Smartphone-based Real-time Speech Enhancement for Improving Hearing Aids Speech Perception

Yu Rao, *IEEE Student Member,* Yiya Hao, Issa M.S. Panahi, *IEEE Senior Member,* Nasser Kehtarnavaz*, IEEE Fellow*

***Abstract*—In this paper, the development of a speech processing pipeline on smartphones for hearing aid devices (HADs) is presented. This pipeline is used for noise suppression and speech enhancement (SE) to improve speech quality and intelligibility. The proposed method is implemented to run in real-time on Android smartphones. The results of the testing conducted indicate that the proposed method suppresses the noise and improves the perceptual quality of speech in terms of three objective measures of perceptual evaluation of speech quality (PESQ), noise attenuation level (NAL), and the coherent speech intelligibility index (CSII).**

I. Introduction

Personal hearing devices such as hearing aid devices (HADs) are wearable devices that are widely being used, especially by the hearing impaired people. According to [1], approximately 15% of American adults aged 18 and over report some trouble in hearing. One in eight people in the United States aged 12 years or older has hearing loss in both ears. 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 [2]. Fortunately, more than 90% of individuals with hearing loss can be helped with hearing instruments [3]. 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. However, such limitations do not exist for the widely used and computationally powerful smartphones. Currently, over 4 billion people worldwide use smartphones. Among them are the people with hearing loss who can use smartphones as an assistive device in two possible ways: as an standalone listening device just similar to a HAD, or as an additional device in conjunction with a HAD.

In this paper, we present an adaptive SE method which runs on Android smartphones in real-time and improves speech quality and intelligibility for hearing aid users. We exploit the benefits of a dual microphone SE technique as the first processing stage. A modified single microphone SE algorithm is then applied as the second processing stage. A tuning factor is introduced in the estimation of a-priori signal-to-noise ratio (SNR) which has an impact on improving the quality and intelligibility of enhanced speech. The proposed SE method is implemented on an Android platform running in real-time with a graphical users interface (GUI) allowing users to adjust the enhancement for different types of background noise.

The rest of the paper is organized as follows. Section II provides an overview of the SE speech processing method implemented on the smartphone platform. Section III covers the real-time implementation aspect of the proposed method on Android smartphones. Section IV covers the results obtained by experimental tests in terms of three objective measures and their discussion. Conclusion is in Section VI.

II. Smartphone-Based Speech Enhancement

As an assistive device for HADs, a smartphone can capture and process audio signals and transfer them to a HAD. In this section, we describe the developed two-stage speech processing pipeline that is implemented on the smartphone platform. In the first stage, we use the two microphones of a smartphone and a block-based normalized least mean square error algorithm to enhance the speech quality. In the second stage, we modify and use the minimum mean square error-Log Scale Amplitude (MMSE-LSA) estimator [4]. To track the noise power, a look-up table technique [5] is applied and a tuning factor is introduced for the a-priori SNR estimation. The algorithms are then optimized to run in real-time on the smartphone platform.

A. Two-Microphone pre-processing - Stage one

The conventional normalized least mean square (NLMS) error algorithm is known for its robustness and stable performance. Generally, it is used for active noise control and feedback cancellation. Nevertheless its computational complexity is prohibitive in practice due to its sample based processing which is not suitable for mobile applications. Several block-based LMS algorithms have been proposed in [6, 7, 17] which demonstrate the possibility of block-based LMS. In [8], a block normalized least mean square (BNLMS) algorithm has been introduced. The BNLMS is thus used here to serve as a pre-processing for our smartphone-based speech enhancement method without affecting the noise suppression part. This preprocessing operation filters the input noisy speech signal received through the two microphones with less feedback effect and provides an output signal of higher SNR

*Research reported in this publication was supported by NIDCD Institute of the National Institutes of Health under award number R56DC014020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Authors are with the Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas 75080, USA.

for the second stage of our single input SE algorithm. The modified MMSE-LSA is mentioned next.

