Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2020 August 03.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2018 July ; 2018: 417–420. doi:10.1109/EMBC.2018.8512369.

Influence of MVDR beamformer on a Speech Enhancement based Smartphone application for Hearing Aids

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

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

Abstract

This paper presents the minimum variance distortionless response (MVDR) beamformer combined with a Speech Enhancement (SE) gain function as a real-time application running on smartphones that work as an assistive device to Hearing Aids. It has been shown that beamforming techniques improve the Signal to Noise Ratio (SNR) in noisy conditions. In the proposed algorithm, MVDR beamformer is used as an SNR booster for the SE method. The proposed SE gain is based on the Log-Spectral Amplitude estimator to improve the speech quality in the presence of different background noises. Objective evaluation and intelligibility measures support the theoretical analysis and show significant improvements of the proposed method in comparison with existing methods. Subjective test results show the effectiveness of the application in real-world noisy conditions at SNR levels of −5 dB, 0 dB, and 5 dB.

Keywords

MVDR; Beamforming; Speech Enhancement; Hearing Aids; Smartphone

I. Introduction

Researchers have developed numerous solutions for hearing impaired in the form of Hearing Aid Devices (HADs) and other hearing assistive devices. Performance of HADs and Cochlear Implants (CI) degrade in the presence of background noise, thus reducing the quality and intelligibility of speech. Speech Enhancement (SE) plays a vital role in suppressing the noise in various stationary and non-stationary environments while preserving the speech features. Already existing HADs do not have the power to handle computationally complex signal processing algorithms [1]–[4]. Lately, HADs manufacturers are using auxiliary devices as an external microphone in the form of a pen or a necklace to capture speech in real-time and transmit it to the HADs through wired or wireless connection [5]. The drawbacks of these existing devices are that they are expensive and not portable. One solution is to use a smartphone as an assistive device to HADs by capturing noisy signal using two microphones and performing complex computations. Smartphones also solve the problem of cost and portability.

imp015000@utd.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In recent times, Apple iPhone and Android smartphones are coming up with sophisticated HA features such as Live Listen by Apple [6], and many third-party applications are being developed to enhance the quality and intelligibility of the speech perceived by the hearing-impaired people. In many HA applications on the smartphone, a single microphone is used to avoid audio input/ output latency. The proposed algorithm uses two microphones on the smartphone to suppress the background noise without distorting the clean speech.

Numerous studies have shown that SE is a key module in the HAD signal processing pipeline and would improve the listening comfort for the hearing impaired. Existing SE methods like Spectral Subtraction [7] and statistical model based methods proposed by Ephraim and Malah [8–9] can be implemented on a smartphone, but they do not improve the speech quality and intelligibility to a satisfactory extent. An ideal binary mask in SE can improve intelligibility [10], but precise estimation of the binary mask is challenging, especially in lower SNR conditions.

Over the last few decades, researchers have developed beamforming algorithms, which can be classified into fixed and adaptive beamformers. Fixed beamformers have static filter coefficients and signal independent spatial response [11]. The isotropic model, which is a first-order approximation of real noise fields, is commonly used in these beamformers and the noise field is not known. This limitation makes the fixed beamformer less efficient for real applications. As a solution, the beamforming filter coefficients can be changed, leading to the second class of adaptive beamformers. Among several SE techniques, adaptive beamformers are commonly used to improve the performance of the algorithm. The MVDR beamformer has wide range applications for extraction of desired speech signals in noisy environments [12]. The MVDR beamformer, known as Capon beamformer [13], dating back to 1980s, minimizes the output power of the beamformer under a single linear constraint on the response of the array towards the preferred signal. This spatial filtering process plays a critical role in extracting the signal of interest, suppressing ambient noise, and separating multiple sound sources. MVDR beamformer requires less a priori knowledge, which makes it practical for implementing it as a smartphone-based SE application for HADs.

