On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms

Nasir Saleem[1]*, Muhammad Irfan Khattak[1], Elena Verdú[2]

[1] Department of Electrical Engineering, University of Engineering & Technology, Peshawar (Pakistan)
[2] Universidad Internacional de La Rioja, Logroño (Spain)

Received 20 September 2019 | Accepted 27 November 2019 | Published 18 December 2019

Abstract

Many forms of human communication exist; for instance, text and nonverbal based. Speech is, however, the most powerful and dexterous form for the humans. Speech signals enable humans to communicate and this usefulness of the speech signals has led to a variety of speech processing applications. Successful use of these applications is, however, significantly aggravated in presence of the background noise distortions. These noise signals overlap and mask the target speech signals. To deal with these overlapping background noise distortions, a speech enhancement algorithm at front end is crucial in order to make noisy speech intelligible and pleasant. Speech enhancement has become a very important research and engineering problem for the last couple of decades. In this paper, we present an all-inclusive survey on unsupervised single-channel speech enhancement (U-SCSE) algorithms. A taxonomy based review of the U-SCSE algorithms is presented and the associated studies regarding improving the intelligibility and quality are outlined. The studies on the speech enhancement algorithms in unsupervised perspective are presented. Objective experiments have been performed to evaluate the potential of the U-SCSE algorithms in terms of improving the speech intelligibility and quality. It is found that unsupervised speech enhancement improves the speech quality but the speech intelligibility improvement is deprived. To finish, several research problems are identified that require further research.

Keywords

Unsupervised Speech Enhancement, Speech Quality, Speech Intelligibility, Noise.

DOI: 10.9781/ijimai.2019.12.001

I. Introduction

SINGLE channel speech enhancement (SCSE) [1]-[16] is one of the significant researched problems in many speech related applications; such as, Automatic Speech Recognition (ASR) [17], Speaker Identification (SI) [18], Human-Machine interaction [19], etc. The problem occurs whenever an interfering noise signal degrades the target speech signal. The interfering noise signals could be convolutive [20] or additive. The convolutive noise signal is produced because of the reverberation. However, additive noise signals are usually supposed since this supposition expresses the uncomplicated solutions and practically more adequate results have been attained with the algorithms structured on such theory [21] [22]. The additive noise distortions significantly aggravate the quality and intelligibility of the speech signals. For this reason, the objective of the speech enhancement algorithms is to quantify the estimate of the underlying clean speech from the noisy speech to increase the intelligibility and quality of the noisy speech signal [21] [22]. A fundamental structure of the SCSE is shown in Fig. 1. A variety of speech related applications exists in our everyday situations where speech enhancement is required as for example: (i) the humans are present in the noisy environments and communicating on the mobile phones, (ii) listening to a call in the noisy street or in the factory, (iii) sitting in subway or travel in a car. In these situations, a speech enhancement could be used to ease the communication by reducing the noise signals. A number of speech enhancement methods have been designed at the front-end to create robust ASR systems by decreasing the discrepancies between the training and testing stages. In ASR, a speech enhancement method is applied to minimize the noise prior to the feature extraction phase. An additional imperative application of the speech enhancement system is for those individuals using hearing aid devices. The speech signals show extremely redundancy and normal hearing listeners can comprehend the target speech signals even in adverse signal-to-noise ratios (SNRs) [23]-[26]. For instance, a normal hearing individual can comprehend approximately 50% of the words spoken in a multitalker corrupted speech at signal to noise ratio equal to 0 dB [27]. However, for individuals with hearing problem (hearing loss), various speech parts could totally be inaudible or significantly distorted. Therefore,

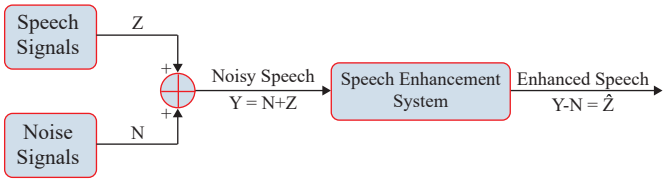

Fig. 1. Single Channel Speech Enhancement.

* Corresponding author.
E-mail address: nasirsaleem@gu.edu.pk

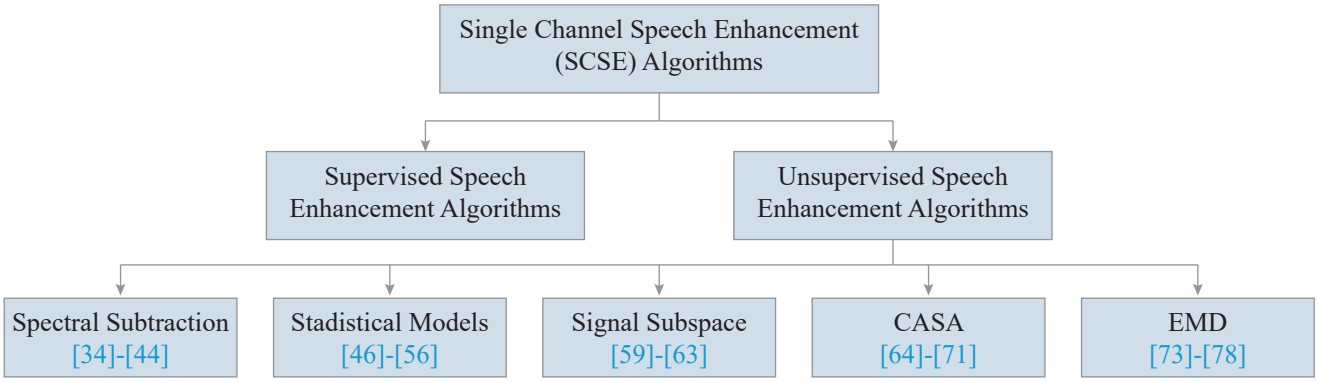

Fig. 2. General Classification of the U-SCSE Algorithms.

the perceived speech signals have small redundancy. Consequently, the individuals with hearing loss feel problem in the noisy environments [28]-[30]. Large attention towards designing the robust speech enhancement algorithms is given to decrease the listening effort and improve the speech intelligibility [31] [32]. The combinations of such algorithms with contemporary digital signal processing systems are implemented in a number of speech related devices.

Single-channel speech enhancement algorithms are divided into two major categories: Supervised SCSE (S-SCSE) algorithms and Unsupervised SCSE (U-SCSE) algorithms. In U-SCSE algorithms, a statistical model is used for speech/noise and the estimate of the underlying clean speech is quantified from the noisy speech devoid of prior facts about speaker identity and noise. Thus, no supervision and classification of the signals is required. Alternatively, the S-SCSE algorithms use models for speech and noise. The model parameters are learned through training of the speech and noise samples and models are defined by mixing the separate models for the speech and noise and the speech enhancement task is performed. In this category, therefore, prior supervision and classification of the speech or noise type is a requisite. The emphasis of this paper is to present a survey on the U-SCSE algorithms.

The remaining paper is organized as follows: Section II shows an extensive review of U-SCSE algorithms in terms of the speech intelligibility and quality. Section III presents experiments performed to evaluate the speech intelligibility and quality potentials of U-SCSE algorithms. Section IV presents the concluding remarks of the survey. Finally, section V presents important research problems which require further study.

II. Classification of U-SCSE Algorithms

This category includes a wide range of U-SCSE algorithms; however, general classification is not limited to the presented algorithms. In U-SCSE algorithms, a statistical model is used. The estimated underlying clean speech is quantified from the input noisy speech utterances devoid of previous facts about speaker identity and noise. A general classification and fundamental framework of the U-SCSE algorithms is shown in Fig. 2-3. In subsequent sub-sections, we provide a taxonomy based review of the U-SCSE algorithms.

