HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2020 July 30.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2018 July ; 2018: 5503–5506. doi:10.1109/EMBC.2018.8513674.

Smartphone based real-time super Gaussian single microphone Speech Enhancement to improve intelligibility for hearing aid users using formant information

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

Statistical Signal Processing Research Laboratory (SSPRL), The University of Texas at Dallas

Abstract

In this paper, we present a Speech Enhancement (SE) technique to improve intelligibility of speech perceived by Hearing Aid users using smartphone as an assistive device. We use the formant frequency information to improve the overall quality and intelligibility of the speech. The proposed SE method is based on new super Gaussian joint maximum *a Posteriori* (SGJMAP) estimator. Using the priori information of formant frequency locations, the derived gain function has *"tradeoff"* factors that allows the smartphone user to customize perceptual preference, by controlling the amount of noise suppression and speech distortion in real-time. The formant frequency information helps the hearing aid user to control the gains over the non-formant frequency band, allowing the HA users to attain more noise suppression while maintaining the speech intelligibility using a smartphone application. Objective intelligibility measures and subjective results reflect the usability of the developed SE application in noisy real world acoustic environment.

I. Introduction

Personal hearing devices such as hearing aid devices (HADs) are used by hearing impaired. In the United States about 2 percent of adults aged 45 to 54 have disabling hearing loss. The rate increases to 8.5% for adults aged 55 to 64. Nearly 25% of those aged 65 to 74 and 50% of those who are 75 and older have disabling hearing loss [1]. HAD and Cochlear implant (CI) are viable solutions for hearing loss. Speech Enhancement (SE) is an elemental block in HAD signal processing pipeline. Existing HADs have limited computing powers due to their size, processor, and power consumption. Therefore, these limitations make it impractical to implement complex signal processing algorithms on them in order to improve their performance. One practical solution is to use smartphone as an assistive tool for HADs as they have superior processing power and large population anyways possess smartphones. The microphone on the smartphone captures the noisy speech. The SE algorithm running on the processor of the smartphone reduces the background noise and the enhanced speech is wirelessly transmitted to the HADs. Recently, extensively used smartphones such as Apple iPhone have come up with new HA features such as Live Listen [2] to enhance the overall quality and intelligibility of the speech perceived by hearing impaired.

Literature offers extensive studies where SE algorithms are developed to improve the performance of HADs in the presence of background noises for the purpose of gaining better hearing capability by users of these devices. However, the prime challenge in single microphone SE is to suppress the background noise without inducing any sort of speech distortion. Traditional SE methods like spectral subtraction [3] and statistical model based methods proposed by Ephraim and Malah [4–5] can be implemented on a smartphone in real-time. But, these algorithms induce musical noise and do not substantially improve speech intelligibility. There are some computationally efficient alternatives for [4–5] which are proposed in [6–7]. But, none of these methods provide a control to suppress the amount of noise reduction in real-time. Recent development in SE algorithms [8] gives provision to control the amount of noise reduction in real time. Studies on speech intelligibility [8] show that maintaining speech intelligibility and also reducing the background noise is inadequate in many widely used algorithms. Studies in [9] have shown some good methods to improve the intelligibility but the performance deteriorates at lower SNR conditions. Formant frequency trajectories are important acoustical cues and researchers [10–11] have made use of this information to improve intelligibility.

In this work, The SE method makes use of formant information in the given speech to define acoustically significant frequency bands and the gain function of the new super Gaussian Joint Maximum *a Posteriori* (SGJMAP) based SE method explained in [8] is applied to suppress more noise in acoustically unimportant bands without inducing distortion in clean speech and residual noise. The *'tradeoff'* factors that are introduced in the proposed method allows us to control the gains over formant and non-formant locations in real time allowing the hearing-impaired smartphone user to attain more noise suppression without speech distortion and thereby maintaining the speech intelligibility. The proposed method is inexpensive and computationally efficient. Objective evaluations show good improvements in quality and intelligibility showing the effectiveness of the developed method. Subjective evaluations show the usefulness of the developed smartphone application in real-world noisy conditions.