The proposed algorithm is a combination of MVDR beamformer and Minimum mean square error Log spectral amplitude estimator (Log-MMSE) SE gain function, for suppressing noise and extracting the desired speech. This method is computationally efficient and helps in achieving minimal speech distortion for the hearing-impaired. Performance of the proposed method is compared against standard techniques of SE for speech quality and intelligibility. Subjective evaluations show promising results of the real-time application.

II. Proposed SE Gain Function

In the smartphone application, signals captured by the two microphones is composed of both clean speech and the background noise. Figure 1 shows the block diagram of the proposed method implemented on the smartphone. We consider a signal model with the first microphone as the reference point, the signal received by the n^{th} microphone ($n = 1, 2$) can be written as,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$y_n(t) = s_n(t) + w_n(t) = s(t - \tau_n) + w_n(t) \tag{1}$$

where $y_n(t)$, $s_n(t)$ and $w_n(t)$ are noisy speech, clean speech and noise signals respectively picked up by the n^{th} microphone at time t. Let τ_0 be the relative time delay between the two microphones given by $\tau_0 = \delta/c$ with δ as the spacing between the two microphones and c being the speed of sound in air. The signals are considered to be zero mean and real, noise signal $w_n(t)$ are assumed to be uncorrelated with $s_n(t)$.

For efficient performance, the signals are transformed to frequency domain and are re-written as,

$$\begin{aligned} Y_n(\omega) &= S_n(\omega) + W_n(\omega) \\ &= e^{-j(n-1)\omega\tau_0\cos(\theta_d)} S_1(\omega) + W_n(\omega) \end{aligned} \tag{2}$$

where $Y_n(\omega)$, $S_n(\omega)$, $W_n(\omega)$ are the Fourier transforms of $y_n(t)$, $s_n(t)$, $w_n(t)$ respectively. $\omega = 2\pi f$ is the angular frequency. Equation (2) can be rearranged into vector form as follows,

$$\begin{aligned} Y(\omega) &\triangleq [Y_1(\omega)\ Y_2(\omega)]^T \\ &= d_{\theta_d}(\omega) S_1(\omega) + W(\omega) \end{aligned} \tag{3}$$

where the superscript ^{T} is the transpose operator,

$$d_{\theta_d}(\omega) \triangleq \left[1\ e^{-j\omega\tau_0\cos(\theta_d)}\right]^T \tag{4}$$

is the steering vector and the noisy signal $W(\omega)$, is defined similar to $Y(\omega)$. θ_d is the angle of incidence of the source at the plane of microphones. Since the signals are assumed to be uncorrelated, the correlation matrix of $Y(\omega)$ can be determined by the method explained in [11].

A. MVDR Beamformer

The goal of beamforming is to extract the desired speech signal $S_1(\omega)$, by applying a linear filter $h(\omega)$ to $Y(\omega)$. This can be shown as follows,

$$Z(\omega) = h^H(\omega) d_{\theta_d}(\omega) S_1(\omega) + h^H(\omega) W(\omega) \tag{5}$$

where $Z(\omega)$ is the output of the beamformer, $h^H(\omega) S_1(\omega)$ is the filtered speech signal, and $h^H(\omega) W(\omega)$ is the residual noise.

The MVDR beamformer output can be obtained by minimizing the variance on either side of (5), or the residual noise with the constraint that the signal from the desired direction is without any distortion. In this work, we consider the minimization of variance of the residual noise.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\min_{h(\omega)} E[|h^H(\omega)W(\omega)|^2] \text{ subject to } h^H(\omega)d_{\theta_d}(\omega) = 1 \tag{6}$$

E[.] denotes mathematical expectation. Using a Lagrange multiplier to adjoin the constraint to the objective function, then differentiating with respect to $h(\omega)$, and equating the result to zero, (6) can be reduced to,

$$h(\omega) = \frac{\Gamma_W^{-1}(\omega)d_{\theta_d}(\omega)}{d_{\theta_d}^H(\omega)\,\Gamma_W^{-1}(\omega)d_{\theta_d}(\omega)} \tag{7}$$

where $\Gamma_w(\omega) = \Phi_w(\omega)/\phi_{w1}(\omega)$ is the pseudo-coherence matrix of the noise with $\Phi_w(\omega) = E[W(\omega)W^H(\omega)]$ and $\phi_{w1}(\omega) = E[|W_1(\omega)|^2]$.