A. Spectral Subtraction-based Speech Enhancement Algorithms

Spectral subtraction (SS) based speech enhancement is simple, effective and traditionally one of the pioneer methods proposed for reducing noise distortion. Noise signals are assumed to be additive. Spectral subtraction based speech enhancement algorithms were initially proposed by Boll [33]. In SS, the estimate of the underlying clean speech spectrum could be obtained by subtracting the estimate of noise spectrum from the noisy spectrum. The noise spectrum

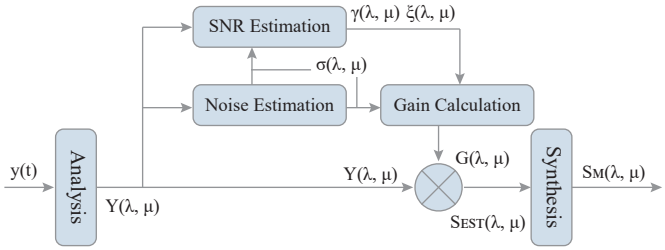

Fig. 3. U-SCSE System.

is estimated and updated during pause periods i.e., absence of the speech signals. The hypotheses for designing such algorithms are: **(i)** the stationary or slowly varying process and, **(ii)** the noise spectra do not vary drastically during updating periods. The enhanced speech is acquired by using inverse transform of the estimated spectrum using noisy phase. According to the basic principle of SS, let us assume that a noisy signal $z(n)$ is composed of the clean speech $s(n)$ and the additive noise signal, $e(n)$

$$z(n) = s(n)+e(n) \tag{1}$$

Computing the STFT of (1), we obtain:

$$Z(\omega,k)=S(\omega,k)+E(\omega,k) \tag{2}$$

Subtract noise magnitude spectrum $|D(\omega,k)|$ from the noisy speech magnitude spectrum $|Y(\omega,k)|$ and finally take the inverse Fourier transform of the difference spectra using the noisy phase to produce the enhanced speech signal, given by equation as:

$$\hat{S}(\omega,k)=[Z(\omega,k)-E(\omega,k)]e^{j\phi_Z(\omega,k)} \tag{3}$$

Since, noise signals are non-stationary and time-variant in the real-world environments; the SS-based enhancement approaches produce negative values for the estimated magnitude spectrum of the clean speech and result in musical noise artifact in enhanced speech. The research is done in near past to reduce the musical noise artifact. Some highly ranked researches on the SS for the speech enhancement are reviewed.

Lu and Loizou [34] proposed a spectral subtraction algorithm based on the geometric approach for the speech enhancement which addressed the inadequacy of the traditional SS algorithm. An efficient scheme to estimate the cross-terms is proposed which is involved in the phase differences between the speech and noise signals. After analyzing the suppression function of the proposed algorithm, it is examined that the algorithm holds the properties of the conventional minimum mean square error (MMSE) algorithm. The evaluation confirmed that geometric approach for the speech enhancement performed considerably better than the conventional spectral subtractive algorithm. A similar approach is also presented in [35].

Paliwal *et al.* [36] examined SS in the modulation domain, an

unconventional acoustic domain for the speech enhancement task, and showed capability of SS in the new domain. Analysis-modification-synthesis (AMS) framework is included and reduced musical noise artifact by applying the modulation-domain based SS algorithm. Moreover, consequences of the frame duration on speech quality have been examined. The outcomes of research indicated that frames with duration with 180-280 msec provided optimized results in terms of the spectral distortions and temporal slurring. For further improvements in the speech quality, a fusion with the MMSE principle has been presented in the short-time spectral domain by joining the magnitude spectrum of the proposed speech enhancement algorithm. Consistent improvements in speech quality have been achieved for different SNRs.

Zhang and Zhao [37] proposed an approach, and performed subtraction on the real and imaginary spectrum independently in modulation-domain. An enhanced magnitude and phase is achieved through the SS approach. Inoue *et al*. [38] provided a theoretical investigation of the musical noise artifact created by the SS on higher order statistics. It is assumed that power SS approach is a common used form. Generalization of SS for the unpredictable exponent parameters has been provided and the quantity of the musical noise artifact has been compared between several exponent-domains. A less musical noise artifact has been observed for a lower exponent spectral-domain and offered good quality and intelligible speech.

Miyazaki *et al*. [39] provided a theoretical examination of the musical noise artifact with an iterative SS method. Iteratively weak-nonlinear signal processing technique has been used to obtain a high quality speech with low musical noise artifact. The generation of musical noise artifact has been formulated by marking changes in kurtosis of the noise spectrum. Optimal internal parameters have been derived theoretically in order to produce no musical noise and explained that with a fixed point in kurtosis yield no musical artifacts.

Antonio *et al*. [40] proposed an improved algorithm based on the SS for real-time noise cancellation and applied the algorithm to the gunshot acoustical signals. A pre-processing approach based on spectral suppression algorithm is applied instead of post-filtering, which requires *a priori* information concerning the direction of arrival of desired signals. Ban and Kim [41] proposed an algorithm for reducing the reverberant noise to the application of remote-talking speech recognition. The SS has been used and the spectra of late reverberant signals are estimated by considering the delayed and attenuated versions of reverberant signals. The unpredictable weight sequences have been estimated via a Viterbi-decoding method based on the reverberation model. The weight sequences are then replaced with fixed weights in SS without estimating the reverberation time.

Hu and Wang [42] proposed a novel algorithm to separate the unvoiced speech signals from the non-speech interfering signals. The voiced speech and periodic parts of interfering signals have been firstly removed. The interference became stationary and the noise energy has been estimated in unvoiced intervals utilizing the separated speech in adjacent voiced intervals. The SS is applied to create time-frequency segments in unvoiced intervals and the unvoiced segments are then grouped. The grouping of segments is based on the frequency characteristics of unvoiced segments by considering thresholding and Bayesian classification.

Kokkinakis *et al*. [43] described and evaluated the capabilities of SS to suppress the late reflections and compared to ideal reverberant masking (IRM) approach. Speech intelligibility outcomes indicated that SS approach can suppress additive reverberant energy to a degree similar to that attained by the IRM. Hu and Yu [44] proposed an adaptive noise spectral estimator to deal with subtraction-based techniques for speech enhancement. The proposed method derived the noise spectrum from a primary estimate of noise spectrum together with the current noisy speech spectrum in an adaptive style. The fundamental framework of SS remained uninterrupted even in case of the gain for all spectral components is altered. The listening tests confirmed the superiority of the noise adaptation technique in suppressing the musical noise artifact and quality improvements.

B. Statistical Model-based Speech Enhancement Algorithms

In the statistical model based speech enhancement algorithms, speech and noise signals are assumed stationary and the resultant filter coefficients remain unchanged. The suppression of noise signals could effortlessly be realized utilizing Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filters. However, noise sources and particularly the speech signals are highly non-stationary. The speech generation trails a time-varying process. By using the noisy spectrum $Z(\omega,k)$, the short-time noise power spectral density (PSD) and the frequency-domain signal-to-noise ratio (SNR) are quantified to determine the weighting gains. The actual spectral weighting is achieved by multiplying the noisy spectrum $Z(\omega,k)$ by weighting gains $G(\omega,k)$ resulting in quantifying DFT coefficients of underlying clean speech according to the following equation:

$$\hat{S}(\omega,k)=G(\omega,k)Z(\omega,k) \tag{4}$$

The computation of the weighting gains rely on the particular speech enhancement algorithms and is usually a function of short-term noise PSD estimate $P_D^2(\omega,k)$ and the SNR estimates $\gamma(\omega,k)$ and $\xi(\omega,k)$ as:

$$\gamma(\omega,k)=\frac{|Z(\omega,k)|^2}{P_E^2(\omega,k)}, \quad \xi(\omega,k)=\frac{P_S^2(\omega,k)}{P_E^2(\omega,k)} \tag{5}$$

Where $\gamma(\omega,k)$ and $\xi(\omega,k)$ indicate *a posteriori* and *a priori* SNR estimate, $P_S^2(\omega,k)=E\{S(\omega,k)\}$ and $P_E^2(\omega,k)=E\{E(\omega,k)\}$ show variance of the clean speech and noise signals. $P_E^2(\omega,k)$ is calculated during the non-speech/ pauses-periods by using standard recursive equation, given as:

$$\hat{P}_E^2(\omega,k)=\beta\hat{P}_E^2(\omega,k-1)+(1-\beta)\hat{P}_Z^2(\omega,k-1) \tag{6}$$

Where, β is the smoothing factor and $\hat{P}_Z^2(\omega,k-1)$ is the noise estimate in the previous frame. The *a priori* SNR can be estimated by using Decision Direct (DD) [1] approach, given as:

$$\xi(\omega,k)=\alpha.\frac{\hat{X}^2(\omega,k-1)}{\hat{P}_D^2(\omega,k-1)}+(1-\alpha).\max\left[\frac{Y^2(\omega,k)}{\hat{P}_D^2(\omega,k)}-1,0\right] \tag{7}$$

Where, α is weighting parameter, $\hat{X}^2(\omega,k-1)$ and $\hat{P}_D^2(\omega,k-1)$ represent the power spectrum estimation of the clean speech and noise at *k-1* frame, respectively. In the following subsections distinguished and latest statistical speech enhancement algorithms based on the Wiener filtering (WF), minimum means square error (MMSE), Gaussian and super-Gaussian models are surveyed.