II. Proposed Formant based Customizable SE gain

A. SGJMAP based Speech Enhancement

In the SGJMAP [7] method, a super-Gaussian speech model is used by considering non-Gaussianity property in spectral domain noise reduction framework. The speech spectral amplitude estimator using super Gaussian model allows the probability density function (PDF) of the distribution to be parameterized by μ and v. Let $x(n)$, $d(n)$ and $y(n)$ be clean speech, noise signal and noisy signal respectively. Considering that noise is additive and uncorrelated with the speech signal, the noisy speech is given by,

$$y(n) = x(n) + d(n) \tag{1}$$

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

$$Y_k(\lambda) = X_k(\lambda) + D_k(\lambda) \tag{2}$$

where, X and D are the clean speech and noise DFT coefficients respectively. In polar coordinates, (2) can be written as,

$$R_k(\lambda)e^{j\theta Y_k(\lambda)} = A_k(\lambda)e^{j\theta X_k(\lambda)} + B_k(\lambda)e^{j\theta D_k(\lambda)} \tag{3}$$

where, $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectrmns of noisy speech, clean speech and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{X_k}(\lambda)$. $\theta_{D_k}(\lambda)$ are the phase spectmms of noisy speech, clean speech and noise respectively. The goal of any SE technique is to estimate clean speech magnitude spectrum $A_k(\lambda)$ and its phase spectrum $\theta_{X_k}(\lambda)$. The magnitude and phase estimator of the SGJMAP jointly maximize the probability of magnitude and phase spectrum conditioned on the observed complex coefficient given by,

$$\hat{A}_k = \arg\max_{A_k} \frac{p\left(Y_k \middle| A_k, \theta_{X_k}\right) p\left(A_k, \theta_{X_k}\right)}{p(Y_k)} \tag{4}$$

$$\hat{\theta}_{X_k} = \arg\max_{\theta_{X_k}} \frac{p\left(Y_k \middle| A_k, \theta_{X_k}\right) p\left(A_k, \theta_{X_k}\right)}{p(Y_k)} \tag{5}$$

But due to the inaccuracies in the model [8] where it depends on parameters like μ, v and accuracy of the voice activity detector (VAD), it is very difficult to estimate the magnitude spectrum of the clean speech in the real world rapidly fluctuating acoustical enviromnent. In [8], these inaccuracies are compensated by introducing a tradeoff factor β into the cost function for estimating the optimal clean speech magnitude spectrum. Taking natural logarithm of (4), and differentiating with respect to A_k gives,

$$\begin{aligned} &\frac{d}{dA_k}\log\left(p\left(Y_k \middle| \beta A_k, \theta_{X_k}\right) p\left(\beta A_k, \theta_{X_k}\right)\right) \\ &= \frac{-\left(Y_k^* - A_k\beta e^{-j\theta X_k}\right)\left(-jA_k\beta e^{j\theta X_k}\right) + \left(Y_k - A_k\beta e^{j\theta X_k}\right)\left(jA_k\beta e^{-j\theta X_k}\right)}{\hat{\sigma}_{D_k}^2} \end{aligned} \tag{6}$$