B. Log-Spectral Amplitude Estimator

In the Log-MMSE method, speech and noise models are considered to be statistically independent Gaussian Random Variables [14]. The aim is to minimize the mean squared error of log magnitude spectra between estimated and true speech. The input is taken to be the output of the MVDR beamformer $z(n)$, which contains filtered speech signal $s'(n)$, and some residual noise $w'(n)$,

$$z(n) = s'(n) + w'(n) \tag{8}$$

The noisy kth Discrete Fourier Transform (DFT) coefficient of y(n) for frame λ is given by,

$$Z_k(\lambda) = S'_k(\lambda) + W'_k(\lambda) \tag{9}$$

Where S′ and W' are the input speech and noise DFT coefficients. In polar coordinates, (9) can be written as,

$$R_k(\lambda)e^{j\theta_{z_k}(\lambda)} = A_k(\lambda)e^{j\theta_{s'_k}(\lambda)} + B_k(\lambda)e^{j\theta_{w'_k}(\lambda)} \tag{10}$$

Where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectra of noisy speech, input signal and noise respectively. $\theta_{z_k}(\lambda)$, $\theta_{s'_k}(\lambda)$, $\theta_{w'_k}(\lambda)$ are the phase spectra of noisy, input speech and noise respectively. Looking at the estimator $\hat{A}_k$, which minimizes the distortion measure as explained in [8], the mean-square error of the log-magnitude spectra is given by,

$$E\left\{\left(log\ A_k - log\ \hat{A}_k\right)^2\right\} \tag{11}$$

Where, A_k is the kth bin of magnitude spectrum, and $\hat{A}_k$ is the kth 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,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$log\ \hat{A}_k = E\{log\ A_k\ |Z_k(\lambda)\} \tag{12}$$

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

$$\hat{A}_k = exp\ (E\{log\ A_k|\ Z_k(\lambda)\}) \tag{13}$$

Solving the above expectation, the final estimate of speech magnitude spectrum according to [8] is given by,

$$\begin{aligned}\hat{A}_k &= \frac{\xi_k}{\xi_k + 1}\exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k \\ &\triangleq G_k R_k\end{aligned} \tag{14}$$

Where $\nu_k = \frac{\xi_k}{1+\xi}\gamma_k$ here $\xi_k = \frac{\sigma^2_{s'k}}{\sigma^2_{\bar{w}_k}}$ is the *a priori* SNR and $\gamma_k = \frac{R_k^2}{\sigma^2_{\bar{w}_k}}$ is the *a posteriori* SNR. $\sigma^2_{\bar{W}_k}$ is estimated using a voice activity detector (VAD) [15]. $\sigma_{s'k}$ is the estimated instantaneous clean speech power spectral density. The optimal phase spectrum is the noisy phase spectrum itself $\theta_{S_k} = \theta_{Y_k}$. The final clean speech estimate is,

$$\hat{S}'_k = G_k Z_k \tag{15}$$

The time domain reconstruction signal $\hat{s}'(n)$ is obtained by taking inverse Fourier Transform of $\hat{S}'_k$.

