Single Channel Speech Enhancement using a Complex Spectrum Method

A. M. Mutawa

***Abstract*—**Speech enhancement plays an important role in speech communication systems. Speech signal enhancement in an additive noise environment in speech recognition and speaker verification system is still a challenging task. In speech enhancement process, the spectral analysis method has more advantageous than other methods due to its simplicity and effective localization of noise components in the signal. But, this method does not analyze the phase information for efficient speech enhancement. This present work proposes a modified phase spectrum compensation method for speech enhancement in a single channel environment that analyzes both the magnitude and phase spectrum of the speech signal. The performance of the proposed method is compared with that of three conventional methods (Spectral Subtraction (SSUB), Minimum Mean Square Error (MMSE) estimator, and Phase Spectrum Compensation (PSC)) through four different objective measures: Log spectral distance (LSD), log likelihood ratio (LLR), itakura-siato (IS) measure, and short-time objective intelligibility (STOI). The experimental results show that the three objective measures (LLR, LSD, and STOI) of the proposed method gives better results over the conventional methods in four different noise Signal to Noise Ratio (SNR).

***Keywords*—**Complex spectrum, Noise reduction, Objective measures, Speech enhancement

I. Introduction

SPEECH enhancement is a field of research that focuses on improving the quality of a speech signal under different environments, and it has become a very popular research topic in recent decades. Speech enhancement plays a key role in the speech communication process. Speech communication is one of the important modes of communication between humans and between human and machine. In real life, the users are expecting that the speech communication system should be so robust to work on any environment at any time [1]. Speech communication involves two important processes namely speaker verification and speech recognition. Interferences due to noise is significantly affecting the performance of any speech communication system and it affects the quality of the original speech signal by some ratio and it is measured using Signal to Noise Ratio (SNR). In specific, improving the speech quality of the signal under moderate to high noise (-5 dB to 15 dB) is highly challenging [1], [2]. There are several kinds of research in the literature over the past several decades discusses the speech enhancement process in a noisy environment [3]–[5].

Besides, the type of environment or surrounding also plays a key role in speech enhancement process. Because, the signal can be acquired from a different environment such as additive noise, reverberation, filtering, and clipping, etc. Hence, it is highly important to estimate the original speech information from the noise effects in the speech signal processing for developing intelligent speech communication systems. Recent developments in speech communication devices such as speech assisting devices, speech communication systems (mobile phones), hearing aids, and cochlear implants are highly sensitive to noise information present in the signal, which must be carefully removed from the original signal for efficient sound reproduction. This present work mainly focused on analyzing the degradation due to additive noises of different SNR from moderate to high values (0 dB, 5 dB, 10 dB, and 15 dB).

The major interest on speech enhancement methodology is to suppress the effect of noise in the original speech signal to improve its quality. It is one of the common problems in either single channel (one microphone) system or multi-channel (more than one microphone) system. Most of the earlier work in the literature focused on investigating the speech enhancement process in single channel microphone system due to its size, cost and computational efficiency [3]–[6]. The most successful method of speech enhancement depends on two major factors: (i) how effectively the method localizes the noise component in the signal and (ii) how intelligently it reduces the effects of noise to enhance the speech signal. Most conventional speech-enhancement methods detect the unvoiced region in the signal as noise or vice-versa [6], [7]. Noise is mainly due to artefacts or environmental factors, and the spectral analysis method can effectively distinguish the unvoiced region from the signal. Furthermore, in place of reducing the effect of noises, the speech enhancement algorithms remove the original signal information. Thereby, the performance of any speech enhancement algorithms depends on, (a) parameter settings of the algorithms (b) the value of SNR (c) type of noise and its environment and (d) calculation of noise estimation [7]. Therefore, it is always challenging to design and develop an intelligent speech enhancement algorithm that is suitable for environments with different noise backgrounds. Hence, it is highly evident to develop an intelligent and adaptive speech enhancement method for efficient speech enhancement in real-life

This work was funded by Kuwait Foundation for the Advancement of Science (KFAS) grant number PR1718SM05

A. M. Mutawa is with the Computer Engineering Department, Kuwait University, 13060, +(965) 24987160, Kuwait (e-mail: dr.mutawa@ku.edu.kw).

applications and it is an open question for researchers over the world.

Some of the most popular speech enhancement algorithms in the literature are the spectral subtraction (SSUB) method, power spectrum compensation (PSC) method and minimum mean square error (MMSE) method [5], [8]–[13]. In order to analyze the short-time stationary properties of the speech signals, many of the speech enhancement algorithms utilized a time-frequency representation of the signals in specific short-time Fourier transform (STFT) for speech enhancement process [14]. In general, the output of the STFT is a complex coefficient which has both magnitude and phase values. Most of the research work in the literature utilized a magnitude component for speech enhancement [15]. But, recently, the phase value also considered for efficient noise suppression in speech enhancement [5], [16].

The spectral subtraction method is the most common, popular and traditional method of additive noise cancellation used in speech enhancement. In this method, the noise spectrum is assessed during the silence periods in speech sample and it is subtracted from the noisy speech signal spectrum to estimate clean speech. Here, the method had an assumption that, the magnitude spectrum of the noise is constant and the speech signal is stationary over short-time. Thereby, the effective noise cancellation through this method depends on the calculation of noisy spectrum magnitude and phase values are not considered for speech enhancement. However, this method has a limitation in introducing spectral artifacts in noise cancellation process and various works in the literature have addressed this issue on improving the performance of spectral subtraction method in speech enhancement methods [12], [17].

$$|Y(f)| = |X(f)| - \alpha\left|\overline{N(f)}\right| \quad (1)$$

where *Y(f)* is the spectrum of an original speech signal, *X(f)* is the spectrum estimate of a noisy speech signal, *N(f)* is the average spectrum estimate of the noise signal, and α is a constant. In this work, the value of α is equal to one for spectrum subtraction; a value greater than 1 denotes over-spectral subtraction.

In this work, for a given noisy speech sample, we determine the phase value of a voiced clean speech signal using STFT. Furthermore, we assume that the phase is uniformly distributed and independent of amplitude [6]. Under these assumptions, all four performance measures are computed from the given speech samples. However, we find in this work that the voiced sound neighboring phase values are highly correlated and that the phase trajectories are highly correlated with spectral amplitude. Thus, we conclude that using the noisy phase is only optimal under the limiting assumptions of independence and a uniform phase distribution. In STFT, the window is selected to trade off the width of its main lobe and attenuation of its side lobes to preserve most of spectral information of the signal.