B. *Single microphone speech enhancement - Stage two*

After the first pre-processing stage, a single microphone SE technique is used. An MMSE-LSA estimator is used and a look-up table technique is applied for noise power estimation. MMSE-LSA can be stated as:

$$\hat{X}(k,l) = G(\tilde{\xi}(k,l), \tilde{\gamma}(k,l))Y(k,l) \qquad (1)$$

where $\hat{X}$ is the estimated speech magnitude spectrum, Y is the received noisy speech magnitude spectrum, k and l represent the frequency bin index and frame index, respectively. $\tilde{\xi}$ and $\tilde{\gamma}$ are known as the a-priori and a-posteriori SNRs. $G(.)$ is the non-linear gain function. To get a better estimate of the a-priori SNR, we introduce a tuning factor ρ $(0 < \rho < \infty)$ to the cost function discussed in [4, 16] as follows:

$$\xi(k,l) = \alpha \frac{|\hat{X}(k,l-1)|^2}{\rho|\tilde{N}(k,l-1)|^2} + (1-\alpha)\max[\frac{|Y(k,l)|^2}{\rho|\tilde{N}(k,l)|^2} - 1, 0] \qquad (2)$$

The tuning factor controls noise distortion in speech and can be found empirically. By increasing ρ, the noise power is overestimated. Thus less residual noise will remain in the enhanced speech. As ρ is decreased, the noise power is underestimated. Therefore, a large portion of noise will remain in the enhanced speech. When $\rho = 1$, then (2) will become the same as the well-known conventional decision-directed method [4].

III. Real-Time Implementation on Smartphone

For real-time smartphone implementation, an Android based smartphone (Nexus 6) was used together with the Android application programming interface (API) 15 [13]. The Native Development Kit (NDK) [14] was used for code optimization. Sampling rate was set to 16kHz. The two microphones on the back of the smartphone were used. The one close to the camera was regarded as Mic 1, the one which was near the micro USB port on the smartphone was regarded as Mic 2. The processed signal from the smartphone was transmitted to an audio output device. Fig.1 shows the real-time framework of the developed approach on the smartphone. The Mic 1 is assumed to be closer to the speech source and Mic 2 to be closer to the noise source. The shift size of the input buffer is *N*. *l* is assumed to be the current frame index, and *n* is sample index. Buffer 1 and Buffer 2 are used for storing the input audio data from Mic 1 and Mic 2 for the frames *l*, *l*-1 and also last *J* samples in frame *l-2*. Each buffer has a length of *2N+J*, where *J* denotes the length of the BNLMS error adaptive filter $\bar{\boldsymbol{w}}$. $y[n]$ denotes the output of the adaptive filter. By overlapping and adding the current filtered output with the previous filtered output by 50%, the output $\tilde{y}[n]$ for the frame *l-2* is obtained. Note that the $\tilde{y}[n]$ is an estimate (in the mean square sense) of the additive noise, uncorrelated with speech, in the received noisy speech signal via Mic 1. $e[n]$ is the error between the filter output $\tilde{y}[n]$ and Mic 1 data samples. As shown in Fig. 3, $e[n]$ and $x_2[n]$ are used to obtain the coefficients of the adaptive filter $\bar{\boldsymbol{w}}$. Also, note that $e[n]$ is a noisy speech signal with SNR higher than that received via Mic 1. The signal $e[n]$, as a single noisy speech signal, goes into the second stage of the single microphone SE (SMSE) algorithm. In this stage, error frames *l-2* and *l-3* are concatenated and treated by the analysis window and fast Fourier transform (FFT). The SMSE gains are applied (which were computed by the modified SNR estimation) to the resulting frame. To get the output of the frame *l-2*, inverse FFT and synthesis window are applied to the data. The synthesis window is used for recovering the analysis window effect, which is computed as the inverse of the overlapping analysis window.