Setting up (6) and simplifying the equation by solving the quadratic equation, the gain function of the new super-Gaussian amplitude estimator is given by [8],

$$G_k = \left[\left(\frac{1}{2\beta} - \frac{\mu}{4\sqrt{\hat{\gamma}_k\hat{\xi}_k}}\right) + \sqrt{\left(\frac{\mu}{4\sqrt{\hat{\gamma}_k\hat{\xi}_k}} - \frac{1}{2\beta}\right)^2 + \frac{v}{2\hat{\gamma}_k\beta^2}}\right] \tag{7}$$

where, $\hat{\xi}_k = \frac{\hat{\sigma}_{X_k}^2}{\hat{\sigma}_{D_k}^2}$ is the *a priori* SNR and $\hat{\gamma}_k = \frac{R_k^2}{\hat{\sigma}_{D_k}^2}$ is the *a posteriori* SNR. $\hat{\sigma}_{D_k}^2$ is estimated using a VAD. $\hat{\sigma}_{X_k}$ is the estimated instantaneous clean speech power spectral density. In [7]. $v = 0.126$ and $\mu = 1.74$ is shown to give better re suits. The optimal phase spectrum is the noisy phase itself $\hat{\theta}_{X_k} = \theta_{Y_k}$.

B. Formant frequency Band Estimation

In this work, the fonnant frequency bands are approximated by calculating the exact fonnants to improve the intelligibility of speech. The pitch and the fonnant 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]. We require a frequency range to approximate the presence of speech and apply considerably less noise suppression over that band and more noise suppression on the other bands. Therefore, the mean of fonnants $f_0 - f_3$ are calculated for large data sets and mean absolute enor for each fonnant is detennined over the data sets to find the frequency band of probable fonnant 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 fonnant. f_X represents mean fonnant frequency computed over entire database for $X = 0$, 1, 2 and 3 [10], f_a is the mean absolute enor detennined for each fonnant. f_X can be estimated in the real time for the noisy speech and f_a which is calculated over the large datasets can be used to find frequency band F_X in real time. Thus, we estimate four frequency bands (F_0 *to* F_3) from the respective mean fonnant locations. The FFT bins conesponding to the four frequency bands are thus calculated.

C. Gain function customization based on Frequency Bands

The block diagram of developed SE method is shown in Fig. 1. In traditional SE methods, the gain function shown in (7) is applied over the entire frequency range inducing speech distortion due to inaccuracies in gain function estimation. The proposed method allows more noise suppression on acoustically unimportant bands and far lowering noise suppression on significant fonnant frequency bands to retain clean speech intelligibility. Thus, we obtain two different gain functions based on the frequency bands. The gain function for the fonnant frequency bands is given by,

$$G_{k_F} = \left[\left(\frac{1}{2\beta_F} - \frac{\mu}{4\sqrt{\hat{\gamma}_k \hat{\xi}_k}}\right) + \sqrt{\left(\frac{\mu}{4\sqrt{\hat{\gamma}_k \hat{\xi}_k}} - \frac{1}{2\beta_F}\right)^2 + \frac{v}{2\hat{\gamma}_k \beta_F^2}}\right] \tag{9}$$

$$\hat{G}_k = \begin{cases} G_{k_F}, & if\ \mathrm{k} \in \mathrm{F}_X \text{ for X=0,1,2,3} \\ G_k, & otherwise \end{cases} \tag{10}$$

Where k represents the k^{th} frequency bin, F_X represents the bins associated with fonnant frequency bands, β and β_F allow the hearing-impaired smartphone user to obtain more noise suppression without speech distortion. The β_F can be kept constant or can be adjusted along with β by HAD user in real time based on his/her listening preference under continuously varying acoustical enviromnent. $\beta_F < \beta$ and β can be varied from 0.5 to 5.

We reconstruct the signal by considering the phase of the noisy speech signal. The final clean speech estimate is,

$$\widehat{X}_k = \widehat{G}_k Y_k \tag{11}$$

The time domain reconstruction signal $\hat{x}(n)$ is obtained by taking Inverse Fast Fourier Transfonn (IFFT) of $\widehat{X}_k$. As we consider band approximations, the inaccuracy in calculating the exact fonnant frequency does not affect the proposed method to a great extent. Near estimation of the fonnant frequency bands can improve the speech intelligibility substantially. Another advantage of the proposed method is it does not induce any musical noise, eliminating the need of any post filter after the enhancement. This reduces computational complexity and latency in real time.