III. Real-Time Implementation on Smartphone as an Assistive Device to HADs

In this paper, Google Pixel running Android 7.1 Nougat operating system is considered as an assistive device. Two microphones (13 cm apart) on the smartphone capture the audio signal, process the signal and transmit the enhanced signal to the HADs. The smartphone device considered has an M4/T4 HA Compatibility rating and meets the requirements set by Federal Communications Commission (FCC). Android Studio [16] is used for implementation of the SE algorithm on the smartphone. An inbuilt android audio framework was used to carry out dual microphone input/output handling. The input data is acquired at 48 KHz sampling rate and a 10ms frame, with FFT size to be 512 is considered as the input buffer. Figure 2 shows the screenshot of the proposed SE method implemented on Pixel smartphone. When the button is in “ON” mode, the microphone will record the audio signal and playback to the HADs without any processing. There is another button present on the screen to apply the developed SE algorithm to enhance the audio stream. The enhanced output signal is then played back to the HADs. Initially, when the SE algorithm is turned on, the algorithm uses approximately 3 seconds to estimate the noise variance. Hence, we assume there is no speech activity during this time. The smartphone application is computationally efficient and consumes less power.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

IV. Experimental Results

A. Objective Evaluation

The performance of the proposed method is evaluated by comparing with dual microphone coherence [17] and Log-MMSE [9] methods, promising results are seen. The objective evaluations are performed for 3 different noise types: machinery, multi-talker babble, and traffic noise.

The plotted results are the average over different speech signals from the HINT database. The audio files are sampled at 16 kHz, and 10 ms frames with 50% overlap are considered. Perceptual evaluation of speech quality (PESQ) [18] and short time objective intelligibility (STOI) [19] are used to measure the quality and intelligibility of the speech respectively. PESQ ranges between 0.5 and 4.5, with 4.5 being high perceptual quality. Higher the score of STOI better the speech intelligibility. Figure 3 shows the plots of PESQ and STOI versus 3 different SNR for the 3 noise types. PESQ and STOI values show substantial improvements over other methods for all three noise types considered. Objective and Intelligibility measures state the fact that the proposed SE method suppresses more noise with minimal speech distorting.

B. Subjective test setup and results

Subjective measures give information about the practical usability of our application in real-time. Thus, Mean Opinion Score (MOS) tests [20] was performed on 10 normal hearing subjects including 5 male and 5 female adults. They were presented with noisy speech and enhanced speech using the proposed, coherence and Log-MMSE methods at different SNR levels of −5 dB, 0 dB, and 5 dB. The audio files were played on headphones for the subjects. Each subject was instructed to rate between 1 and 5 for each audio file based on the following criteria: 5 being excellent speech quality and imperceptible level of distortion. 1 having the least quality of speech and intolerable level of distortion. This test provided a good comparison between the proposed method and other existing methods. Subjective test results in Fig. 4 illustrate the effectiveness of the proposed method in reducing the background noise, simultaneously preserving the quality and intelligibility of the speech.

V. Conclusion

An MVDR beamformer based dual microphone SE algorithm was developed and implemented on a smartphone as a real-time application. This method can act as an assistive device for HADs. Objective and Subjective evaluations verify that the proposed method can be used as a solution to enhance the speech in real-world noisy enviromnents.

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

References

[1]. 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." in Proc. IEEE Int. Symp. Circuits Syst, 5 2008, pp. 1472–1475.

[2]. Karadagur Ananda Reddy C, Shankar N, Shreedhar Bhat G, Charan R and Panalii 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]

[3]. Klasen TJ, Bogaert den TV, Moonen M, and Wouters J, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," IEEE Trans. Signal Process, vol. 55, no. 4, pp. 1579–1585. Apr. 2007.

[4]. Reddy CKA, Hao Y, and Panahi I, "Two microphones spectral coherence based speech enhancement for hearing aids using smartphone as an assistive device," in Proc. IEEE Int. Conf. Eng. Med. Biol. Soc, Oct. 2016, pp. 3670–3673.

[5]. Edwards B, "The future of hearing aid technology," J. List, Trends Amplif, vol. 11, no. 1, pp. 31–45, Mar. 2007.