1. Wiener Filtering

Wiener filtering based speech enhancement minimizes the mean square error (MSE) between the estimated speech magnitude spectrum and the original signal magnitude spectrum. The formulation of the optimal wiener filter gain is as follows: [45]

$$G(\omega,k)=\frac{\xi(\omega,k)}{\xi(\omega,k)+1} \tag{8}$$

Over the years, Wiener filtering and its variants are used for the speech enhancement task. We discuss and review some of the highly ranked research studies on WF algorithms.

Huijun *et al* [46] proposed a SCSE algorithm which exploited connections between various time-frames to minimize residual noise. Contrasting to the traditional speech enhancement methods that apply a post-processor after standard algorithms like spectral subtraction, the proposed method applied a hybrid Wiener spectrogram filter (HWSF)

to reduce noise, trailed by a multi-blade post-processor that exploited two-dimension features of the spectrograms to retain the speech quality and to further reduced the residual noise. Spectrograms comparison showed that the proposed method significantly reduced the musical noise distortions. The usefulness of the proposed method is additionally confirmed by the use of objective assessments and unceremonious subjective listening tests.

Jahangir and Douglas [47] proposed a frequency-domain optimal linear estimator with perceptual post-filtering. The proposed method incorporated the masking properties of human hearing system to make the residual noise inaudible. A modified way is presented to quantify the tonality coefficients and relative threshold offsets for the best possible estimation of noise masking threshold. The proposed speech enhancement method has been evaluated for noise reduction and speech quality under many noisy conditions and yielded better results than [1].

Almajai and Milner [48] examined the visual speech information to enhance the noisy speech. The visual and audio speech features are analyzed which identified a pair with the highest audio-visual connection. The research revealed that high audio-visual connections exist inside individual phoneme rather entire speech. This connection is used in the application of a visually-driven Wiener filtering, which achieved clean speech and noise power spectrum statistics from the visual features. Clean speech statistics are quantified from the visual features using a maximum *a posteriori* structure and is incorporated inside the states of hidden Markov network to afford phoneme localization. Noise statistics are achieved by using a novel audio-visual voice activity detector, which used visual speech features to formulate the robust speech/nonspeech classifications. The efficiency of the proposed method is evaluated subjectively and objectively which confirmed the superiority.

Marwa *et al.* [49] presented adaptive Wiener filtering approach for speech enhancement. The proposed approach depended on the adaptation of the filter transfer function from sample-to-sample based speech signal statistics (the local mean and variance). The method is implemented in the time-domain to contain time-varying nature of the speech. The approach is evaluated against conventional frequency domain spectral subtraction, wavelet denoising methods and Wiener filtering using different speech quality metrics. The results showed superiority of the proposed Wiener filtering method.

Xia and Bao [50] proposed a Weighted Denoising Auto-encoder (WDA) and noise classification based speech enhancement approach. Weighted reconstruction loss function is established into standard Denoising Auto-encoder (DAE) and link between the power spectrums of underlying clean and noisy speech is expressed by WDA structure. The sub-band power spectrums of underlying clean speech are quantified using the WDA structure from the noisy speech. The *a priori* SNR is quantified using *a Posteriori* SNR Controlled Recursive Averaging (PCRA) approach. The enhanced speech is achieved by the Wiener filter in the frequency-domain. Moreover, GMM-based noise classification method is engaged to make the proposed method appropriate for various conditions. The experimental results demonstrated that the proposed method achieved improved objective speech quality. Effective noise reduction and SNR improvements are attained with less speech distortion.

Kristian and Marc [51] investigated speech-distortion weighted inter-frame Wiener filters for the SCSE in a filterbank configuration. The filterbank configuration utilized a regularization parameter as a tradeoff between speech distortion and noise reduction. The method depends on the quantification of inter-frame correlation coefficients, and it is shown that these coefficients could be robustly estimated using a secondary higher resolution filterbank. It is then demonstrated that real-valued scalar gains can be applied directly in higher resolution filterbank rather than inter-frame filtering in the primary filterbank, which leads to a robust noise reduction performance for any value of regularization parameter.

2. MMSE Estimators

The minimum means square error (MMSE) estimator [1] inheres to vital class of the estimators and quantifies the spectral magnitudes. The MMSE estimator reduces the quadratic error of the spectral speech amplitudes according to the following equation:

$$E\left\{(S(\omega,k)-\hat{S}(\omega,k))^2\right\} \rightarrow Min \tag{9}$$

Considering the Gaussian model of the speech and noise, the final weighting rule is given according to [1] as:

$$\hat{S}(\omega,k) = E\left\{S(\omega,k)/Z(\omega,k)\right\} \tag{10}$$

$$\hat{S}(\omega,k) = \frac{\sqrt{\nu(\omega,k)}}{\lambda(\omega,k)}\Gamma(1.5)\,F_1(-0.5,1,\nu).Z(\omega,k)$$
$$\nu(\omega,k) = \frac{\xi(\omega,k)}{\xi(\omega,k)+1}.\gamma(\omega,k) \tag{11}$$

Γ(.) and F1(.) shows Gamma function and Hypergeometric function, respectively. We discuss and review some of the highly ranked research studies on MMSE algorithms.

Basheera *et al.,* [52] proposed novel optimum linear and nonlinear estimators. They are derived based on the MMSE sense to reduce the distortion in original speech. Linear and nonlinear bilateral Laplacian gain estimators are proposed. The observed signal is first decorrelated through a real transform to achieve its moment coefficients and then applied to the estimated speech signal in the decorrelated domain. The mathematical aspect of MSE of estimators is evaluated suggesting significant improvement. Kandagatla and Subbaiah [53] derived joint MMSE estimation of speech coefficients provided phase uncertainty by assuming the speech coefficients. Uncertain phase is used for amplitude estimation. Furthermore new Phase-blind estimators are designed utilizing the Nagakami power spectral density function and the generalized Gamma for speech and noise priors.

Hamid *et al*. [54] addressed the problem of speech enhancement using β-order MMSE-STSA. The advantages of the Laplacian speech modeling and β-order cost function are taken in MMSE estimation. An investigative solution is presented for the β-order MMSE-STSA estimator deeming Laplacian priors for DFT coefficients of the clean speech. A Gaussian distribution for the real and imaginary parts of the DFT coefficients of the noise is presupposed. Using estimates for the joint PDF and the Bessel function, a better closed-form adaptation of the estimator is also presented.

Gerkmann and Krawczyk [55] derived a MMSE optimal estimator for underlying clean speech spectral amplitude. It is shown that the phase contains extra information which can be used to differentiate outliers in the noise from the target signals. *Matthew and Bernard* in [56] proposed a Bayesian STSA stochastic deterministic speech model, which included *a priori* information by utilizing a non-zero mean. For the speech STFT magnitude, investigative expressions are derived in the MMSE principle whereas phase in maximum-likelihood principle. An approach for quantifying *a priori* stochastic deterministic speech model parameters is explained based on the harmonically related sinusoidal parts in the STFT frames and deviations in magnitude and phase of components between succeeding STFT frames.