In the implementation carried out which are reported next, the following parameters were used; $J = 64$, $\mu = 0.05$, and $\varepsilon = 2.22 \times 10^{-16}$. 10 ms shift time or frame size was considered, i.e. $N = 160$ at 16KHz sampling rate. The real-time processing times are listed in Table I, where Stage 1 refers to

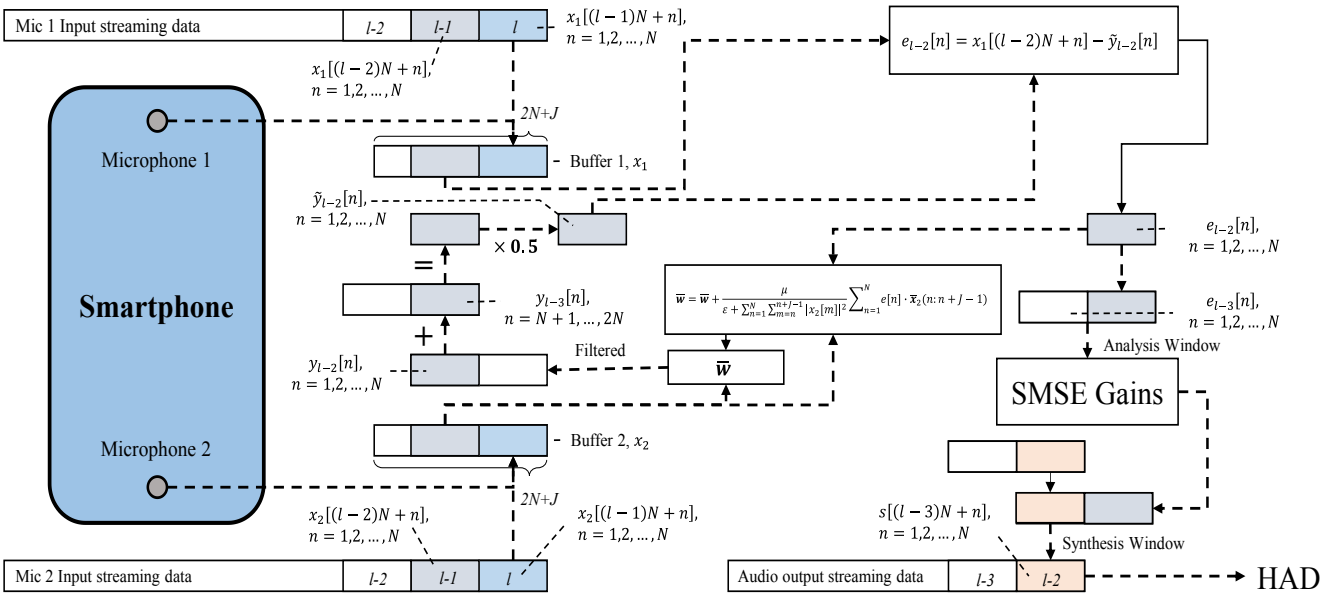

Fig. 1. Framework of real-time implementation on smartphone

the pre-processing stage using two microphones and Stage 2 refers to the SMSE stage.

TABLE I. PROCESSING TIMES ON SMARTPHONE

	Frame	Shift	Stage 1	Stage 2
Time (ms)	20	10	0.14	2.99

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In the experiments conducted, the High-quality INTeractomes (HINT) sentences database sampled at 16 kHz was used. 20 clean HINT speech files were added to three different noise types (babble, driving-car, machinery). To determine the proper value for ρ, the composite measure for speech distortion in [9] was used. Different values of ρ were tested for the three noise types. Csig represents the predicted speech distortion, Cbak represents the predicted noise distortion, and Coval represents the predicted overall quality.