[6]. Dec. 2017 [Online]. Available: https://support.apple.com/en-us/HT203990

[7]. Boll S, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech Signal Process, vol. ASSP-27, no. 2, pp. 113–120. Apr. 1979.

[8]. Ephraim Y and Malah D, "Speech enhancement using a minimum meansquare error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process, vol. ASSP-32, no. 6, pp. 1109–1121. Dec. 1984.

[9]. Ephraim Y and Malah D, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech. Signal Process, vol. ASSP-33, no. 2, pp. 443–445. Apr. 1985.

[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]. Pan C, Chen J and Benesty J, "Performance Study of the MVDR Beamformer as a Function of the Source Incidence Angle," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[12]. Pan C, Chen J and Benesty J, "On the noisereduction performance of the MVDR beamformer innoisy and reverberant environments," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 815–819.

[13]. Capon J, "High resolution frequency-wavenumber spectrum analysis," Proc. IEEE, vol. 57, pp. 1408—14i8, Aug. 1969.

[14]. 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.

[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]. Dec. 2017 [Online]. Available: https://developer.android.com/studio/intro/index.html

[17]. Yousefian N, Kokkinakis K and Loizou PC, "A coherence-based algorithm for noise reduction in dual-microphone applications," 2010 18th European Signal Processing Conference, Aalborg, 2010, pp. 1904–1908.

[18]. Rix AW, Beerends JG, Hollier MP, and Hekstra AP, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in Proc. IEEE Int. Conf. Acoust., Speech. Signal Process, 5 2001, vol. 2, pp. 749–752.

[19]. Taal CH, Hendricks RC, Heusdens R, and Jensen R, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process vol. 19, no. 7, pp. 2125–2136. Feb. 2011.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