C. Signal Subspace-based Speech Enhancement Algorithms

Signal subspace [57] [58] based SE approaches use KLT, SVD and EVD to disintegrate noisy speech signals into the noise plus

signal subspace known as the signal-subspace, whereas eliminates the noise signal that falls within orthogonal noise-subspace. The signal-subspaces are processed separately to remove noise components utilizing a diagonal gain matrix based on uncorrelated components in subspace. The components of the gain matrix are quantified by time-domain or spectral-domain estimators. The covariance matrix R_Z of the noisy speech can be written as:

$$R_Z = R_s + R_E \tag{12}$$

R_s and R_E are the covariance matrices of the clean speech and noise signals. R_Z is supposed to have a higher rank than R_s. The EVD of the covariance matrices is given as:

$$R_S = V \Lambda V^T \tag{13}$$

$$R_D = V(\sigma_W^2 I)V^T \tag{14}$$

$$R_Y = V(\Lambda + \sigma_W^2 I)V^T \tag{15}$$

Λ indicates a diagonal matrix that contains the Eigen-values, V indicates an orthonormal matrix containing eigenvectors; σ shows variance of noise whereas I indicate identity matrix. Speech enhancement process is represented by a filtering operation input speech vector as:

$$z = \{z(1), z(2), y(3), \ldots . , z(n)\}^T \tag{16}$$

$$\hat{S} = \Psi Z \tag{17}$$

The term Ψ is the filtering matrix, given by equation (18) as:

$$\Psi = V_P G_P V_P^T \tag{18}$$

Where G_p holds weighted Eigen values of R_Z, and V_p and V_p^T shows KLT and its inverse matrices, respectively. We discuss and review highly ranked research studies on SigSub algorithms.

Borowicz and Petrovsky [59] examined speech enhancement methods based on the perceptually motivated signal subspace. Lagrange multipliers are used to modify the spectral-domain-constrained (SDC) estimator. The residual noise power spectrums are shaped with an algorithm for accurate computing the Lagrange multipliers. The proposed approach uses masking phenomena for residual noise shaping and is optimal for the case of colored noise. Results show that the proposed method outperformed the competing methods and provided high noise reduction and improved speech quality.

Mohammad *et al.* [60] proposed a non-unitary spectral transformation of the residual noise based on diagonalization of covariance matrices associated to the clean speech and noise signals. Through this transformation, the optimization problem is solvable devoid of any constraints on the structure of contributed matrices.

Vera [61] pointed out that estimation of the dimension of signal subspace is critical and depends on the noise variance as well as SNR. Both fluctuate along temporal segments of speech and frequency bands. It is anticipated to work over frames in all critical bands utilizing the threshold noise variance. Belhedi *et al* [62] used soft mask as a core in the proposed approach. The method produces two separate signals of dissimilar qualities and made them available in two separate channels. The classification of the channels is made via Fuzzy logic that needs two separate parameters. One parameter determines quality and intelligibility whereas the second parameter determines the gender of the speaker via F_0 tracking method. The proposed approach achieved an average 59.5% improvement in SIR, 67.9% progress in PESQ, and 10.5% improvement in TPS.

Sudeep and Kishore [63] proposed a perceptual subspace approach via masking properties of the human auditory system with variance normalization to decide the gain parameters. An estimator is used to determine the filter coefficients. The noise is handled by substituting the noise variance by Rayleigh quotient. Normalization of variance is made by removing the spikes to evade rapid increase or decrease in power of the output samples making the output more intelligible.

D. Computational Auditory Scene Analysis-based Speech Enhancement Algorithms

The field of computational study intends to achieve human performance in the Auditory Scene Analysis (ASA) by using single microphone recordings of the acoustic prospect. This definition describes the biological relevance of the field by limiting the microphone number to two and its functional goal of Computational Auditory Scene Analysis (CASA). The CASA uses perceptually motivated mechanisms. Over the years, CASA based methods are used for the speech enhancement; here we are reviewing some of the work in recent years.

A new ideal ratio mask (IRM) depiction is proposed by Bao and Abdulla in [64] by utilizing inter-channel correlation. The power ratio of the speech and noise during the structuring of ratio mask is adaptively reallocated; therefore more speech components are held and noise components are masked simultaneously. Channel-weight contour is assumed to modify the mask in all Gammatone filterbank channels.

Wang *et al.* [65] proposed IRM estimation that relies on the spectral dependency into the speech cochleagram to enhance noisy speech. A data field representation is established to design time-frequency connection of the cochleagram with adjacent spectral information to estimate IRM. Firstly, a pre-processed section is used to achieve initial time-frequency values of noise and speech. Then the data field model is used to obtain the forms of speech and noise potentials. Subsequently, the optimal potentials that reveal their respective optimal distribution are achieved by the optimal influence factors. Lastly, masking values are obtained via the potentials of the speech and noise for reinstating the clean speech signals.

Wang *et al.* [66] considered a novel approach of speech and noise models, and presented two model-based soft decision methods. A ratio mask is computed by the exact Bayesian estimators of speech and noise. Additionally, a probabilistic mask is estimated with a variable local criterion. Liang *et al.* [67] considered local correlation knowledge from two aspects for improved performance. The time-frequency segmentation-based potential function is derived to represent the local correlation between mask labels of neighboring units directly. It is demonstrated that time-frequency unit that belongs to one segment is mostly dominated by one source. Alternatively, a local noise level tracking phase is integrated. The local level is attained by averaging many neighboring time-frequency units and is considered as a method for accurate noise energy. It is utilized as an intermediary auxiliary variable to signify the correlation. A high dimensional posterior distribution is simulated by a Markov Chain Monte Carlo (MCMC) approach. During iterations, the correlation is fully utilized to quantify the acceptance ratio. The estimated ideal binary mask (IBM) is achieved using the expectation operator. The proposed approach is compared and evaluated with a Bayesian approach and the approach yielded considerably large performance gain in terms of SNR gain and HIT-FA rates.

Narayanan and Wang [68] presented a system for robust SNR estimation based on CASA. The proposed method used an estimate of the IBM to separate a time–frequency illustration of the noisy speech signal into speech and noise dominated sections. Energy inside each region was totaled to gain the filtered global SNR. SNR transformation was established to translate the estimated SNR to the true global SNR of the noisy speech signal.

Hu and Wang [69] proposed a tandem algorithm to estimate the pitch of a target speech utterance and separated the voiced regions

of the target speech. First, a coarse estimate of the target pitch was obtained and then the estimate is used to segregate target speech using harmonicity and temporal continuity. Lee and Kwon [70] proposed a CASA-based speech separation system and matched the missing speech parts by using the shape analysis method.

May and Dau [71] presented a method based on the estimate of the ideal binary mask from noisy speech in supervised learning of AMS features and auditory inspired modulation filterbanks with logarithmically scaled filters were used. Spectro-temporal integration stage was incorporated to obtain speech activity information in neighboring time-frequency units.

E. Empirical Mode Decommission-based Speech Enhancement Algorithms

Empirical Mode Decomposition (EMD) [72] directly extracts the energy related to different intrinsic time scales. EMD is an adaptive approach and follows some necessary steps to decompose nonlinear and nonstationary data. (**i**) First, the EMD obtains the local maxima and minima. (**ii**) Secondly, the EMD finds the local maximum and local minimum envelopes. (**iii**) Third, the EMD finds the mean of the obtained local extrema envelopes and finally subtracts this mean envelope from the input data to attain the residual intrinsic mode function (IMF).

Upadhyay and Pachori [73] proposed a novel speech enhancement method for suppressing stationary and non-stationary noise sources. The variational mode decomposition (VMD) and EMD approaches are combined to develop the new idea for speech enhancement. Firstly, the EMD decomposes the input noisy speech into the IMFs. The VMD is then applied on the summation of preferred IMFs. The Hurst exponent was used to select the IMFs. The proposed speech enhancement method reduced low and high-frequency noise sources and showed enhanced speech quality.

Khaldi *et al.* [74] presented a speech enhancement method that exploited the combined effects of EMD and the local statistics of the speech signal by utilizing the adaptive centre weighted averaging filter. The speech signals were segmented into frames and all frames were segmented down by EMD into IMFs. The filtered IMFs depend on the voiced or unvoiced frame. An energy norm was utilized to classify the voiced frames and a stationarity index was used between unvoiced and transient chain. Zao *et al.*, [75] proposed a speech enhancement scheme based on the adoption of Hurst exponent during the selection of IMFs to reconstruct the target speech.

Hamid *et al.*, [76] proposed a novel data adaptive thresholding approach. The noisy speech signals and fractional Gaussian noises were mixed to generate the complex noisy signal. Bivariate EMD was used to decompose the complex noisy signal into complex-valued IMFs and all IMFs were segmented into short-time frames for processing. The variances of the IMFs of fractional Gaussian noise computed inside the frames were used as the reference to categorize subsequent frames of noisy speech into signal-dominant and noise-dominant frames, respectively. A soft thresholding method is used at noise-dominant frames to decrease the effects of noise. Every frame and IMF of the speech signals were combined to yield the enhanced speech signal.