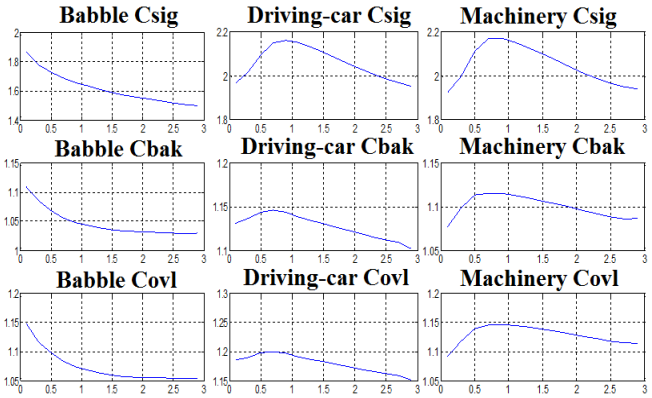

Fig. 2. Composite measure comparison versus different ρ values

Fig.2 gives the composite comparison for different values of ρ for the three noise types at 0 dB SNR. As shown in Fig. 2, the overall quality gives the best result when $\rho = 0.7$ for driving-car and machinery noise. In babble noise condition, the overall quality decreases as ρ increases. This is attributed to the statistical noise characteristics, that is Babble noise is more non-stationary than the other two types of noise. Thus, the noise power estimation is less accurate especially when the tuning factor is increased. Also as seen in this figure, underestimating the noise power yields less speech distortion.

To evaluate the performance of the single microphone SE stage (stage 2) with and without the dual microphone pre-processing (stage 1), synthetic data for dual microphone were generated via MATLAB. The two microphones were separated by 0.13 meters on the back of the smartphone Nexus 6. Two sound sources (speech and noise) were used, each 3 feet away from smartphone. Speech source was close to microphone 1 (Mic 1) and noise source was close to (Mic 2). The azimuth angle was at 90° for speech source and -90° for noise source. The elevation angle was considered to be at 0° for both speech and noise sources. This is because we assume the user is able to adjust the smartphone towards the best position in practice. The far-free field was assumed in conducting the experiments. The experimental setup is shown in Fig. 3. For performance analysis, three objective measures were considered: The perceptual evaluation of speech quality (PESQ) [10] was used to measure speech quality, noise attenuation level (NAL) was used to measure noise suppression [11], and the coherent speech intelligibility index

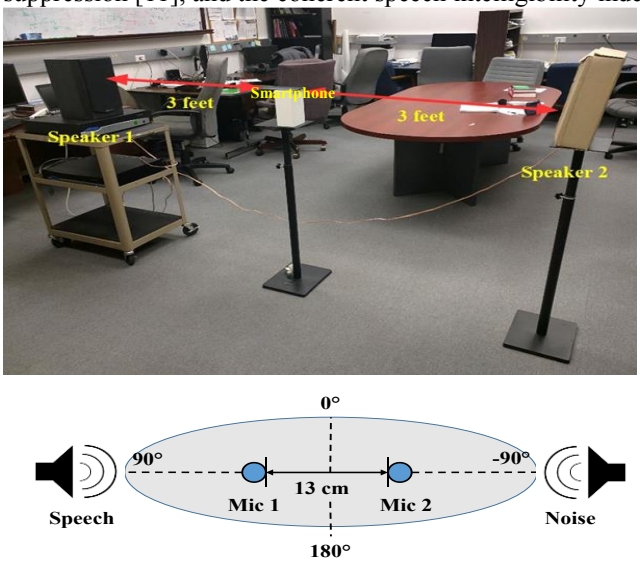

Fig. 3. Experimental setup

(CSII) was used to measure speech intelligibility [12]. The results of the experiments are shown in Table II, where Noisy represents the result of noisy speech (unprocessed), Stage 2 represents the result of the SE pipeline without Stage 1. Stage 1+2 represents the result of the SE pipeline with the first and second stages.