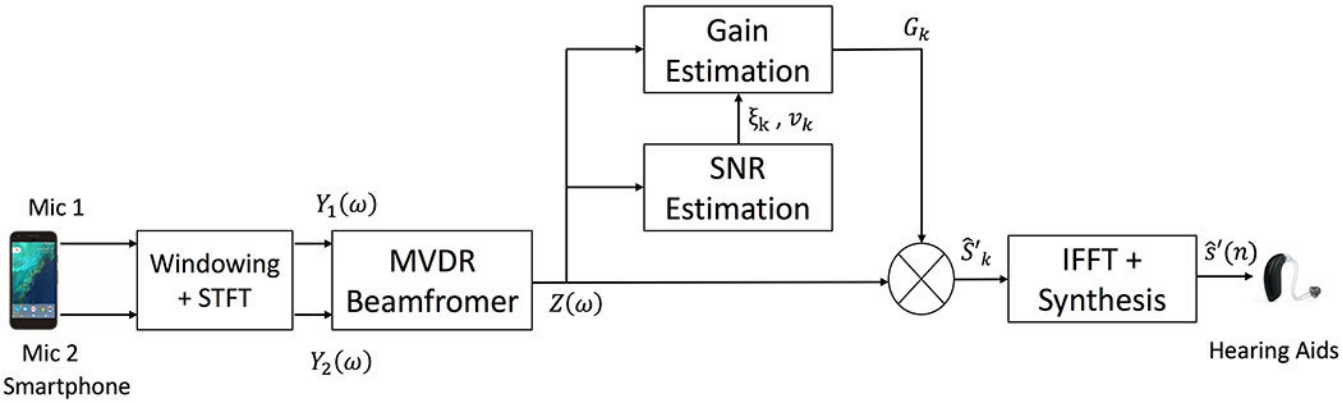

Fig. 1:
Block Diagram of Proposed SE method

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

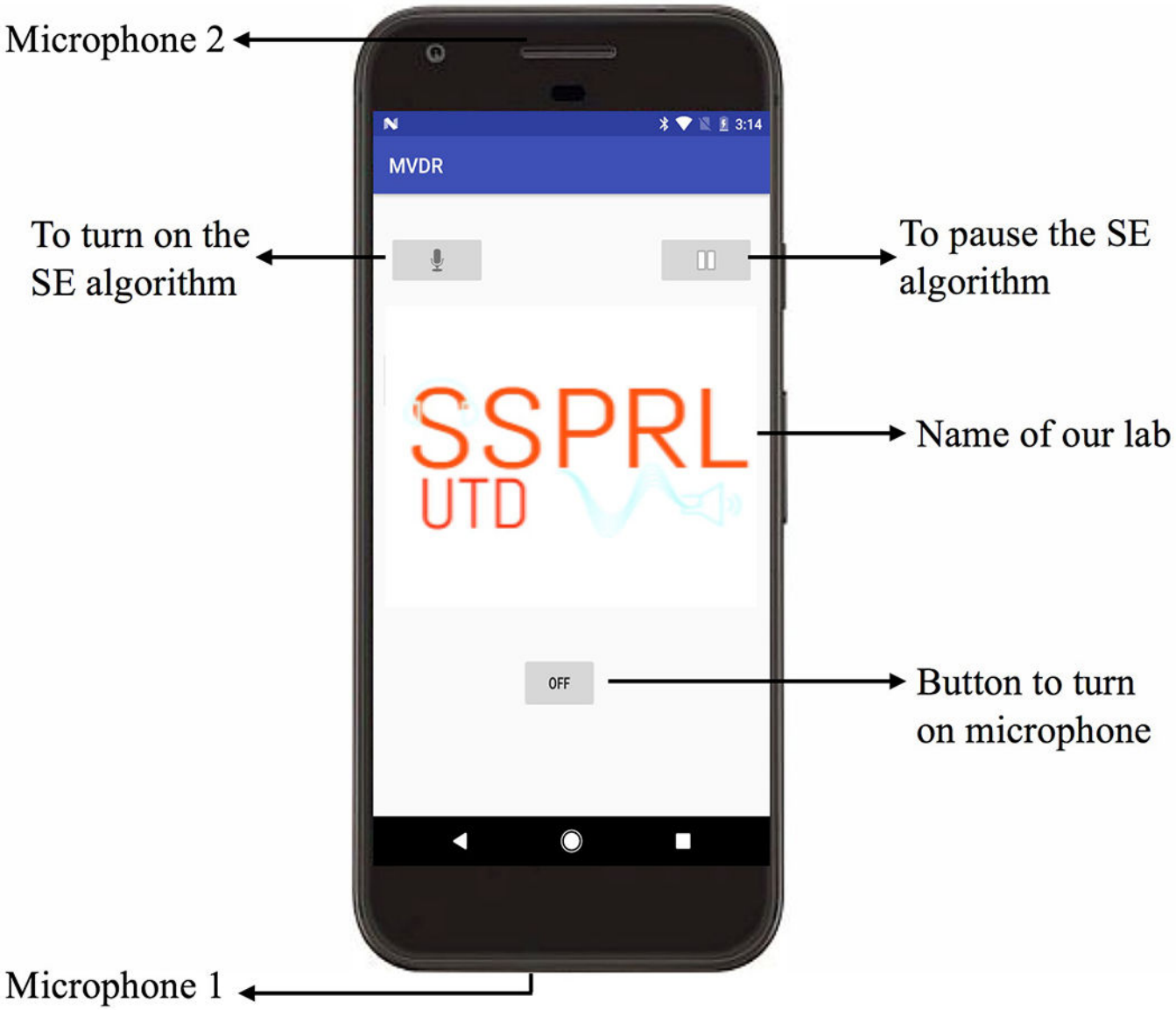