Chatlani and Soraghan [77] used the EMD as a post-processing stage for filtering low frequency noise. An adaptive approach was designed to choose IMF index for sorting out the noise component from speech components. This separation was carried out by using a second-order IMF statistics. The low-frequency noise components were removed by the biased reconstruction from the IMFs. Khaldi *et al.*, [78] used EMD for fully data-driven based approaches for noise reduction. Noisy speech signal was decomposed adaptively into IMFs using sifting process. The signal reconstruction with IMFs was done using the MMSE filter and thresholded using a shrinkage function.

The U-SCSE algorithms provide acceptable speech quality and noise reduction in many real-world noise sources. The U-SCSE algorithms along with several advantages also came with some limitations. The TableI and Table II provides advantages and limitations of various U-SCSE algorithms. These limitations will point out several

TABLE I. Problem Statements, Methodologies, Contributions and Limitations of U-SCSE Algorithms

Method	Problem Statement	Methodology	Contribution	Limitation
GA-SS [34]	Speech enhancement to improve speech quality and to reduce the musical noise distortion.	Compute the magnitude spectrum of the noisy signal using the FFT. The noise spectrum is updated using noise estimators. The gain is estimated using modified gain and multiplied with noisy spectrum to enhance speech.	Performed significantly better than the traditional spectral subtraction algorithm in terms of speech quality and musical noise artifact.	Speech intelligibility is not evaluated. Additionally informal tests were conducted for evaluations. Noise reduction impact on speech intelligibility research is required.
MOD-SS [36]	Speech enhancement to improve speech quality and intelligibility in Modulation domain.	The SE method used AMS-based modulation domain. Each frequency component of the acoustic magnitude spectra is processed frame-wise across time using a modulation AMS framework, and the enhanced modulation spectrum is computed.	New Speech enhancement domain in terms of SS is explored. Better speech quality and speech intelligibility is obtained. Better noise reduction is offered.	Although the proposed method offered better results, the combination with other domains produces complexity in the proposed method. The complexity of the method is not discussed.
MOD-SS [37]	Speech enhancement to improve speech quality in Modulation domain.	The magnitude subtraction is adopted and extended into the modulation frequency domain for the separate enhancements of the real and imaginary spectra. The noise is estimated in real and imaginary spectra and the estimated speech is recreated.	Perform subtraction on the real and imaginary spectra separately in the modulation frequency domain. Better noise reduction and speech quality is achieved.	The speech intelligibility potential of the proposed method is not discussed. The method estimated the phase, thus the complexity of the method is not discussed.
VAD-SS [39]	The Speech enhancement for better results and musical noise reduction in the Kurtosis of noise spectra.	Iteratively weak-nonlinear method is used to obtain quality speech with less musical artifact. The generation of musical artifact is formulated by marking changes in kurtosis of the noise spectrum. Optimal internal parameters are derived theoretically to produce no musical artifact in kurtosis.	The proposed method provided better results and generation of musical noise artifact is formulated in the Kurtosis of noise spectra.	No theoretical explanation is given, only experimental results are presented. Speech quality and intelligibility is not discussed.

TABLE II. Problem Statements, Methodologies, Contributions and Limitations of U-SCSE Algorithms

Method	Problem Statement	Methodology	Contribution	Limitation
SDW-IFWF [51]	Speech enhancement for better quality and to reduce the musical noise distortion	Speech-distortion weighted inter frame Wiener filters for noise reduction is implemented in a filter bank structure. The filters utilized a regularization parameter as a tradeoff between speech distortion and noise reduction. The method depends on the estimation of inter frame correlation coefficients and these coefficients are more robustly estimated using a secondary higher resolution filter bank.	The contribution of the paper is the implementation of the scalar SDW-IFWF gain in a HRFB, matching a principle in the crucial lower-resolution filter bank to improve the speech quality and noise reduction with less musical artifact	The algorithm provided improved results in terms of the speech quality However, speech intelligibility potential of the proposed algorithm is not discussed and evaluated.
LBLG-NBLG [52]	Speech enhancement for better quality speech and low speech Distortion	The estimators are derived on the basis of MMSE to reduce the distortion of the fundamental speech. The musical artifact is reduced without affecting the noise reduction. LBLG and NBLG estimator are proposed. The input signal is decorrelated to obtain moment coefficients. The estimators are applied to estimate the clean signal in the decorrelated domain. The original signal is obtained in time domain.	The proposed method obtained better speech quality and noise reduction. Non-linear and linear bilateral Laplacian estimators are derived to improve the speech quality.	Although method produced better speech quality as compare to traditional methods; however, the speech intelligibility and complexity potentials are not fully explored.
EPW-Sub [60]	Speech Separation in optimized subspace for improved quality and intelligibility.	The separation is achieved by optimizing the subspace via decomposing the mixture signal into three subspaces: sparse, sub-sparse and low-rank subspaces. Soft masking is used for the final verdict. Two signals of different qualities are provided in two separate channels. The channel classification is made by using Fuzzy logics with two parameters. F0 tracking algorithm is proposed to classify gender.	Embedded pre-whitening subspace method is proposed based on controlled spectral-domain for better speech quality and noise reduction in colored noises.	Although the proposed method offered better results but the speech intelligibility in non-stationary noise sources is not discussed.
CASA-SE [65]	The Speech enhancement for improved quality and intelligibility in the data driven field of cochleagram.	Iteratively weak-nonlinear method is used to obtain quality speech with less musical artifact. The generation of musical artifact is formulated by marking changes in kurtosis of the noise spectrum. Optimal internal parameters are derived theoretically to produce no musical artifact in kurtosis	Ideal Ratio Mask is estimated in the data driven field of cochleagram to enhance the noisy speech. The proposed method obtained considerable gain in speech quality. Better results in terms of energy loss and residue noise are contributed.	The proposed algorithm has not incorporated the DF model into the STFT domain. The complexity of the algorithm is not discussed.

research areas which need further research.

III. Speech Intelligibility and Quality Potential of Various U-Scse Algorithms

The Table I-II illustrates the problem statements, methodologies, contributions and limitations of U-SCSE algorithms. It is clear from Table I-II that the U-SCSE algorithms addressed the problem of the speech enhancement effectively for noise reduction, musical noise artifact and speech quality. Speech enhancement is usually used as the front-end to Automatic speech recognition systems where speech intelligibility is the more important attribute. It is observed from the survey of the above different classes that speech intelligibility attributes is not fully explored in most of the U-SCSE algorithms. This section provides an intense experimental evaluation to observe the quality and intelligibility potentials of the U-SCSE approaches.

A. Methods

The experiments represent the measures used to evaluate and validate the performance of speech enhancement algorithms. In experiments, the U-SCSE algorithms are evaluated by using a set of 60 noisy speech sentences belonging to female and male speakers in terms of the speech intelligibility and quality. The noisy stimuli are generated by adding four real-time background noises to the clean speech utterances at several signal-to-noise ratios (SNR). The clean speech sentences are selected from the standard IEEE database [85] randomly. Four nonstationary noise sources (street, exhibition hall, airport, and multitalker babble noise) are chosen from the Aurora database [86]. The speech utterances are mixed at four SNR from 0dB to 15dB, spacing 5dB applying the ITU-T P.51. The sampling rate is fixed at 8 kHz. Five classes of U-SCSE algorithms are included in the experiments performed for speech quality and intelligibility. The U-SCSE classes include Spectral Subtraction (SS), Wiener Filtering (WF), Minimum Mean Square Error (MMSE) estimators, Signal Subspace (SigSub) and EMD type. Table III provides the details of speech enhancement algorithms used in the experiments. Two evaluation measures are quantified in order to access the U-SCSE algorithms. The PESQ [87] is preferred for the speech quality; an ITU-T P.862 standard that substituted the obsolete ITU-T P.861 standard because of inadequate performance to evaluate the speech enhancement. The PESQ score follows the range of −0.5 and 4.5, but, during experiments the score follows the mean opinion score (MOS), that is, a range of 1.0 to 4.5. The PESQ scores are calculated using the following equation:

$$PESQ = \eta 0 + \eta 1.DSYM + \eta 2.DASYM \quad (18)$$

Where $\eta 0 = 4.5$, $\eta 1 = -0.1$ and $\eta 2 = -0.039$.