This paper is organized as follows. An introduction to speech enhancement algorithms and their significance is provided in section 1. The materials and methods used to improve speech quality using the proposed methodology are described in section 2. Section 3 presents the experimental results and discussion of algorithm performance under different types of noises. Finally, the conclusions and limitations of the present work are presented in section 4.

II. Experimental Methods

A. Database

This present work used the international standard NOIZEUS database for speech enhancement [17]. This database contains the speech recordings of 30 sentences from the IEEE sentence database produced by three male and three female speakers [17], [18]. Each speaker has spoken five sentences and recorded using Tucker Davis Technology (TDT) in a Speech Processing Lab at University Texas, Dallas, USA at a sampling frequency of 25 kHz and later its downsampled to 8 kHz. Since most of the speech intelligibility application involve the processing signal frequency to a maximum of 10000 Hz to cover most important frequency components for signal intelligibility [17]. Each sentence is corrupted by eight different types of real-world noises (Babble, Car, Exhibition hall, Restaurant, Street, Airport, Train station, and Train), and all the sentences include all of the phonemes in the American English language. The intermediate reference system (IRS) filters are used to obtain clean and noisy signals. A noise segment of the same length as that of the filtered clean speech signal was obtained by randomly from the noise recordings. The extracted noises segments are artificially added to the fileted clean speech signal in order to reach the desired SNR levels. A short description of the database is given in Table I. For more details regarding the database can be found from [17].

B. Proposed Speech Enhancement Method

The proposed method employs the analysis modification and synthesis (AMS) framework. Analysis-modification-synthesis (AMS) framework is used in most of the single-channel speech enhancement process for effective speech enhancement in the spectral domain [17]. The AMS involve three-stage process namely, (i) analysis – here the input speech samples are processed using short-term Fourier Transform (STFT), (ii) modification – the noisy spectrum undergoes some modification in its spectrum to reduce its effect in the original signal (iii) synthesis – extraction of original speech signal using inverse STFT and overlap method. This present work analyzed the modification stage by using a complex spectrum method to effectively reduce the interference from the noise in the original speech signal for speech enhancement.

Speech is assumed to be quasistationary and is analyzed framewise. We assume that at each time instance *n*, the clean speech signal $x_t(n)$ is degraded by additive noise $v_t(n)$, and the noisy signal is derived as $y_t(n)$.

In an additive noise model,

$$y_t(n) = x_t(n) + v_t(n) \quad (2)$$

where $y(n)$ is the noisy speech signal, $x(n)$ is the clean speech signal, and $v(n)$ is the additive noise in the time domain. t is the frame number, t=1,2, 3, ….N, and N is total number of frames.

TABLE I. DESCRIPTION OF THE SPEECH DATABASE.

Segment no	Parameter	Values
1.	Total number of speech signals	30
2.	Total number of additive noises	8
3.	Total number of noisy signals	960 samples: 30 clean signals × 8 types of noises × 4 SNRs
4.	SNR ranges	0 dB, 5 dB, 10 dB, and 15 dB
5.	Min duration of clean/noisy signals	2.116 sec
6	Max duration of clean/noisy signals	3.508 sec
6.	Sampling frequency	25000 Hz
7.	Downsampling frequency	8000 Hz

In the frequency domain, Equation (1) becomes

$$Y_t[\omega_k] = X_t[\omega_k] + V_t[\omega_k] \quad (3)$$

where $Y_t[\omega_k]$, $X_t[\omega_k]$ and $V_t[\omega_k]$ are Discrete Short Time Fourier Transform (DSTFT) representations of output, input and noise spectrum, respectively, and k is the k^{th} discrete frequency. In this work, we assume that the harmonic frequencies and amplitudes are constant for a given length of speech signal using STFT.

The proposed method is illustrated in Fig. 1. It is based on Phase Spectrum Compensation method given by Paliwal et al. [13], [14]. In which conjugate symmetry of DSTFT is used to cancel noise components by changing angle of DSTFT spectrum.

The proposed method uses power spectral density (PSD) of the speech signal, which gives power per unit frequency, to detect the presence of speech and noise in a given frame. An average power spectral density, P_t, is computed for each frame t.

The phase spectrum compensation function is

$$\Psi_t[\omega_k] = \phi[\omega_k]\ Z_t[\omega_k] \quad (4)$$

where $\psi(\omega_k)$ is an antisymmetric function based on antisymmetric property of phase and is kept same for all frames, given as

$$\phi(\omega_k) = \begin{cases} 1, & 0 < \frac{k}{N} < \frac{1}{2} \\ -1, & \frac{1}{2} < \frac{k}{N} < 1 \\ 0, & otherwise \end{cases} \quad (5)$$

$Z_t(\omega_k)$ is the inverted power spectral density function scaled by a constant σ and computed as

$$Z_t[\omega_k] = \sigma \bar{P}_t \quad (6)$$

Inverted Power spectral density $\bar{P}_t$ is obtained by subtracting average power spectral density of each speech frame t from the maximum power spectral density of frames. The higher value of PSD of frames shows the high noise content and smaller value of PSD of frames indicates the high speech content. This allows suitable change for compensation.

Therefore, the modified DSTFT Noisy spectrum is given as,

$$Y_\Psi[\omega_k] = Y_t[\omega_k] + \Psi_t[\omega_k] \quad (7)$$

The modified phase spectrum is computed as,

$$\angle Y_\Psi[\omega_k] = ARG\,|\,Y_\Psi[\omega_k]\,| \quad (8)$$

where ARG is complex angle function, $Y_t[\omega_k]$ is the output spectrum, and $\Psi_t[\omega_k]$ is the noisy spectrum.

The enhanced complex spectrum is estimated as

$$\hat{X}[\omega_k] = |Y_\Psi[\omega_k]|\, e^{j\,\angle Y_\Psi[\omega_k]} \quad (9)$$

It is then converted into time domain using Inverse STFT . Finally, the overlap add method is applied to get enhanced time domain signal, $\hat{x}_t(n)$.

C. Objective Speech Quality Measures

In general, the impact of noise in a signal degrades its quality and it is always non-uniform. Objective speech quality measures are used to analyze the distortion levels in a signal

on each frame over time [19]. In specific, the speech frequency varies over the time and a sequence of phonemes are used to produce the speech. Thereby, the magnitude of background noise effect varies in a speech. Though numerous performance measures are used in the literature for assessing the performance of speech enhancement methods, we focused to analyze the four most important measures such as Log spectral distance (LSD), Log Likelihood Ratio (LLR), Itakura Siato (IS) measure, and Short-time objective intelligibility (STOI) for performance comparison. These objective measures used to quantify the effect of background noises in the signal and to compare the performance of speech enhancement in different algorithms [20], [21].