Fig 2:
Screenshot of developed SE method

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

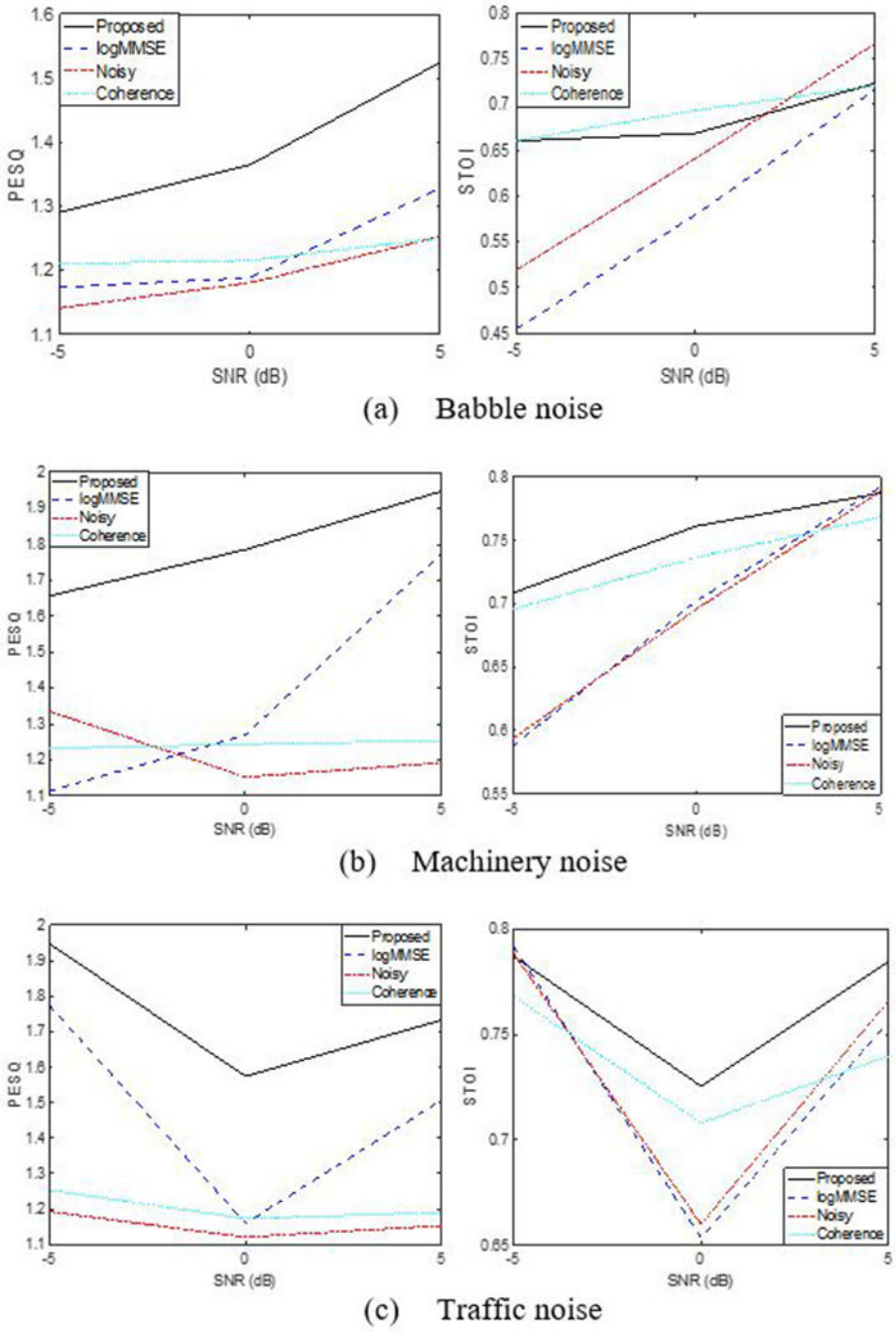

Fig.3.