TABLE III. List of U-SCSE Algorithms in Experiments

S. No	Speech Enhancement Class	Speech Enhancement Algorithm
1	Spectral Subtractive (SS)	SS [79] SS-RDC [80] MBSS [81]
2	Wiener Filtering (WF)	WF [45] WWF [82]
3	Minimum Mean Square Estimation (MMSE)	MMSE-SPU[1] LMMSE [2]
4	Signal Subspace (SigSub)	KLT [83] PKLT [84]

TABLE IV. PESQ Analysis of U-SCSE Algorithms

Noise Type	SNR (dB)	Spectral Subtractive			Wiener Type		Statistical-Model		Signal Subspace		EMD
		SS	RDC	MBSS	WF	WWF	MMSE	LMMSE	KLT	PKLT	H-EMD
Airport	0dB	1.59	1.69	1.81	1.92	1.18	1.23	1.95	1.78	1.51	1.84
	5dB	2.03	2.16	2.20	2.12	2.03	1.43	2.12	2.13	2.02	2.01
	10dB	2.39	2.35	2.54	2.43	2.27	1.54	2.45	2.29	2.08	2.63
	15dB	2.95	2.74	3.12	3.05	2.62	1.65	3.03	2.79	2.42	2.93
Babble	0dB	1.45	1.68	1.98	1.78	1.16	1.26	1.92	1.34	1.34	1.91
	5dB	2.07	2.16	2.28	2.12	2.13	1.53	2.12	2.11	1.98	2.19
	10dB	2.42	2.36	2.59	2.46	2.34	1.67	2.53	2.37	2.25	2.72
	15dB	2.60	2.61	2.75	2.67	2.55	1.85	2.71	2.61	2.51	2.88
Exhibition Hall	0dB	1.25	1.49	1.43	1.69	1.33	1.33	1.72	1.37	1.63	1.77
	5dB	1.87	1.91	2.01	2.01	1.81	1.61	1.95	1.89	1.50	1.93
	10dB	2.47	2.14	2.44	2.40	2.39	1.79	2.46	2.44	2.28	2.57
	15dB	2.82	2.46	2.82	2.78	2.65	1.95	2.79	2.86	2.50	2.89
Street	0dB	1.49	1.51	1.54	1.60	1.53	1.53	1.72	1.59	1.55	1.79
	5dB	2.05	1.98	2.14	2.06	2.08	1.88	2.04	2.12	2.12	2.14
	10dB	2.49	2.36	2.61	2.60	2.33	2.03	2.52	2.32	2.14	2.63
	15dB	2.92	2.53	2.89	2.74	2.65	2.25	2.77	2.84	2.55	2.92

TABLE V. Across-Class Comparative Analysis of U-SCSE Algorithms in terms of PESQ

Noise Type	SNR (dB)	Spectral Subtraction			Wiener Filtering		MMSE Estimation		Signal Subspace		EMD
		SS	RDC	MSS	WF	WWF	MMSE	LMMSE	KLT	PKLT	H-EMD
Airport	15dB			*			*	*			
	10dB			*			*	*			*
Babble	15dB			*	*		*	*			*
	10dB			*	*		*	*			*
Exhibition Hall	15dB			*	*		*	*			*
	10dB			*	*		*	*			*
Street	15dB			*	*		*	*			*
	10dB			*			*	*			

TABLE VI. Across-Class Comparative Analysis of U-SCSE Algorithms in terms of STOI

Noise Type	SNR (dB)	Spectral Subtraction			Wiener Filtering		MMSE Estimation		Signal Subspace		EMD
		SS	RDC	MSS	WF	WWF	MMSE	LMMSE	KLT	PKLT	H-EMD
Airport	15dB	*		*			*	*		*	
	10dB			*			*	*			*
Babble	15dB		*	*	*	*	*	*	*	*	*
	10dB			*	*		*	*			*
Exhibition Hall	15dB	*	*	*	*		*	*	*		*
	10dB			*	*		*	*			*
Street	15dB	*		*	*	*	*	*	*		*
	10dB			*			*	*			

Note: Algorithms specified by asterisks sign executed equally well whereas algorithms without asterisks sign executed poorly.

5	Empirical Mode Decomposition (EMD)	H-EMD [75]

A separate evaluation metric is used to access the intelligibility of the enhanced speech. The short-time speech intelligibility (STOI) [88] is considered for this purpose. The STOI sores are calculated by the equation given as:

$$\mathrm{f(STOI)} = \frac{100}{1+\exp(a\mathrm{STOI}+b)} \tag{19}$$

The parameters a, b are set according to [8], $a = -17.4906$ and b = 9.6921.

B. Results and Discussion

A performance comparison analysis at two levels is presented in this section. First, within-class performance comparison of the U-SCSE algorithms is established. The five classes are Spectral Subtractive, Statistical-models, Wiener-Filtering type, Subspace and EMD-type. This performance comparison was conducted to observe the significant performance differences within-class algorithms. Secondly, across-classes performance comparison is conducted to evaluate and find the algorithm(s) that performed better in all noisy situations.

1. Within-Class Algorithm Comparison

Table IV provides the results for PESQ (speech quality) whereas average speech intelligibility results are demonstrated in Fig. 4. Of three tested spectral-subtractive algorithms, the multi-band spectral subtraction (MBSS) [81] performed constantly the best across all noisy situations in terms of the speech quality. The MBSS and SS-RDC [80] methods performed equivalently well excluding 0dB exhibition hall

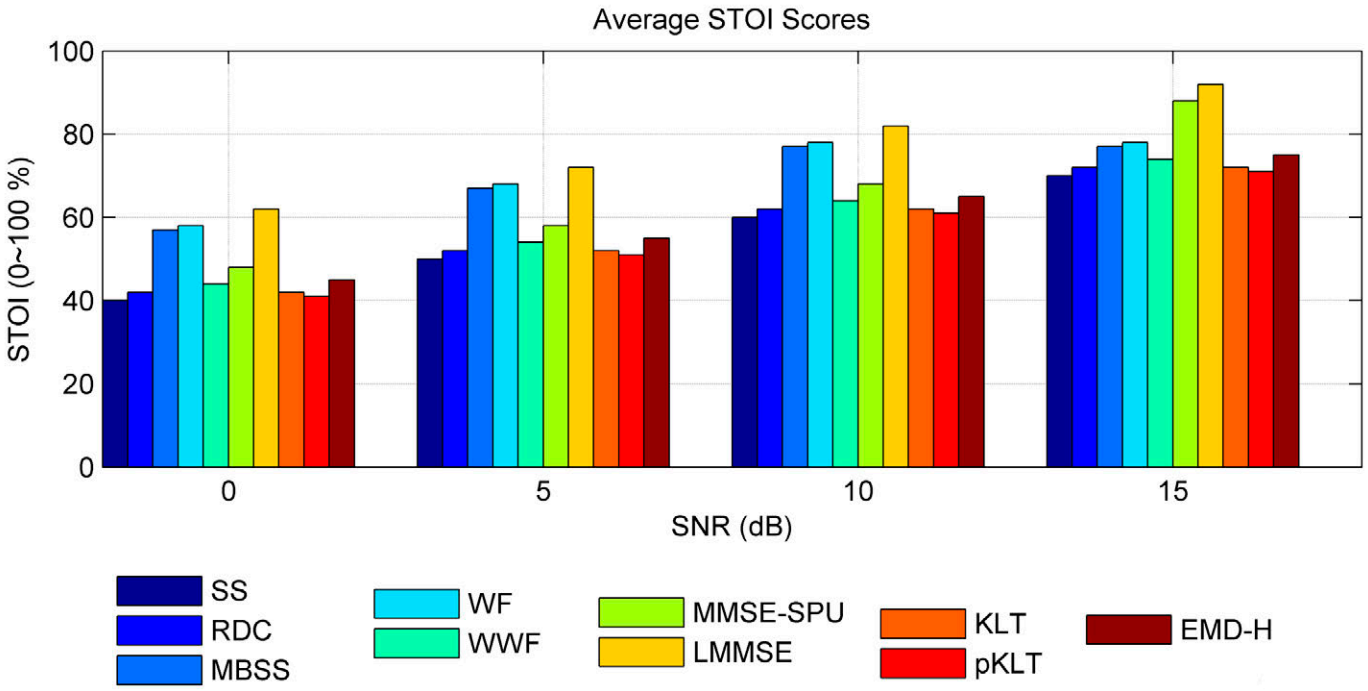

Fig. 4. Average Speech Intelligibility prediction for U-SCSE algorithms in terms of STOI.

noise and 0dB street noise conditions. Noise distortion of SS-RDC algorithm was considerably less than the MBSS and SS [79] approaches in all noisy situations. In terms of speech intelligibility, the MBSS and SS-RDC approaches equally performed in most of the noisy situations excluding 0dB exhibition hall noise and 0dB street noisy situations, where MBSS algorithm performed notably superior and presented less speech distortion. In brief, MBSS performed better than SS-RDC and SS, providing better overall speech intelligibility and quality. For speech quality, the two subspace approaches performed equally for the most of SNRs and noise types, excluding 0 dB babble noise.