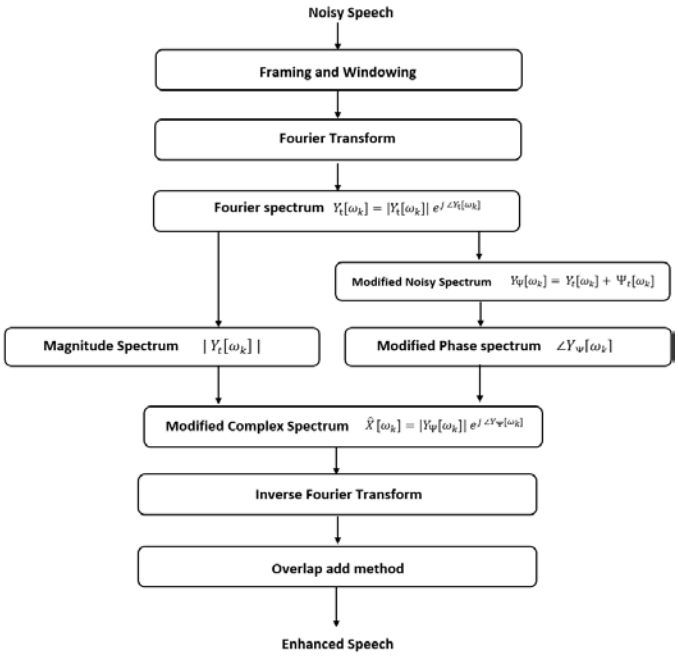


Fig. 1: Overview of the proposed speech enhancement method

1) Log Likelihood Ratio

The speech production process can be effectively modeled using linear prediction (LP) models. Most of the objective performance measures depend on the calculation of a distance between two sets of linear prediction coefficients (LPC) calculated on the original and the enhanced speech. Log likelihood ratio (LLR) is one of the most common types of distance measure used in speech recognition applications. It was first introduced by Itakura as a distance measure for speech recognition applications [22], [23]. This method has since been applied in many speech-processing applications, such as speaker verification, speaker recognition, and speech recognition.

The log likelihood Ratio is defined as

$$LLR(\overrightarrow{s_e}, \overrightarrow{s_o}) = log\left(\frac{\overrightarrow{s_e} R_c \overrightarrow{s_e}^T}{\overrightarrow{s_o} R_c \overrightarrow{s_o}^T}\right) \tag{10}$$

where $\overrightarrow{s_o}$ is the Linear Predictive Coding (LPC) vector coefficient of the original signal frame, $\overrightarrow{s_e}$ is the LPC vector coefficients of an enhanced signal frame and R_c is the autocorrelation coefficient matrix of the original speech signal. In this work, only the smallest 95% of frame LLR values are considered to compute the average LLR value, and values of LLR within the range of 0 to 2 are considered in this work to avoid the influence of outliers.

2) Log Spectral Distance

The distance measure between the feature vectors of the original speech signal spectrum to the enhanced speech signal spectrum. It is always a symmetric measure unlike to IS and LLR measure [22]. It is defined as

$$LSD(\overrightarrow{s_e}, \overrightarrow{s_o}) = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\, log_{10}\left(\frac{S_e(\omega)}{S_o(\omega)}\right)\right] d\omega} \tag{11}$$

The average value of spectral distortion over a large number of frames between the LPC log power spectrum of the original signal, $S_o(\omega)$, to the LPC log power spectrum of an enhanced speech signal, $S_e(\omega)$, gives us the quantity of LSD.

3) Itakura-Siato Distance Measure

The Itakura-Siato (IS) distance measure is defined as [23]

$$IS(\overrightarrow{s_e}, \overrightarrow{s_o}) = \frac{{\sigma_o}^2}{{\sigma_e}^2}\left(\frac{\overrightarrow{s_e} R_c \overrightarrow{s_e}^T}{\overrightarrow{s_o} R_c \overrightarrow{s_o}^T}\right) + log\left(\frac{{\sigma_o}^2}{{\sigma_e}^2}\right) - 1 \tag{12}$$

Here, ${\sigma_o}^2$ and ${\sigma_e}^2$ are the gain of the LPC coefficients of the original and enhanced speech signals, respectively. Here, the IS were limited to the range of [0,100] to reduce the number of outliers in the signal.

4) Short-Time Objective Intelligibility

The short-time objective intelligibility (STOI) is a performance measure used to find the correlation between the temporal envelopes of the clean speech signal to the enhanced speech signal in short-time overlapped segments. Initially, the speech samples are short-time segmented using the windowing process, normalizing the windowing coefficients; then, the value of the correlation coefficient is calculated for each segment. Later, the average value of correlation coefficient over all the time segmented speech signal represents the value of speech intelligibility measure. STOI can be considered as an alternative to the speech intelligibility index (SII) or the speech transmission index (STI), when you are interested in the effect of nonlinear processing to noisy speech, e.g., noise reduction, binary masking algorithms, on speech intelligibility [20]–[23].

III. Experimental Results and Discussions

This section presents the experimental results of the present work. All the experiments were performed on the NOIZEUS speech corpus database to analyze the performance of the proposed speech enhancement method and compare it with that of three conventional speech enhancement methods (SSUB, MMSE and PSC). This database has been used by several researchers for speech enhancement applications and

its gender-matched database, and it has 30 IEEE sentences. One of the speech sample (Train) at 15 dB SNR is excluded from this work for analysis. Since the original signal source in the database is corrupted and could not be used for analysis.

Initially, the speech samples from the database are framed using a hamming window method with a window length of 25 ms over the 40% overlapping between the frames. This windowing is applied to all the speech samples in the database. Later, the speech samples are added with an additive Gaussian noise of different SNR of the same window length of original speech signals. Then, the STFT is applied to the noisy speech signals and extracted its magnitude and phase spectrum values.