III. SMARTPHONE IMPLEMENTATION TO FUNCTION AS AN ASSISTIVE DEVICE TO HA

Present work considers iPhone 7 as an assistive device to HA. The data is captured using the default mic on the smartphone at a sampling rate of 48 kHz. After acquiring the input, the 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. Switching the 'ON' button enables SE module to process the incoming audio stream by applying the proposed SE algorithm on the magnitude spectrum of noisy speech. The enhanced signal is then played back through the HAD. Once the noise suppression is on, we have provided parameters, which allow more noise suppression without inducing speech distortion and musical noise. Varying β will change the gains over non-formant regions and varying β_F will change the gains over fonnant regions. From our experiments, we have seen that by setting β high and β_F low, can achieve more noise suppression and maintain the speech integrity, as we do not distort the significant bands. The user can control these parameters by adjusting its values on the touch screen panel of the smartphone to attain more noise suppression based on their level of hearing comfort. For user's ease, we have also provided a button where he/she can keep β_F set automatically based on the value of β. However, β has to be adjusted manually as keeping the β constant is not feasible in real world noisy acoustic conditions. The processing time for the frame size of 10 ms (480 samples) is 1.4 ms. The smartphone application consumes very less power because of the computational efficiency of the developed algorithm. Through our experiments, we found that a fully charged smartphone can ran the application for 6.2 hours on iPhone 7 which has 1960 mAh battery. We use Starkey live listen [12] to stream the data from iPhone to the HAD. The audio streaming is encoded for Bluetooth Low Energy consmnption.

IV. Experimental results and discussion

A. Objective Evaluation

To objectively measure the performance of the proposed SE method, we use Perceptual Evaluation of Speech Quality (PESQ) [13] as it had a correlation with subjective tests than any other objective measures. The amount of noise reduction and the residual noise

reduction is generally measures using segmental SNR (SegSNR), so we also use this measure to evaluate the performance of our proposed method [14]. The proposed method is compared to Log-MMSE [5] and traditional SGJMAP [7] method to evaluate the performance. The fonnant frequency bands were calculated by detennining the mean of the fonnants and mean absolute enor for over 300 clean speech files from TIMIT database. The experimental evaluations are perfonned for 3 different noise types: machinery, multi-talker babble, and traffic noise. We considered these noise types as most of the enviromnental noises can be correlated to any one of the 3 noise types. The reported results are the average over 20 sentences from TIMIT database. For objective evaluation, noisy speech files sampled at 16 kHz, and 20 ms frames with 50% overlap were considered. All the noisy files have reverberation time of about 200 ms (RT 60). The β and β_F were adjusted empirically to give the best values for both PESQ and SegSNR and for each noise type. PESQ and SegSNR values show significant improvements over Log-MMSE and SGJMAP methods for all three noise types considered. Objective measures shown in Fig. 3 reemphasize the fact that the proposed method achieves comparatively more noise suppression without distorting speech by using fonnants and varying user adjustable parameters in real time.

B. Subjective Evaluations

Objective tests provide useful evaluation during the development phase of the proposed method. However, the practical usability of the application can be assessed by subjective tests. We perfonned mean opinion scores (MOS) [16] tests on 10 nonnal hearing adults including both male and female subjects. The subjects were presented with the noisy speech and enhanced speech using the Log-MMSE, SGJMAP and the proposed SE methods at the SNR levels of +5 dB, 0 dB and −5 dB for 3 different noise types. Subjects were asked to rate between 1 to 5 for each speech file based on how pleasant it was and how many words they could recognize. They were also allowed to go back and change the scores after listening to other speech files. Before starting the test, the subjects were instructed regarding the parameters β and β_F and were asked to set these parameters according to their listening preference. Different subjects chose to vary β and β_F for different types of noise. This test supported our claim that proposed SE method and its application is user adaptive and noise dependent. We also did field testing of our application where the acoustic enviromnent changed dynamically. Subjective evaluation is illustrated in Fig. 4.