The two Wiener-type algorithms performed well for most SNR conditions and four types of noise except 0dB airport noise and 0dB babble noise. For speech quality, the WF [45] performed significantly better than the WWF [82] approach at all SNRs and noise sources. WWF performed poorly in all noise sources at almost all SNRs and significant residual noise is experienced in the enhanced speech. On the other hand WF-as offered better speech quality and the noise reduction capabilities were significant. For speech intelligibility, the WF-as performed well at all SNRs and noise sources as compared to WWF method. There is significant speech distortion observed in the output speech utterance of the WWF approach.

The two statistical-model based approaches performed good for most of SNRs and noise types. The log-MMSE (LMMSE) [2] performed significantly better than the MMSE-SPU [1] approach at all SNRs and noise sources. MMSE performed poorly in all noise sources at almost all SNRs, and significant residual noise observed in the enhanced speech. On the other hand LMMSE offered better speech quality and noise reduction capabilities were significant. For speech intelligibility, the MMSE-SPU performed very poorly at all SNRs and noise sources. The small speech intelligibility signifies the higher speech distortion offered by MMSE-SPU. LMMSE offered better speech intelligibility and comparatively less speech distortion is experienced in the output speech.

The generalized subspace approach, KLT [83] performed significantly better than the pKLT [84] approach at all SNRs and noise sources except 0dB exhibition hall noise. The KLT approach was more successful in suppressing the background noise and perceptual speech quality. In terms of speech intelligibility, KLT and pKLT approaches performed equally well at all SNRs and noise sources except 0dB exhibition hall noise. There is no significant improvement in speech intelligibility observed for pKLT approach. On the other hand, KLT improved speech intelligibility marginally.

In terms of the speech quality, the EMD-H [75] algorithm performed well for all SNRs and noise types, except at 0dB exhibition hall noise and 0dB street noise. The EMD-H was successful in suppressing the background noise and improving the perceptual quality and speech intelligibility at all SNRs and the noise sources.

2. Across-Class Algorithm Comparison

Table V-VI indicates the results achieved by using ANOVA statistical analysis for the speech quality and intelligibility. Asterisk sign in Table V-VI show lack of statistical significant difference between algorithms with the utmost scores and the denoted algorithms. The U-SCSE algorithms marked by the Asterisk sign in Table V performed similarly. Table V indicates no single algorithm is categorized as the best, and several speech enhancement algorithms performed equally well across SNRs situations and noise types. In terms of the speech quality, MMSE-SPU, LMMSE, WF, EMD-H and MBSS performed equally well across all SNRs situations. Table VI indicates the results achieved from the ANOVA statistical analysis for speech intelligibility. The MMSE-SPU, LMMSE, MBSS and WF performed well. All algorithms produced low speech distortion (high intelligibility) across all SNRs situations and noise sources. KLT, SS-RDC and WWF algorithms also performed well in isolated SNR situations.

IV. Conclusion

This paper presented a comprehensive review of the different classes of the single-channel speech enhancement algorithms in unsupervised perspective in order to improve the intelligibility and quality of the contaminated speech. Various classes of the unsupervised speech enhancement approaches for enhancing the noisy speech have been discussed. We have summarized possible algorithms of the Spectral Subtraction (SS), Wiener Filtering (WF), Minimum Mean

Square Error (MMSE) estimators, Signal Subspace (SigSub) and EMD type, explained state-of-the-art approaches and a many related studies have been reviewed. The review suggested that unsupervised speech enhancement methods show an acceptable speech quality but speech intelligibility potential remains medium. The algorithms of unsupervised class show better noise reduction however; decrease of the residual noise artifact and speech distortion requires further research. Different unsupervised speech enhancement approaches have distinctive advantages that make these algorithms appropriate for speech enhancement; in contrast, these algorithms have some serious limitations as well. Table I-II summarized the problem statements, methodologies, contributions and the limitations of many speech enhancement algorithms. On the basis of the limitations extracted from the reviewed papers and also from the experimental results, it is concluded that unsupervised speech enhancement improves the speech quality but the speech intelligibility improvement potential requires further research. The algorithm can use the noise estimators, but accurate estimate is also a difficult task. A too aggressive estimation may lose important speech contents which in turn affect the speech intelligibility whereas too low noise estimation may lead to the residual noise. We have outlined various problems that need research to design robust single-channel speech enhancement algorithms. This rapid progress in the unsupervised speech enhancement algorithms will possibly persist in the future. To conclude, some following open research problems are outlined that are extracted from research studies:

1. Generalization to the Nonstationary Noise Sources: Although U-SCSE algorithms provide promising speech quality results in stationary noise sources, however, their performance in nonstationary noise sources is not high. Effective noise estimation must be integrated with U-SCSE algorithms for better speech quality and noise reduction results.

2. Speech Intelligibility in Nonstationary Noise Sources: U-SCSE provides enhanced speech with very low speech intelligibility. More effective algorithms are required that can improve speech intelligibility in nonstationary noise sources.

3. Musical Noise Artifact and Speech Distortion: Unsupervised speech enhancement algorithms provide acceptable noise reduction, however reduction of the residual noise artifact and speech distortion requires further research.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[2] E. Yariv and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443-445, 1985.

[3] E. Yariv and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing,* vol. 3, no. 4, pp. 251-266, 1995.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, vol. 9, no.1, pp. 12-15, 2002.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403-2418, 2001.

[6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126-137, 1999.

[7] N. Saleem, M. I. Khattak, G. Witjaksono, and G. Ahmad, "Variance based time-frequency mask estimation for unsupervised speech enhancement," *Multimedia Tools and Applications*, 1-25, 2019.

[8] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE signal processing letters*, vol. 8, no. 1, pp. 10-12, 2001.

[9] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 354850, 2005.

[10] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, 2001.

[11] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497-514, 1997.

[12] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845-856, 2005.

[13] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098-2108, 2006.

[14] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters,* vol. 9, no. 4, pp. 113-116, 2002.

[15] Y. Hu and P.C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio processing*, vol. 12, no. 1, pp. 59-67, 2004.

[16] J. H. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795-805, 1991.

[17] S. Watanabe, M. Delcroix, F. Metze, and J.R. Hershey, Eds., *New era for robust speech recognition: exploiting deep learning*, Springer, 2017.

[18] N. Saleem and T. G. Tareen, "Spectral Restoration based speech enhancement for robust speaker identification," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 1, pp. 34-39, 2018.

[19] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*, Morgan Kaufmann, 2017.

[20] N. Saleem, E. Mustafa, A. Nawaz, and A. Khan, "Ideal binary masking for reducing convolutive noise," *International Journal of Speech Technology*, vol. 18, no. 4, pp. 547-554, 2015.

[21] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, 2006.

[22] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[23] D. G. Jamieson, R. L. Brennan, and L. E. Cornelisse, "Evaluation of a speech enhancement strategy with normal-hearing and hearing-impaired listeners," *Ear and hearing*, vol. 16, no. 3, pp. 274-286, 1995.

[24] B. C. Moore, "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms," *Speech communication*, vol. 41, no. 1, pp. 81-91, 2003.