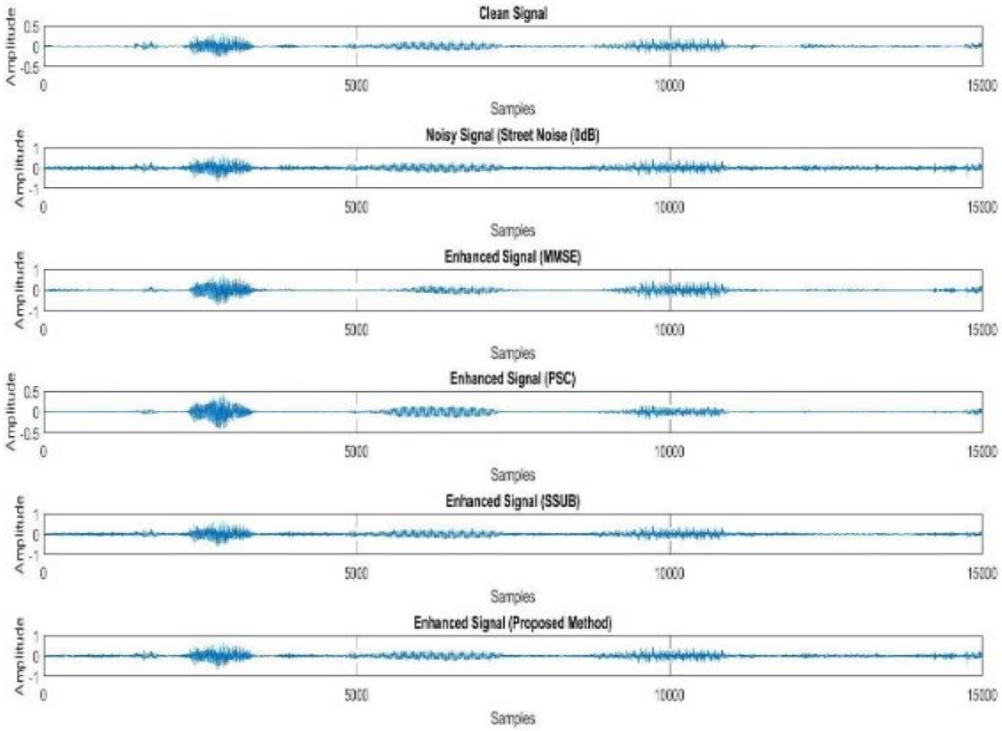
Finally, the speech signals are reconstructed to derive the time domain speech signals using a two-stage approach, namely, the inverse STFT and overlap-add method. Later, four different objective measures are computed from the original (clean) speech signal and enhanced speech signal for performance comparison over each frame. The time series plot of the original (clean) and enhanced speech signal over four different SNRs is shown in Fig. 2. This plot represents the signal variations over the 15000 samples for different SNRs. From the Fig. 2, it is highly evident that, the proposed method has effectively suppressed the additive Gaussian noise through the PSD-PSC speech enhancement method and the output (enhanced) speech signal is highly matched with the original (clean) speech signal. Figs. 3-6 illustrates the performance of log spectral distance measure of four different speech enhancement methods over eight types of noises.

Fig. 3 shows the spectral power distribution over frequency for a speech signal with car noise at 0 dB, 5 dB, 10 dB, and 15 dB. Most studies related to speech enhancement method are limited to a few types of noises, and very few have analyzed the performance of a speech enhancement method over eight types of noises. The average value of LSD over 30 speech signals in 0 dB noise is shown in Figs. 4-7. From the results, it indicates that the proposed speech enhancement method gives the lowest value of LSD over the other three methods. In specific, the performance of the proposed method shown the improvement in speech enhancement process over conventional methods. Also, the average value of LSD close to zero (enhanced signal is almost similar to clean signal) when the noise SNR increases. Among the four different speech enhancement methods, the PSC has a higher LSD value and then followed by SSUB (Spectral Subtraction method).

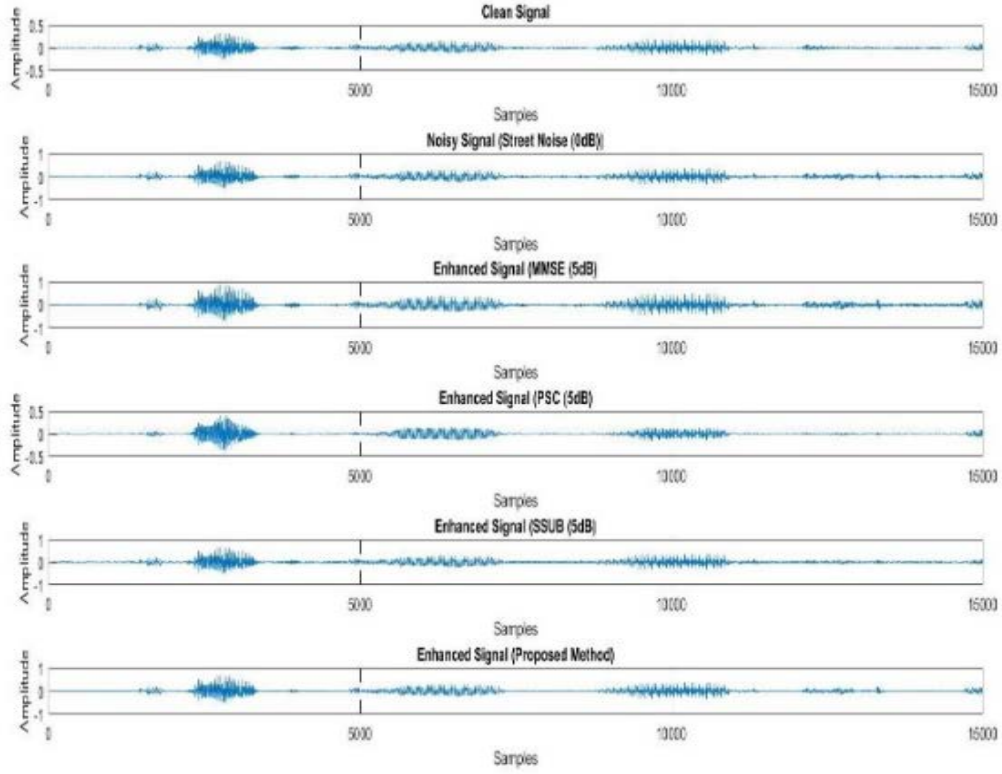
The LSD variations at different noise levels (5 dB, 10 dB and 15 dB) are shown in Figs. 4-7. The experimental results confirm that the proposed method achieves lower LSD values than conventional speech enhancement methods for most types of noise. The PSC method does not perform well in enhancing the quality of the speech signal after filtering under four noise levels (0 dB, 5 dB, 10 dB and 15 dB). There are no large differences between the MMSE, SSUB and proposed method in speech enhancement process when using the LSD measure. However, the proposed method achieves better performance than the other methods due to the inclusion of phase variations in the speech enhancement process.

