Y and Malah D, “Speech enhancement using a minimum meansquare error short-time spectral amplitude estimator,” IEEE Trans.Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121,1984.

[7]. Ephraim Y and Malah D, “Speech enhancement using a minimum meansquare error log-spectral amplitude estimator,” IEEE Trans. Acoustics,Speech, and Signal Processing, vol. 33, no. 2, pp. 443–445, 1985.

[8]. Weninger F, Hershey JR, Roux JL, Schuller B, “Discriminatively trained recurrent neural networks for single-channel speech separation,” IEEE Global Conf. on Signal and Inf Processing, 12 2014.

[9]. Lim J, “Evaluation of a correlation subtraction method for enhancing speech degraded by additive noise,” IEEE Trans. Acoust., Speech.Signal Process, vol. ASSP-37, pp. 471–472, 12 1987.

[10]. Ning L, Loizou PC, “Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction,” J. Acoust. Soc. Amer, vol. 123(3), pp. 1673–1682, 2008. [PubMed: 18345855]

[11]. Loizou PC, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” IEEE Trans. Speech Audio Process, Vol.19, pp. 47–56, 2011.

[12]. Mustafa K and Bruce IC, "Robust formant tracking for continuous speech with speaker variability," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 435–444, 3 2006.

[13]. McLoughlin Ian Vince, and Chance RJ. "LSP-based speech modification for intelligibility enhancement." Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on Vol. 2 IEEE, 1997.

[14]. Tudor-Catalin Zorila, Kandia Varvara, and Stylianou Yannis. "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[15]. Sohn J, Kim NS, and Sung W, “A statistical model-based voice activity detection,” IEEE Signal Processing Letters., vol. 6, no. 1, pp. 1–3, 1999.

[16]. https://developer.apple.com/xcode/

[17]. https://developer.apple.com/library/content/documentation/MusicAudio/Conceptual/CoreAudioOverview/WhatisCoreAudio/WhatisCoreAudio.html

[18]. http://www.starkey.com/blog/2014/04/7-halo-features-that-will-enhanseevery-listening-experience

[19]. Rix AW, Beerends JG, Hollier MP, Hekstra AP, “Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs” IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), 2, pp. 749–752., 5 2001

[20]. ITU-T Rec. P.830, “Subjective performance assessment of telephoneband and wideband digital codecs,” 1996.