Objective evaluation of speech quality and intelligibility

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

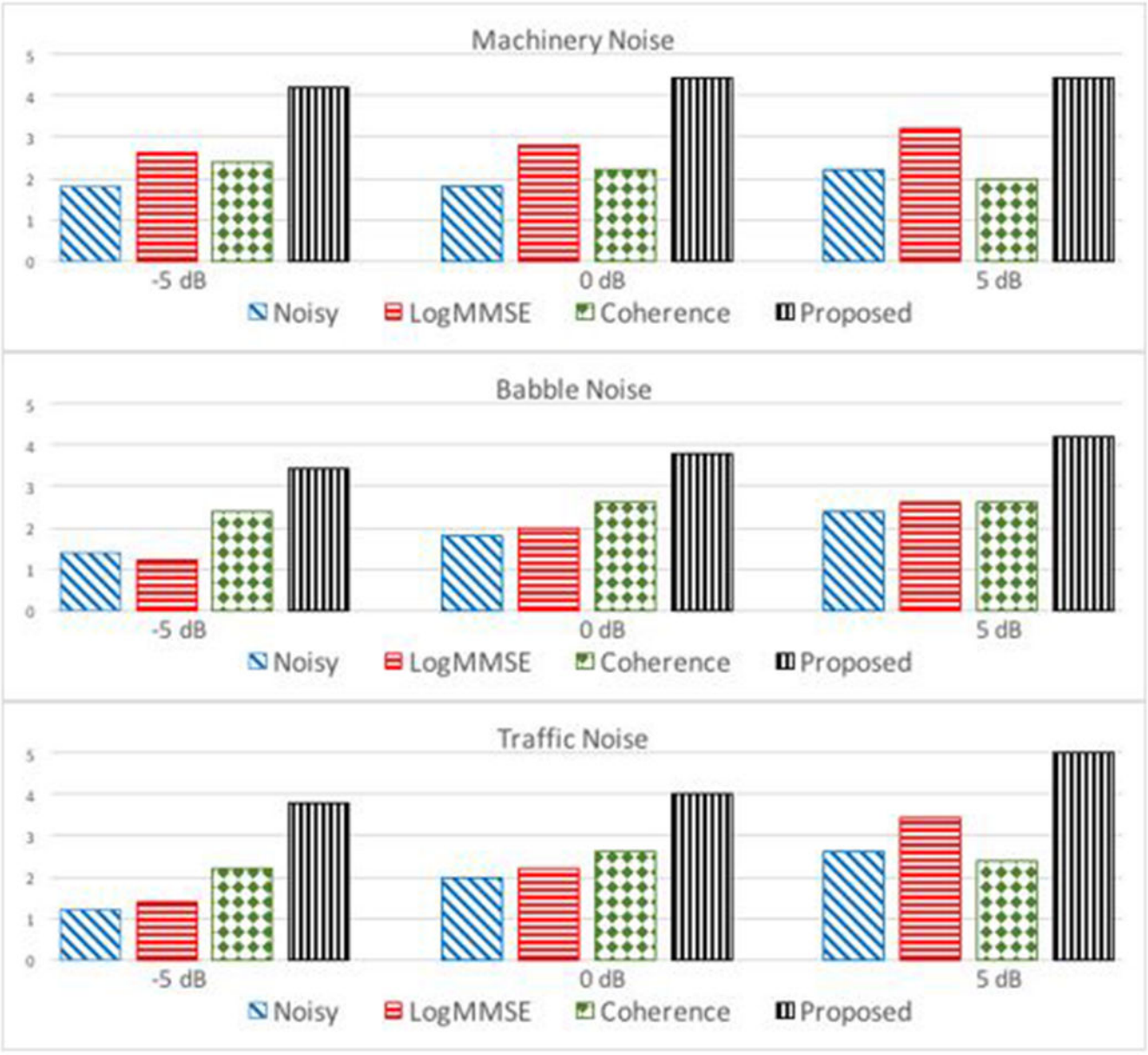

Fig.4.
Comparison of Subjective results