Table II shows the performance of STOI, LLR and IS measure of four different noise SNRs over eight types of noises. The value of STOI, LLR and IS represents the average value over 30 speech signals. From the experimental results, the proposed method achieved a higher STOI value in comparison with other methods over four different SNRs. Also, the value of STOI steadily increases while the noise SNR increases. Though there are no large differences in STOI values among the speech enhancement methods, the proposed method perform well in comparison with the conventional methods. In the case of LLR, the proposed method achieves the lowest value in comparison with the other three speech enhancement methods. But, in the case of Itakura-Siato measure, spectral subtraction and MMSE method give optimal performance in speech enhancement over the proposed method.

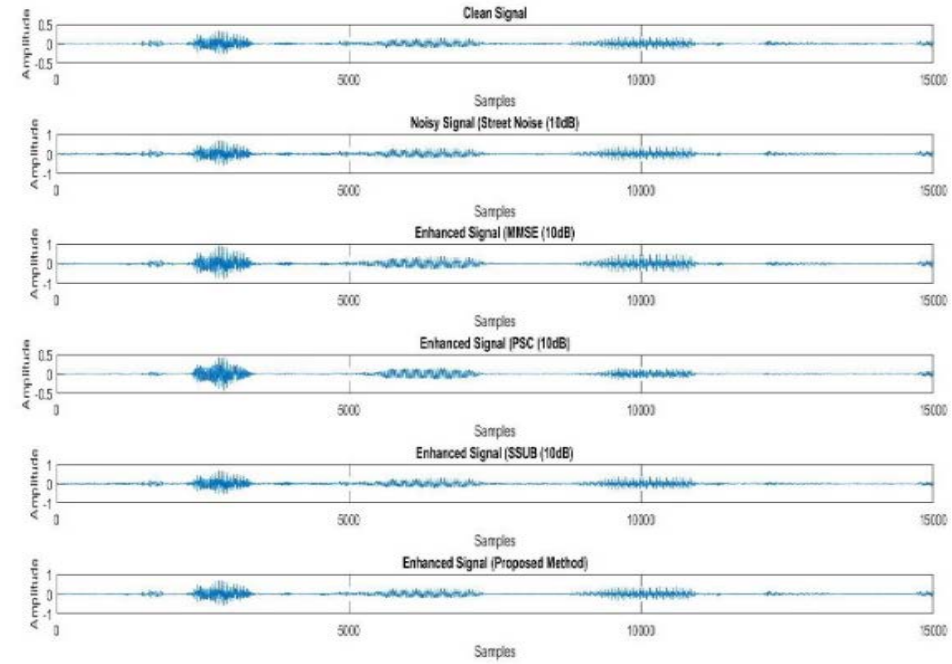
The proposed method is derived from the PSC method, with modification in phase spectrum computation; hence, these two methods yield a higher IS value than do MMSE and SSUB. The experimental results show that the proposed speech enhancement method performs well in three different, realistic SNRs in real time and yields optimal results at lower SNR values (e.g., 0 dB). This strong performance at lower SNR values occurs because the magnitude and phase spectrum of these values are more valuable for the speech enhancement process than are higher values. Though the present work gives better performance over conventional methods on three different noise SNRs over 30 different speech samples, it has some limitations. Firstly, the proposed method analyzes the speech signal of fixed frame duration, it is also important to analyze the performance of the proposed method with different frame durations. Secondly, the proposed method should be tested with other open-source and international standard speech corpus to validate its efficiency in speech enhancement process. Thirdly, the proposed method is evaluated only through objective measures, but, it is also important to analyze the performance of the proposed system using subjective measures and with a combination of subjective and objective measures. Lastly, the proposed method is evaluated through only a selected and most popular objective measures for analyzing its performance. In future, the researchers will work on addressing all the above five limitations to fine-tune the proposed method for improving its robustness on speech enhancement process.


Clean Signal
Noisy Signal (Street Noise (0dB)
Enhanced Signal (MMSE)
Enhanced Signal (PSC)
Enhanced Signal (SSUB)
Enhanced Signal (Proposed Method)
Amplitude
Samples
0
5000
10000
15000

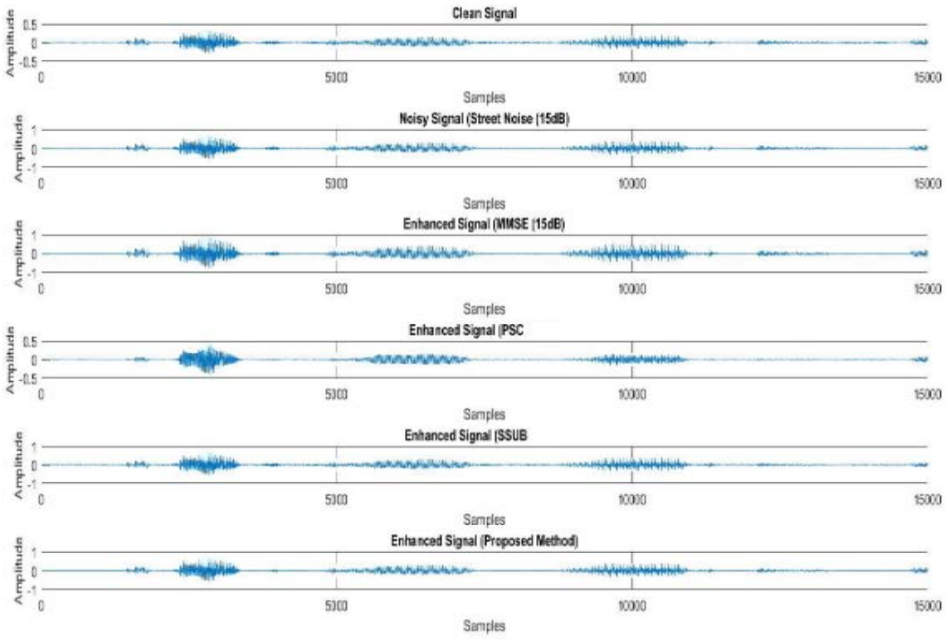
(a)


Clean Signal
Noisy Signal (Street Noise (0dB))
Enhanced Signal (MMSE (5dB)
Enhanced Signal (PSC (5dB)
Enhanced Signal (SSUB (5dB)
Enhanced Signal (Proposed Method)
Amplitude
Samples
0
5000
10000
15000