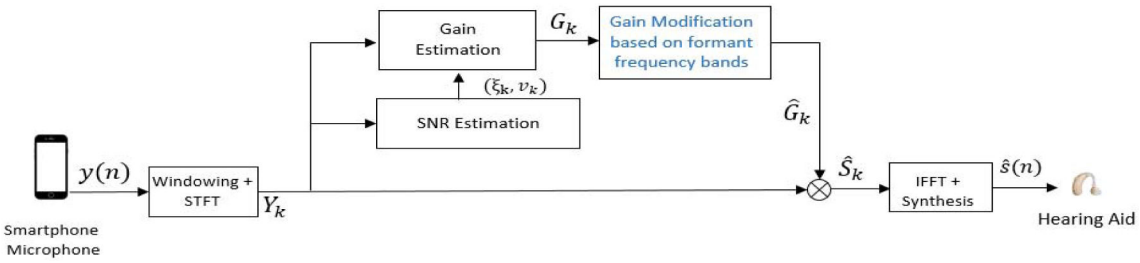

Fig. 1:
Block Diagram of Proposed SE method

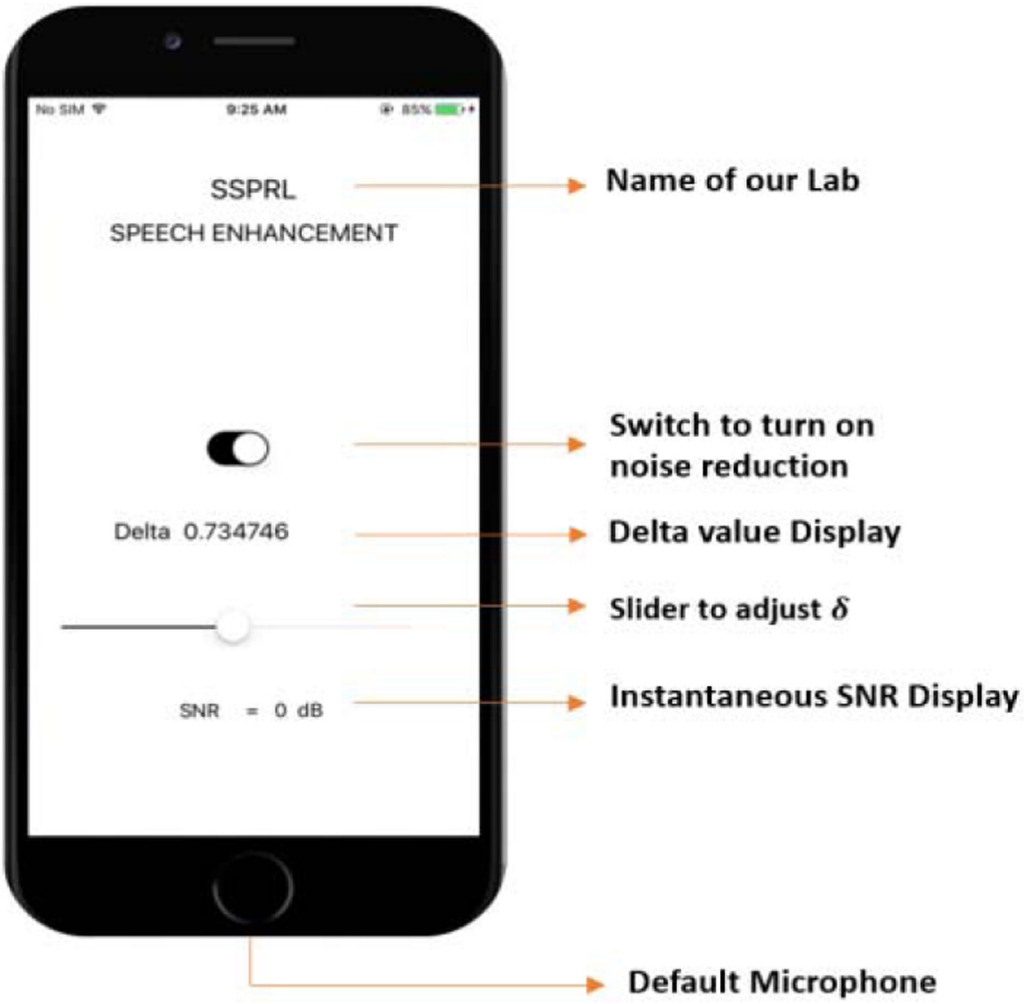

Fig 2:
Snapshot of developed SE method

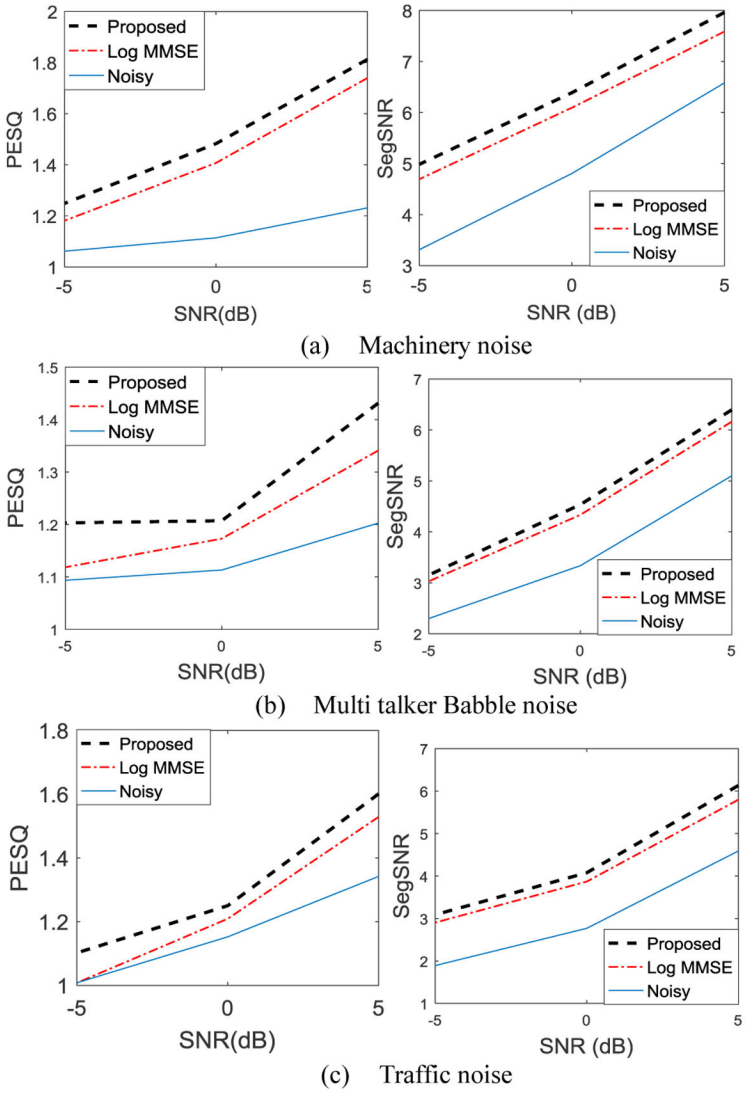

Fig.3.
Objective evaluation of speech quality and intelligibility

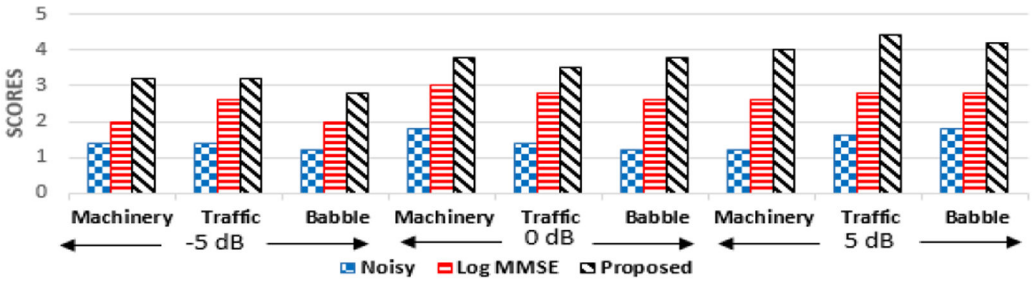

Fig.4.
Comparison of Subjective results