V. Conclusion

We developed a single-channel SE which makes use of fonnant frequency information to improve the intelligibility of speech. We introduced two gain functions depending on acoustical significance. The resulting gain allows the hearing impaired smartphone user to suppress more noise on non-formant bands by retaining speech intelligibility and readjust the amount of noise suppression and speech distortion. The proposed method was implemented on a smartphone, which works as an assistive device for HA. The objective and subjective results validate 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 5R01DC015430-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors are with the Statistical Signal Processing Research Laboratory (SSPRL), Department of Electrical and Computer Engineering, The University of Texas at Dallas. imp015000@utd.edu.

References

[1]. Quick Statistics (n.d.) retrieved from http://www.nidcd.nih.gov/health/statisties/pages/-quiek.aspx

[2]. https://support.apple.com/en-us/HT20399Q

[3]. 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.

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

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

[6]. Wolfe PJ and Godsill SJ, “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, ” EURASIP Journal on Applied Signal Processing, vol. 2003, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia CommunicationsT.

[7]. Lotter P. Vary, “Speech Enhancement by MAP Spectral .Amplitude Estimation using a super-gaussian speech model,” EURASIP Journal on Applied Sig. Process, pp. 1110–1126, 2005.

[8]. Karadagur Ananda Reddy C, Shankar N, Shreedhar Bhat G, Charan R and Panahi I, “.An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device,” in IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1601–1605, Nov. 2017. [PubMed: 29353988]

[9]. 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]

[10]. Bhat GS, Shankar N, Reddy CKA and Panahi IMS, “Formant frequency-based speech enhancement technique to improve intelligibility for hearing aid users with smartphone as an assistive device,” 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), Bethesda, MD, 2017, pp. 32–35.

[11]. Zorila Tudor-Catalin, 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.

[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, March 2006.

[13]. https://develoner.annle.com/library/content/documentation/MusicAudio/Conceptual/CoreAudioOverview/WhatisCoreAudio/WhatisCoreAudio.html

[14]. Rix AW, Beerends JG, P Hollier M, 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., May 2001.

[15]. H Taal C, Hendricks RC, Heusdens R, Jensen R, “.An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE trans. Audio, Speech, Lang. Process. 19(7), pp. 2125–2136., Feb 2011.