(b)



(c)



(d)

Fig. 2: Time series signal plot of the speech signal over four different SNRs: (a) 0 dB (b) 5 dB (c) 10 dB and (d) 15 dB

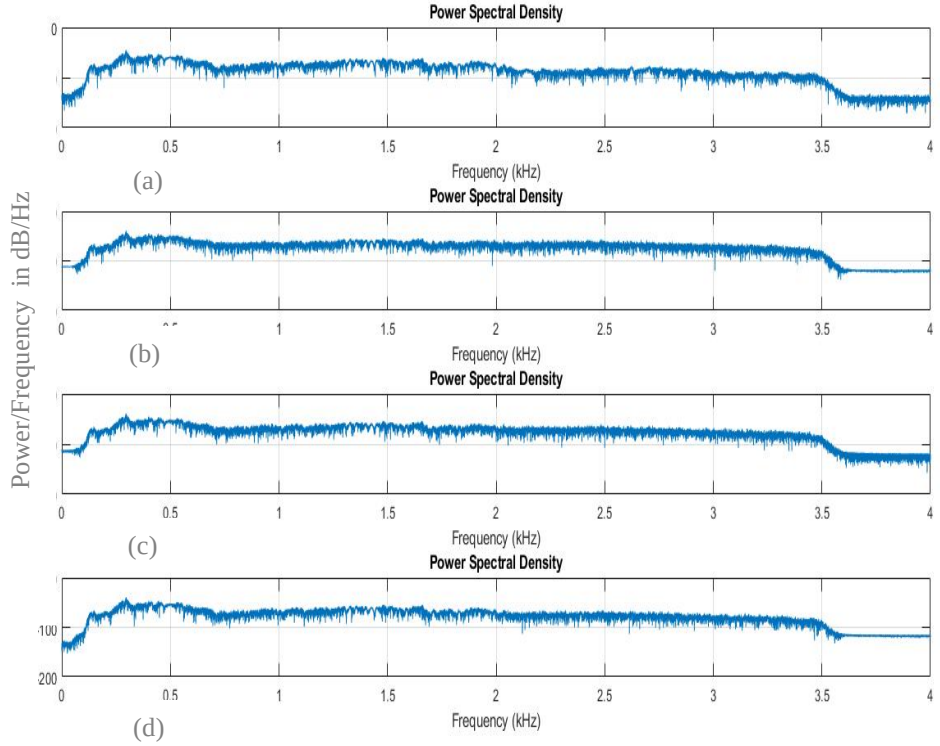


Fig. 3: Power spectral density plot of speech signal with noise (car) at different noise levels: (a) 0 dB (b) 5 dB (c) 10 dB and (d) 15 dB