TABLE II. EXPERIMENTAL RESULTS OF THREE OBJECTIVE MEASURES

Noise Type		Objective Measures		
		PESQ	*NAL*	*CSII*
Babble	Noisy	1.31		0.21
	Stage 2	1.17	21.61	0.18
	Stage 1+2	2.28	52.79	0.39
Driving Car	Noisy	1.37		0.22
	Stage 2	1.59	36.09	0.23
	Stage 1+2	2.38	61.35	0.42
Machinery	Noisy	1.35		0.23
	Stage 2	1.53	40.03	0.25
	Stage 1+2	2.31	58.74	0.4

As shown in Table II, with Stage 1+2, an improved performance was obtained in terms of both speech quality and intelligibility. Especially for babble noise, using only single microphone SE sometimes resulted in reduced speech intelligibility (CSII). Fig.4 shows sample plots of the time domain waveforms and the spectrograms of the input and output signals recorded in the experiments at 0 dB SNR. The speech and noise were played in the same way as shown in Fig. 3. The distance from the smartphone to the speech source and to the noise source was each 3 feet. The sentence "A boy fell from a window" from HINT database was calibrated for the average sound pressure level (SPL) of 65dB. The machinery noise was also calibrated for SPL average of 65dB. As shown in Fig.4, the background noise was attenuated significantly while the speech information was preserved. A NAL of 15 dB was achieved. In addition to the objective results reported in this paper, a clinical setup has been done for subjective testing of the developed speech enhancement solution. Fig.5 shows

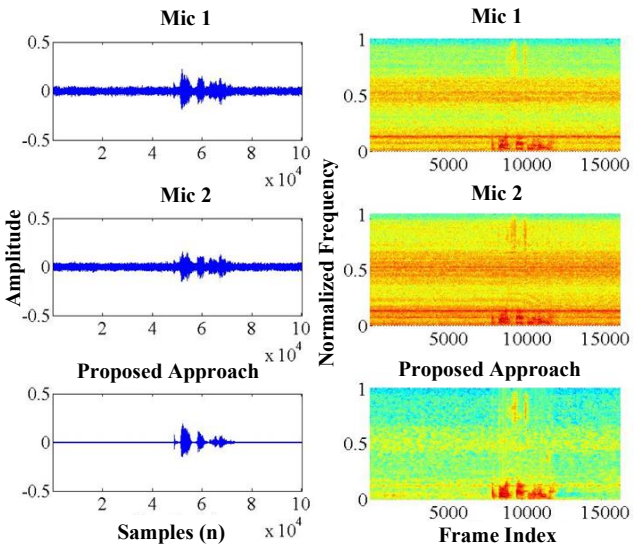

Fig.4. **(a)** Time domain waveforms and **(b)** Spectrograms of recorded speech in machinery noise at 0 dB SNR.

the subjective test results where "UNP " represents the noisy speech, "ENH" represents the enhanced speech. The details of the clinical setup and the subjective results obtained are reported in another accompanying paper in [15]. The subjective results reported in [15] confirm the objective results reported in this paper in terms of improvement of speech perception when deploying the developed speech enhancement on the smartphone platform in a clinical setting with human subjects.