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

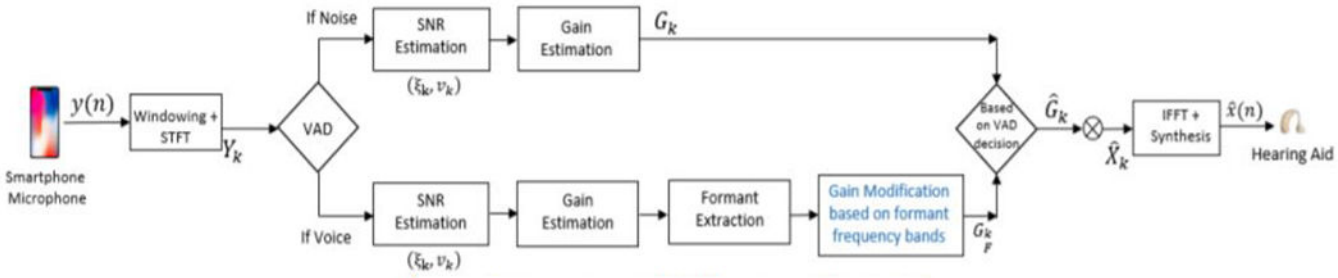

Figure 1:
Block Diagram of the proposed SE Method

Figure 2:
Snapshot of developed SE method

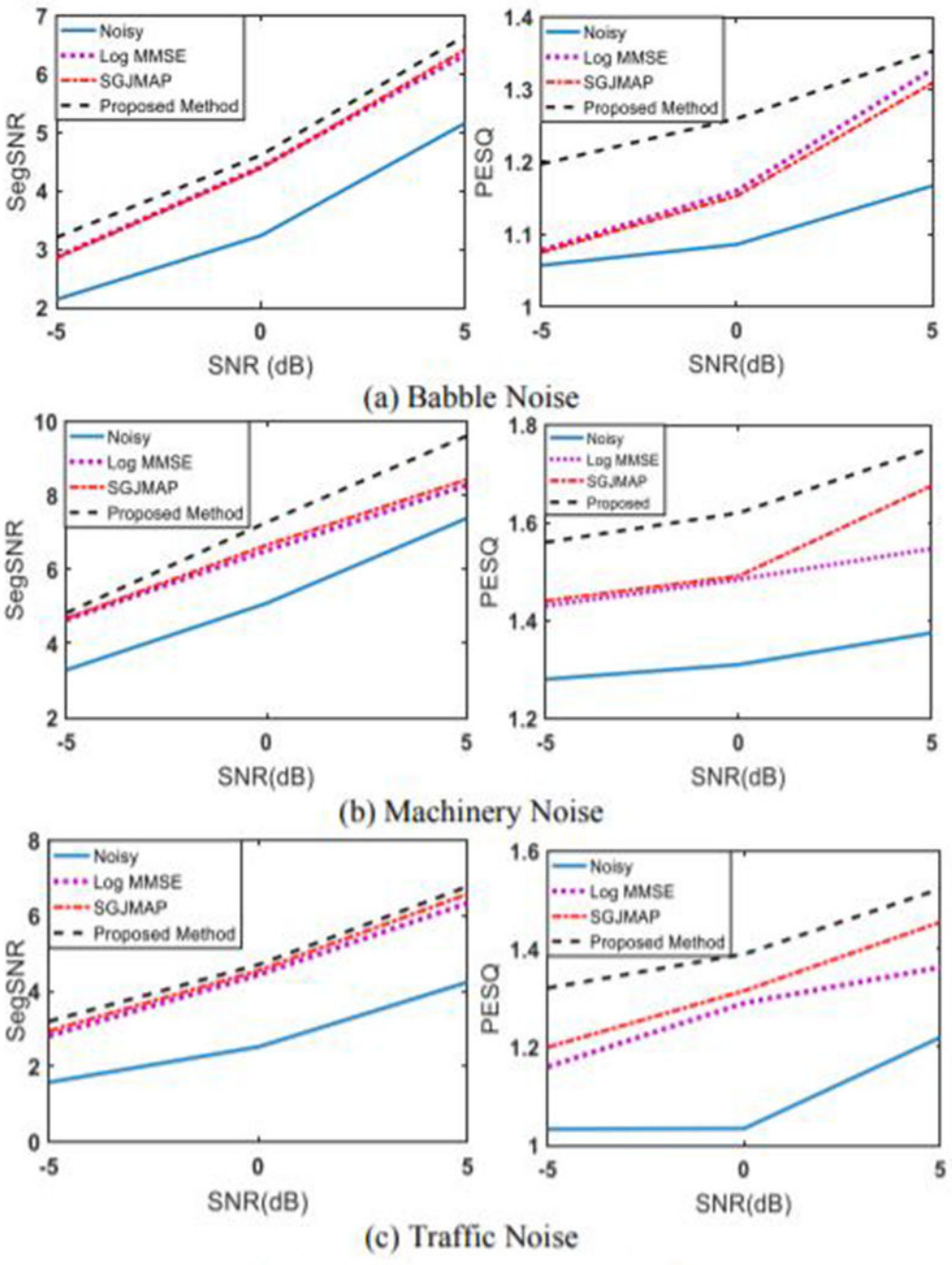

Figure 3:
Objective test results

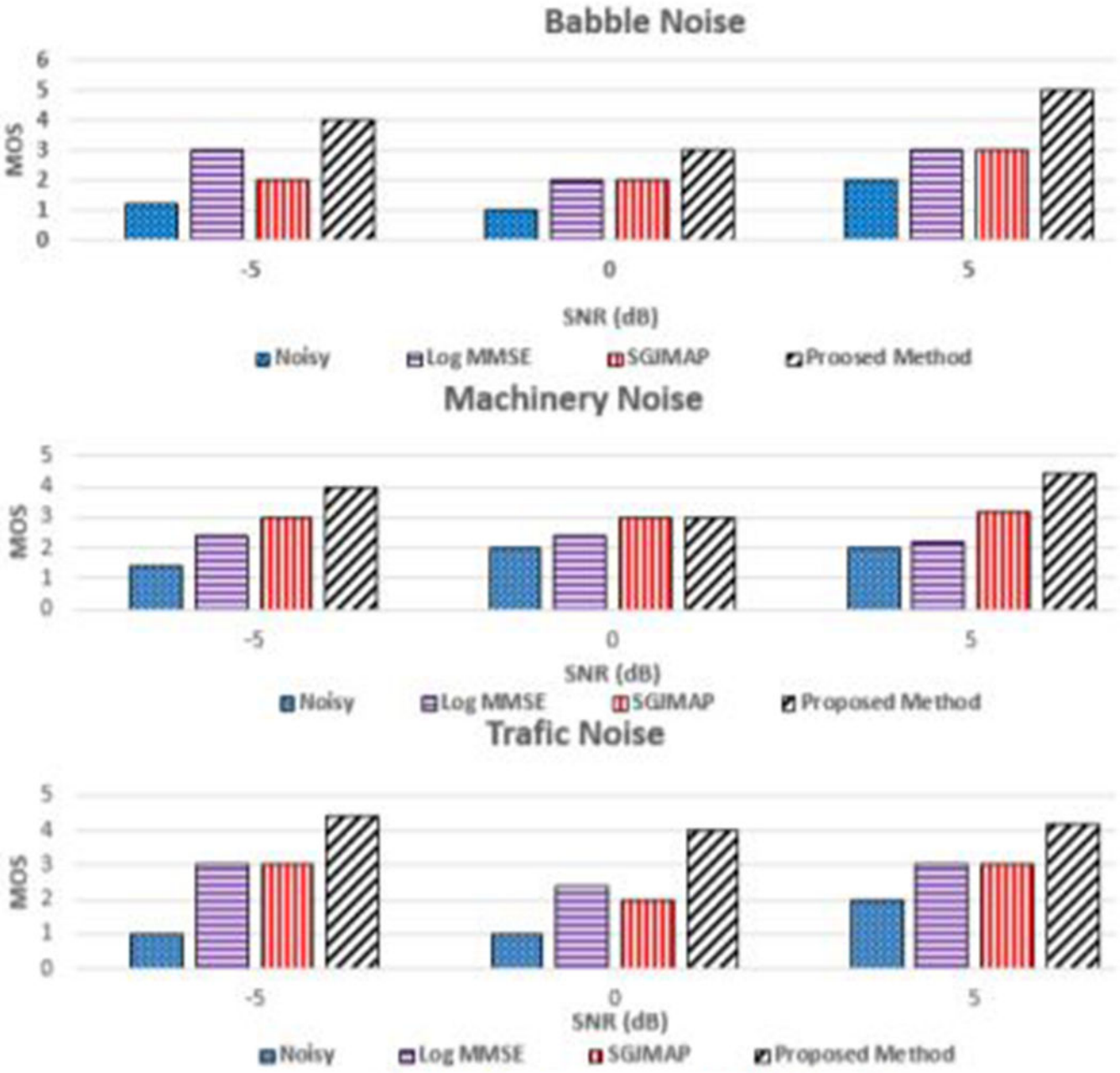

Figure 4:
Subjective test results