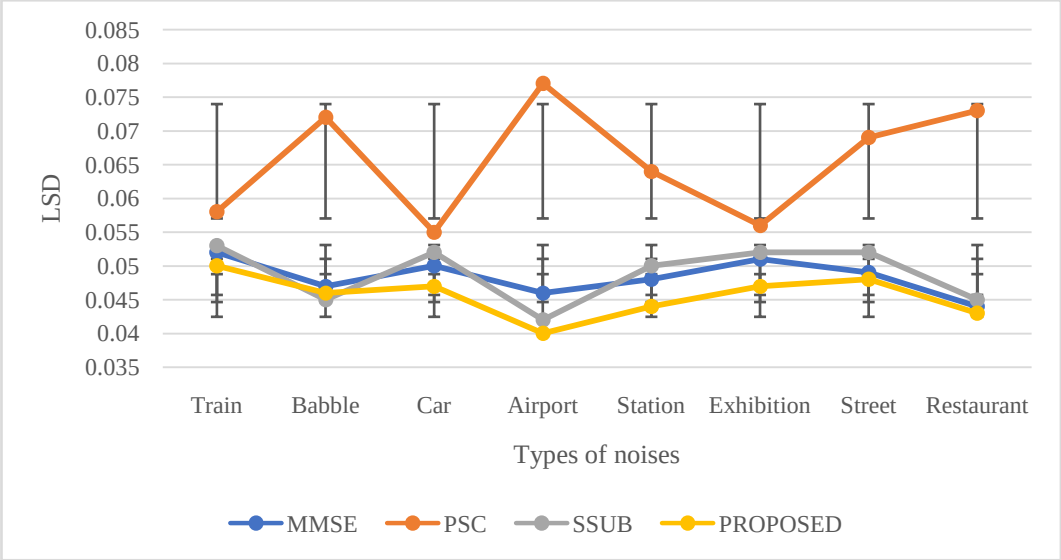


Fig. 4: Average LSD values over eight types of noise at 0 dB

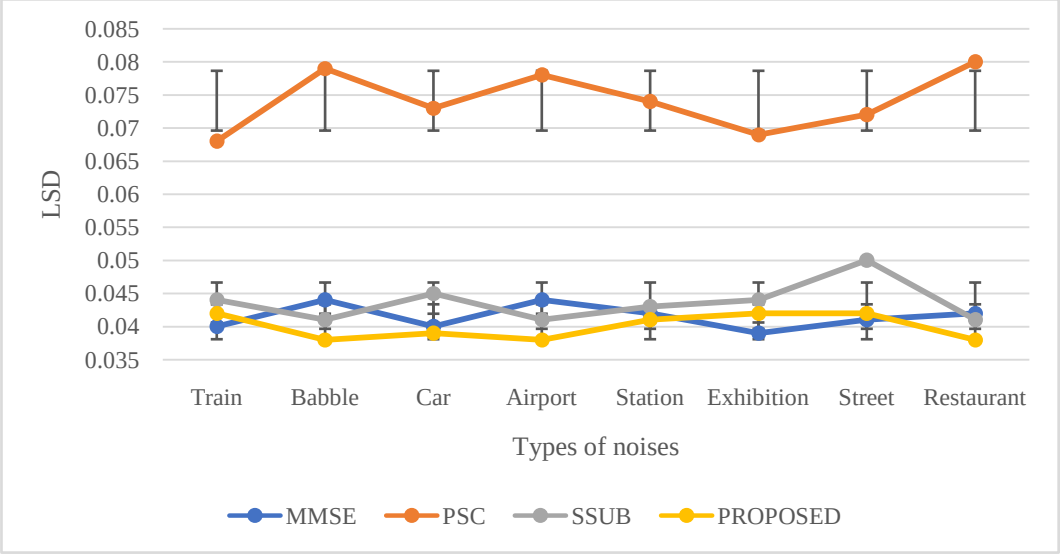


Fig. 5: Average LSD values over eight types of noise at 5 dB

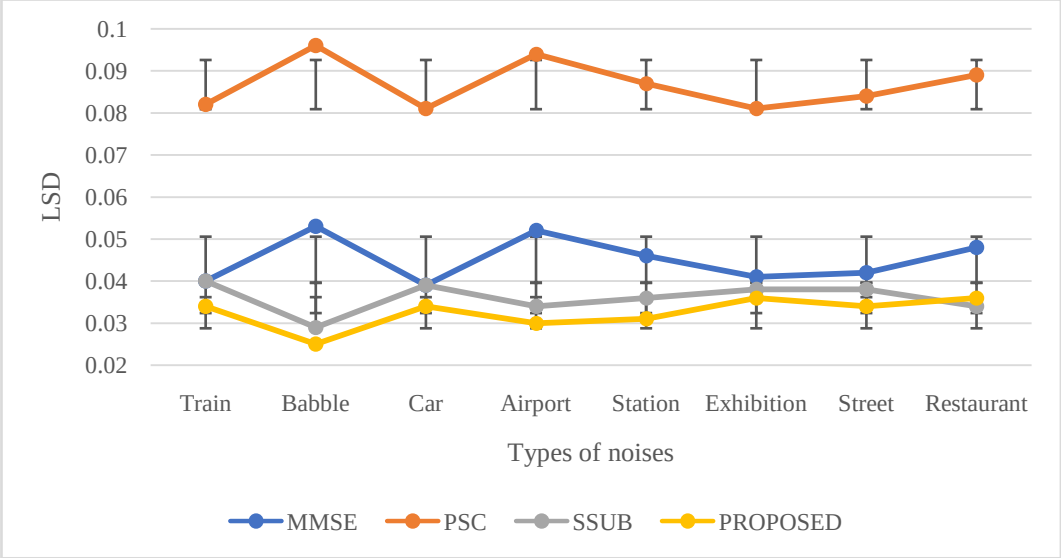


Fig. 6: Average LSD values over eight types of noise at 10 dB

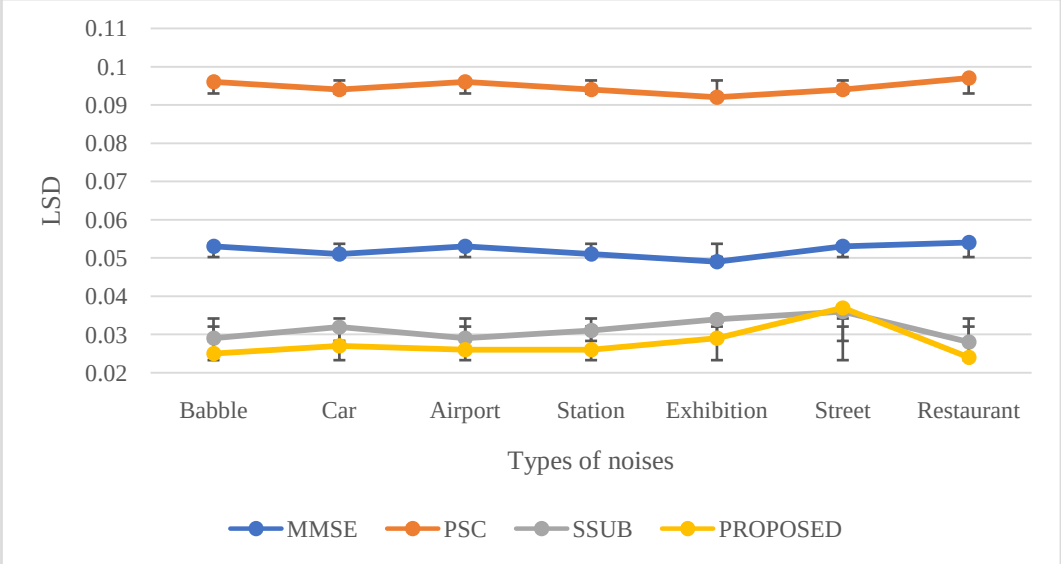


Fig. 7: Average LSD values over eight types of noise at 15 dB

TABLE II. PERFORMANCE OF STOI, LLR AND IS MEASURE OVER FOUR DIFFERENT NOISE SNRS

Noise	STOI				LLR				IS MEASURE			
	MMSE	PSC	SSUB	NEW	MMSE	PSC	SSUB	NEW	MMSE	PSC	SSUB	NEW
0 dB												
Train	0.739	0.771	0.793	0.823	1.805	2.762	1.678	1.436	2.077	13.128	1.577	1.981
Babble	0.745	0.766	0.779	0.807	2.374	3.502	2.299	1.558	4.019	75.040	2.146	7.162
Car	0.727	0.777	0.782	0.808	2.313	2.780	1.390	1.273	3.472	18.983	1.427	1.624
Airport	0.771	0.766	0.797	0.824	2.302	3.635	2.482	1.417	3.272	92.156	1.850	9.579
Station	0.748	0.766	0.779	0.810	2.409	3.608	2.240	1.478	3.415	55.126	1.755	5.534
Exhibition	0.762	0.772	0.815	0.849	2.379	3.732	2.234	1.475	3.708	91.364	1.875	6.338
Street	0.745	0.766	0.783	0.819	2.247	2.954	1.583	1.492	3.138	23.435	3.138	1.945
Restaurant	0.834	0.805	0.862	0.869	1.854	3.692	3.361	1.543	5.349	174.229	1.091	27.803
5 dB												
Train	0.851	0.824	0.873	0.903	1.882	3.234	1.986	1.437	2.088	48.503	1.697	4.860
Babble	0.855	0.827	0.871	0.895	2.072	3.537	2.240	1.516	2.499	87.036	2.305	10.144
Car	0.841	0.828	0.863	0.889	2.577	3.594	2.538	1.509	5.490	100.611	2.297	15.043
Airport	0.867	0.827	0.881	0.899	2.038	3.713	2.809	1.491	2.650	120.237	2.390	22.793
Station	0.847	0.826	0.872	0.897	2.527	3.447	2.070	1.403	5.498	83.060	1.880	7.300
Exhibition	0.871	0.823	0.891	0.912	2.452	3.740	2.385	1.520	3.451	74.432	1.979	8.454
Street	0.841	0.823	0.864	0.895	1.579	3.448	2.219	1.396	1.836	78.429	1.836	8.069
Restaurant	0.873	0.825	0.886	0.906	1.853	3.587	2.774	1.464	2.171	98.614	2.153	22.141
10 dB												
Train	0.921	0.850	0.924	0.939	2.611	3.406	2.106	1.467	6.749	76.698	2.015	7.744
Babble	0.954	0.858	0.946	0.950	2.091	3.686	2.875	1.516	2.885	126.185	2.627	29.644
Car	0.917	0.854	0.921	0.937	1.984	3.487	1.916	1.447	2.431	89.947	2.048	6.593
Airport	0.930	0.851	0.931	0.941	2.122	3.677	2.773	1.572	2.807	123.121	2.733	24.600
Station	0.926	0.850	0.926	0.937	2.331	3.687	2.619	1.607	3.678	124.837	2.876	19.316
Exhibition	0.925	0.850	0.935	0.944	2.433	3.572	2.238	1.521	4.437	106.592	2.475	10.738
Street	0.918	0.852	0.923	0.938	2.080	3.682	2.788	1.568	2.729	121.946	2.729	26.213
Restaurant	0.935	0.851	0.936	0.943	2.075	3.675	2.786	1.467	2.747	125.385	2.470	25.909
15 dB												
Babble	0.955	0.860	0.954	0.965	2.242	3.686	2.815	1.559	3.353	128.217	2.922	28.448
Car	0.954	0.860	0.951	0.959	2.034	3.674	2.791	1.582	2.595	124.422	2.932	27.402
Airport	0.956	0.860	0.957	0.967	1.951	3.687	2.922	1.568	2.361	126.868	2.897	32.750
Station	0.954	0.859	0.954	0.962	1.918	3.686	2.894	1.491	2.359	126.715	2.653	31.287
Exhibition	0.956	0.860	0.960	0.966	2.066	3.672	2.462	1.592	2.816	123.975	2.912	16.649
Street	0.952	0.859	0.948	0.954	2.247	2.954	1.583	1.492	3.138	234.346	3.138	1.945
Restaurant	0.957	0.860	0.961	0.970	1.919	3.693	2.860	1.558	2.348	128.577	2.834	29.718