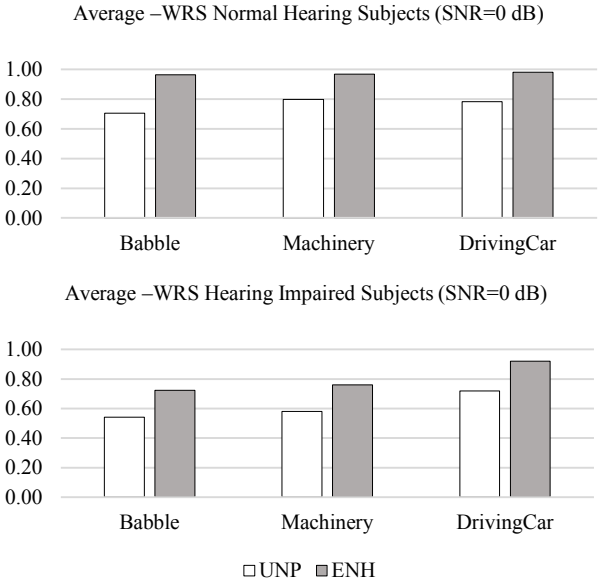

Fig.5 Subjective test results of word recognition rate

V. CONCLUSION

A 2-stage modified speech enhancement (SE) method to improve speech quality and intelligibility for HAD users was introduced in this paper. The algorithms were optimized to run in real-time on Android smartphones. The experiments were conducted to analyze and verify the performance of the proposed SE method using noisy speech signals at 0dB SNRs. Three background noise signals were used; Babble noise, Driving car/traffic noise, and Machinery noise. Three commonly used objective test measures were employed in the experiments for evaluating the quality and intelligibility of the enhanced speech. The experimental test results showed significant improvements of quality and perception of the speech achieved by using the proposed speech enhancement method running on the smartphone platform in real-time. In the future, SE algorithm on smartphone has more benefits than HAD such as real-time monitor, fully control of status and easily upgrading. However, smartphone still cannot fully replace HAD because of power consumption, wearing comfort and manageability by elder people.

References

[1] D. L. Blackwell, J. W. Lucas, T. C. Clarke, "Summary health statistics for U.S. adults: National Health Interview Survey, 2012", *National Center for Health Statistics. Vital Health Stat*. vol. 10, no. 260, 2014.

[2] Quick Statistics (n. d.) retrieved from http://www.nidcd.nih.gov/health/ statistics/pages/-quick.aspx.

[3] Hearing Aids (n. d.) retrieved from http://www.ent.uci.edu/clinical-specialties/ ear-surgery/hearing-aids.

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

[5] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans., Audio, Speech and Lang. Process.*, vol. 16, no. 6, pp. 1112-1123, Aug, 2008.

[6] L. S. Hsieh and S. L. Wood, "Performance analysis of time domain block LMS algorithm," *IEEE International Conference on Acoustic Speech Process.*, Minneapolis, vol. 3, pp 535-538, Apr, 1993.

[7] B. F. Boroujeny and K. S. Chan, "Analysis of the frequency-domain block LMS algorithm," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2332-2342, Aug, 2000.

[8] D. Y. Zhao, X. D. Lu and M.S. Xiang, "Block NLMS cancellation algorithm and its real-time implementation for passive radar," *Radar Conference 2013, IET Int.*, Xi'an, pp. 1-5, Apr, 2013.

[9] ITU-T Recommendation, P. 835-2003, *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm*.

[10] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *ICASSP 01*, vol. 2, pp. 749-752, 2001.

[11] G. Kannan, Ali. A. Milani, Issa M.S. Panahi, R. Briggs "An Efficient Feedback Active Noise Control Algorithm Based on Reduced-Order Linear Predictive Modeling of fMRI Acoustic Noise," *Biomedical Engineering, IEEE Trans.*, vol. 58, no. 12, pp. 3303–3309, Dec. 2013.

[12] P. C. Loizou, "Objective Quality and Intelligibility Measure" in *Speech Enhancement: Theory and practice*, 2nd ed. Boca Raton, CRC press, 2013, ch. 11, sec. 4, pp 561-564.

[13] Android platform versions. Retrieved from https://developer.android.com/ about/dashboards/index.html.

[14] Android native development kit. Retrieved from https://developer.android.com/tools/sdk/ndk/index.html.

[15] I. Panahi, N. Kehtarnavz, L. Thibodeau. "Smartphone-based noise adaptive speech enhancement for hearing aid applications," accepted as an invited paper, IEEE EMBC Conference, 2016.

[16] P.C. Loizou, G. Kim, "Reasons why current speech enhancement algorithms do not improve speech intelligibility and suggested solutions", IEEE Trans. Audio, Speech, Lang. Proc., V. 91, No.1, 2010, pp.47-56.

[17] B. F. Boroujeny and K. S. Chan, "Analysis of the frequency-domain block LMS algorithm," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2332-2342, Aug, 2000.