[25] K. H. Arehart, J. H. Hansen, S, Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," *Speech Communication*, vol. 40, no. 4, pp. 575-592, 2003.

[26] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07,* vol. 4, IEEE, 2007, pp. IV-561.

[27] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777-1786, 2007.

[28] S. Gordon-Salant, "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects," *The Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1199-1202, 1987.

[29] J. B. Allen, "How do humans process and recognize speech?," *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 567-577, 1994.

[30] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of rehabilitation research and development*," vol. 38, no. 1, pp. 111-122, 2001.

[31] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486-1494, 2009.

[32] G. Kim and P.C. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 8, pp. 2080-2090, 2010.

[33] S. Boll "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113-120, 1979.

[34] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech communication*, vol. 50, no. 6, pp.453-466, 2008.

[35] S. Nasir, A. Sher, K. Usman, and U. Farman, "Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 6, pp. 1081-1087, 2013.

[36] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450-475, 2010.

[37] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 6, pp. 1770-1779, 2010.

[38] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, no. 4, pp. 509-522, 2013.

[39] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080-2094, 2012.

[40] A. L. Ramos, S. Holm, S. Gudvangen, and R. Otterlei, "A spectral subtraction based algorithm for real-time noise cancellation with application to gunshot acoustics," *International Journal of Electronics and Telecommunications*, vol. 59, no. 1, pp. 93-98, 2013.

[41] S. M. Ban and H. S. Kim, "Weight-Space Viterbi Decoding Based Spectral Subtraction for Reverberant Speech Recognition," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1424-1428, 2015.

[42] K. Hu and D. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1600-1609, 2010.

[43] K. Kokkinakis, C. Runge, Q. Tahmina, and Y. Hu, "Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 115-124, 2015.

[44] H. T. Hu and C. Yu, "Adaptive noise spectral estimation for spectral subtraction speech enhancement," *IET Signal Processing*, vol. 1, no. 3, pp. 156-163, 2007.

[45] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[46] H. Ding, Y. Soon, S. N. Koh, and C. K. Yeo, "A spectral filtering method based on hybrid wiener filters for speech enhancement," *Speech Communication*, vol. 51, no. 3, pp. 259-267, 2009.

[47] M. J. Alam and D. O'Shaughnessy, "Perceptual improvement of Wiener filtering employing a post-filter," *Digital Signal Processing*, vol. 21, no. 1, pp. 54-65, 2011.

[48] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642-1651, 2010.

[49] M. A. A. El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E. S. M. El-Rabaie, W. Al-Nuaimy, ... and F. E. A. El-Samie, "Speech enhancement with an adaptive Wiener filter," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 53-64, 2014.

[50] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13-29, 2014.

[51] K. T. Andersen and M. Moonen, "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 97-107, 2017.

[52] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussian, S. A. R. Al-Haddad, and W. A. Jassim, "Low-distortion MMSE speech enhancement estimator based on Laplacian prior," *IEEE Access*, vol. 5, pp. 9866-9881, 2017.

[53] R. K. Kandagatla and P. V. Subbaiah, "Speech enhancement using MMSE estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *Speech Communication*, vol. 96, pp. 10-27, 2018.

[54] H. R. Abutalebi and M. Rashidinejad, "Speech enhancement based on β-order MMSE estimation of Short Time Spectral Amplitude and Laplacian speech modeling," *Speech Communication*, vol. 67, pp. 92-101, 2015.

[55] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase,"·*IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129-132, 2012.

[56] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1445-1457, 2013.

[57] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, 1995.

[58] K. Hermus and P. Wambacq, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 045821, 2006.

[59] A. Borowicz and A. Petrovsky, "Signal subspace approach for psychoacoustically motivated speech enhancement," *Speech Communication*, vol. 53, no. 2, pp. 210-219, 2011.

[60] M. Kalantari, S. R. Gooran, and H. R. Kanan, "Improved embedded pre-whitening subspace approach for enhancing speech contaminated by colored noise," *Speech Communication*, vol. 99, pp. 12-26, 2018.

[61] E. V. de Payer, "The subspace approach as a first stage in speech enhancement," *IEEE Latin America Transactions*, vol. 9, no. 5, pp. 721-725, 2011.

[62] B. Wiem, P. Mowlaee, and B. Aicha, "Unsupervised single channel speech separation based on optimized subspace separation," *Speech Communication*, vol. 96, pp. 93-101, 2018.

[63] P. Sun, A. Mahdi, J. Xu, and J. Qin, "Speech enhancement in spectral envelop and details subspaces," *Speech Communication*, vol. 101, pp. 57-69, 2018.

[64] F. Bao, W. H. Abdulla, F. Bao, and W. H. Abdulla, "A New Ratio Mask Representation for CASA-Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 27, no. 1, pp. 7-19, 2019.

[65] X. Wang, F. Bao, and C. Bao, "IRM estimation based on data field of cochleagram for speech enhancement," *Speech Communication*, vol. 97, pp. 19-31, 2018.

[66] X. Wang, C. Bao, and F. Bao, "A model-based soft decision approach for speech enhancement," *China Communications*, vol. 14, no. 9, pp. 11-22, 2017.

[67] S. Liang, W. Liu, and W. Jiang, "A new Bayesian method incorporating with local correlation for IBM estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 476-487, 2012.

[68] A. Narayanan and D. Wang, "A CASA-based system for long-term SNR estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2518-2527, 2012.

[69] G. Hu, and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067-2079, 2010.

[70] Y. K. Lee and O. W. Kwon, "Application of shape analysis techniques for improved CASA-based speech separation," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 1, pp. 146-149, 2009.

[71] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, 3350-3359, 2014.

[72] N. Rehman, C. Park, N. E. Huang, and D. P. Mandic, "EMD via MEMD: multivariate noise-aided computation of standard EMD," *Advances in Adaptive Data Analysis*, vol. 5, no. 02, pp. 1350007, 2013.

[73] A. Upadhyay and R. B. Pachori, "Speech enhancement based on mEMD-VMD method," *Electronics Letters*, vol. 53, no. 7, pp. 502-504, 2017.

[74] K. Khaldi, A. O. Boudraa, and M. Turki, "Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement," *IET Signal Processing*, vol. 10, no. 1, pp. 69-80, 2016.

[75] L. Zao, R. Coelho and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 899-911, 2014.

[76] M. E. Hamid, M. K. I. Molla, X. Dang, and T. Nakai, "Single channel

speech enhancement using adaptive soft-thresholding with bivariate EMD," *ISRN signal processing*, 2013.

[77] N. Chatlani and J. J. Soraghan, "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1158-1166, 2011.

[78] K. Khaldi, A. O. Boudraa, A. Bouchikhi, and M. T. H. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 873204, 2008.

[79] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP'79, IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, IEEE, 1979, pp. 208-211.

[80] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 799-807, 2001.

[81] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *ICASSP*, vol. 4, 2002, pp. 44164-44164.

[82] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59-67, 2004.

[83] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, 2003.

[84] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement,"·*IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700-708, 2003.

[85] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225-246, 1969.

[86] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[87] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of* IEEE *International Conference on Acoustics, Speech, and Signal Processing 2001 (ICASSP'01)*, vol. 2, 2001, pp. 749-752.

[88] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.

Nasir Saleem

Engr. Nasir Saleem received the B.S degree in Telecommunication Engineering from University of Engineering and Technology, Peshawar-25000, Pakistan in 2008 and M.S degree in Electrical Engineering from CECOS University, Peshawar, Pakistan in 2012. He was a senior Lecturer at the Institute of Engineering and Technology, Gomal University, D.I.Khan-29050, Pakistan. He is now Assistant Professor in Department of Electrical Engineering, Gomal University, Pakistan. His research interests are in the area of digital signal processing, speech processing and enhancement.

Muhammad Irfan Khattak

Muhammad Irfan Khattak is working as an Associate Professor in the Department of Electrical Engineering in the University of Engineering and Technology Peshawar. He did his B.Sc Electrical Engineering from the same University in 2004 and did his PhD from Loughborough University UK in 2010. His research interest involves Antenna Design, On-Body Communications, Speech processing and Speech Enhancement.

Elena Verdú Pérez

Elena Verdú received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, speech and image processing, data mining and expert systems.