NEW: Proposed speech enhancement method

IV. CONCLUSION

In this work, we investigated a complex spectrum based speech enhancement for single-channel applications. This proposed method analyzes the phase values of the signal besides the amplitude and frequency in conventional speech enhancement algorithms. The performance of the proposed method is analyzed using four objective speech quality measures such as LLR, LSD, Itakura- Siato distance measure, and STOI. The performance of the proposed method is compared with that of three conventional speech enhancement algorithms: Minimum Mean Square Error (MMSE), Spectral Subtraction (SSUB) and Power Spectrum Compensation (PSC). This present experiment revealed that the proposed method performed well in improving the quality of the speech signal over conventional methods and it is confirmed through the three objective measures namely, STOI, LLR, and LSD. Future work should focus on analyzing the performance of the proposed method under different operating conditions, including different noise environments; developing new

performance (composite) measures by combining the objective and subjective measures to evaluate the performance of the proposed algorithm; and utilizing recent machine learning methods, such as deep learning [23], to improve the proposed method's potential in real-time system design.

ACKNOWLEDGMENT

This work was funded by Kuwait Foundation for the Advancement of Science (KFAS) grant number PR1718SM05

REFERENCES

[1] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, pp. 55–66, Feb. 2015.

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: PTR Prentice Hall, 1993.

[3] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *5th Int. Conf. Spoken Lang. Process.*, Durham, NC, 1998.

[4] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study, eleventh international multi-conference on information processing," *Procedia Comput. Sci.*, vol. 54, pp. 574–584, Jan. 2015.

[5] Z. Li, W. Wu, Q. Zhang, H. Ren, and S. Bai, "Speech enhancement using magnitude and phase spectrum compensation," in *IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Okayama, Japan, 2016, pp. 1–4.

[6] S. Singh, A. M. Mutawa, M. Gupta, M. Tripathy, and R. S. Anand, "Phase based single-channel speech enhancement using phase ratio," in *6th Int. Conf. Comput. Appl. Elect. Eng. Recent Advances (CERA)*, Roorkee, India, 2017, pp. 393–396.

[7] K. Kondo, *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Berlin, Heidelberg: Springer, 2012.

[8] A. Stark, K. Wo´jcicki, J. Lyons, and K. Paliwal, "Noise driven short-time phase spectrum compensation procedure for speech enhancement," in *9th Annu. Conf. Int. Speech Commun. Assoc.*, Brisbane, QLD, Australia, 2008, pp. 549–552.

[9] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[11] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio Speech Lang. Process.*, vol. 24, pp. 2009–2022, Aug. 2016.

[12] M. Weibeg and I. Clasesson, *Spectral Subtraction with Extended Methods, Research Reports HK-R, Aug*, 1996.

[13] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, pp. 465–494, Apr. 2011.

[14] K. Wojcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement," *IEEE Signal Process. Lett.*, vol. 15, pp. 461–464, May 2008.

[15] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, pp. 1931–1940, Sept. 2014.

[16] E. Loweimi, S. M. Ahadi, and S. Loveymi, "On the importance of phase and magnitude spectra in speech enhancement," in *19th Iranian Conf. Elect. Eng.*, Tehran, Iran, 2011, pp. 2439–2444.

[17] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, Aug. 2007.

[18] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. 17, pp. 225–246, Jun. 1969.

[19] ITU-T P.56, *Objective Measurement of Active Speech Level*. Geneva, Switzerland: ITU, 1963.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 4214–4217.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, pp. 2125–2136, Sept. 2011.

[22] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality. Advanced Reference Series*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[23] M. Kolbk, Z.-H. Tan, J. Jensen, M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)*, vol. 25, pp. 153–167, Nov. 2017.

Dr. Mutawa got his Ph.D. from Syracuse University, New York (1999) in the field of Artificial intelligence and his current research interest are in Robotics, Expert Systems Artificial Intelligence, Signal Processing, Patter Recognition, Deep Learning and e-Learning. Dr. Mutawa served as a Director of the Office of Engineering Education at the College of Engineering and Petroleum Kuwait University from 2004-2012, and as Assistant Vice President for academic services for computer systems and distant learning from 2012-2014. Currently, Dr. Mutawa is Vice President of Arab Robotic Association from 2016-present, and Founder and President of ROBOTEC consultations in IT and Technologies 2016-present. Director of the office of E-Services and Engineering Education at college of Engineering and Petroleum, Kuwait University, 2017- present.