JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 03 Issue 02
2022

Research Article
https://doi.org/10.56979/302/2022/50

Spectral Methods for Single Channel Speech Enhancement in Multi-Source Environment

Alamgir Rustrana[1*], Sarmadullah Khan[2], and Sheeraz Ahmad[3]

[1]Department of Electrical Engineering, CECOS University, Peshawar, Pakistan.
[2]Career Dynamics Research Center, Peshawar, Pakistan.
[*]Corresponding Author: Alamgir Rustrana. Email: alamgirustrana@gmail.com

Received: May 12, 2022 **Accepted:** September 13, 2022 **Published:** September 27, 2022.

Abstract: Speech communication for both humans and automatic devices can be negatively impacted by background noise, which is common in real environments. Among many techniques, speech separation using a single microphone is the most desirable from an application standpoint. The resulting monaural speech separation problem has been a central one in speech processing for several decades. However, its success has been limited thus far. This research presents work that develops speech separation systems using combinations of T-F masking, DNNs, and model-based reconstruction. The aim of each system is to improve the perceptual quality of the speech estimates. The performance of many speech processing applications is severely degraded when both noise and reverberation are present. The proposed solution has been tested in the simulation environment and based on the simulation result, it is observed that the speech enhancement can easily be performed through the integration of the solution. This research suggests two-staged noise reducing systems in order to reduce the background noise through a single microphone recording in a low-SNR based on ideal binary masking and Wiener filter. It has two stages. Firstly, for background noise reduction, a Wiener filter with an enhanced SNR is utilised on noisy speech. Secondly, IBM is calculated in each time–frequency channel through utilisation of the pre–processed speech from the first stage and the matching of the time–frequency channels to a pre-selected threshold in order to minimise residual noise. These channels meeting the threshold requirement are conserved while all the other ones are attenuated.

Keywords: Speech enhancement, noise, SNR, Wiener filter, PESQ.

1. Introduction

Due to its numerous features, such as automated speech recognition, hearing aids, voice communication devices, etc., hands-free cell phone use is widely accepted as standard in the present period. In this framework, procurement of unpolluted voice from near as well as far microphones is incredibly necessary for creating top-quality systems based primarily on speech. Due to the problem of noise in terms of being introduced in the background or in the room because of echoes, results in dropping the microphone voice signals. Though, in multisource conditions, extra discourse sources are available, making the issue additionally testable. Such type of challenges encouraged the study on single network voice improvement procedures. This research develops optimal spectral procedures for single network voice improvement in multisource scenarios, which results in an increase in quality along with unambiguousness of corrupted speech. Signal processing has multiple sub-fields. Audio signal processing is one of the subfields which deals with the electronic processing of audio signals. Audio signal processing can be applied in both analogue and digital forms of audio. In analogue audio signal processors, the audio signal is a variation of an

electric signal, while digital audio processing works on a digital representation of the same audio signal. In audio signal processing, digital audio signals are converted to analogue and vice versa. During the process, we can manipulate the frequency of the given signal. Due to the advancement of technology, audio processing can now be performed even on a common household PC.

Speaking is one of the most traditional and effective forms of communication there is. Face-to-face interaction was required to communicate verbally in the beginning, but as time went on, cellphones were added. Telephones made it possible for people to communicate through speech at very large distances. With time, the requirements of communication grew even further, and it all led to further advancement in the field of telecommunication, and eventually we ended up with wireless communication devices, which were later perfected in the form of cellular communication. Today, cell phones are an integral part of our lives. In order to ensure the efficient and smooth transmission of audio through our communication systems, digital audio signal processing is actively used. Due to the ability of our latest communication devices to perform complex computation very efficiently, we have been able to include more and more complex audio processing algorithms into the system. A simple single-channel model for audio enhancement can be seen in figure 1.

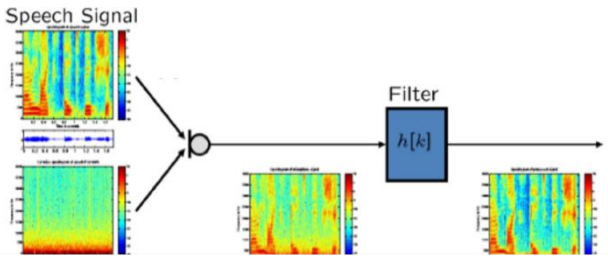

Figure 1. Speech Enhancement Single Channel Mechanism

Due to the fact that speech is considered a direct form of communication, it has become clear that speech is an amazing tool for human interaction [1, 2]. An algorithm that is carried out by a computer programme is used in an auto speech recognition (ASR) technology to modulate a speech signal into a series of words.

The main purpose of enhancement of speech is to lower the noise part of the signal without disturbing the original signal. Up to this point, a lot of study has been conducted on speech augmentation. However, the problems of lingering noise in the form of musical tones and distortions in the original signal are still open. Numerous studies have been carried out to address the issues, and they have addressed the issues to a certain degree of success in each of the research. Moreover, de-noising algorithms are used to leave minimal distortion in a signal during the process of recovering the original signal. This is, however, a complicated process. Since there will always be noise in a system, be it in the time domain or frequency, which will overlap the original signal, the removal of distortion from the original signal becomes complex. Removing noise from a speech signal is called speech enhancement. For this process, the frequency domain is preferred since a lot of work is done in this mode and we can build on the century of research already done in it.

The unique spectral subtraction [3, 4] method removes a noise floor without regard to noise frequency distribution. It has paved the way for increasingly sophisticated noise reduction techniques [5, 6], which employ a frequency-dependent gain function in the spectrum to reduce noise. Since many speech models operate in the frequency domain, many speech augmentation methods rely on parameters of estimated frequency spectra of the input data, and the quality of the estimates affects how effective they are. Short-time spectral amplitude (STSA) type enhancement algorithms have been used in several studies to greatly reduce the presence of musical noise and improve the general quality of the improved speech [4, 7, 12]. When using the signal subspace (SSB) algorithm [10], low-variance multitier spectral estimates (MTSE) [13] are used to produce speech spectra that are free of musical noise.

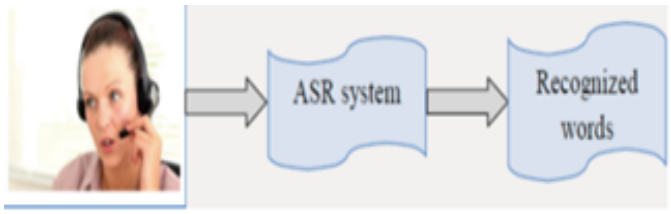

Figure 2. Auto Speech Recognition System

Usage of a hands-free cell phone has now been considered as standard due to its various features like automatic speech recognition, hearing aids, voice communication devices etc. In this framework, procurement of unpolluted voice from near as well as far microphones is incredibly necessary for creating top-quality systems based primarily on speech. Due to the problem of noise in terms of being introduced in the background or in the room because of echoes, it results in dropping the microphone voice signals. Though, in multisource conditions, extra discourse sources are available, making the issue additionally testable. Such type of challenges encouraged the study on single network voice improvement procedures. The objectives of research are to develop optimal spectral procedures for single network voice improvement in multisource scenarios, which results in an increase in quality along with unambiguousness of corrupted speech.

In speech recognition, the focal issue related to research is the way to extricate discriminative and remarkable highlights from discourse signals Some dialogue highlights have been put forth in the writing to help along this journey. Thus, the problem in this situation is to properly include the four categories: Language-related characteristics (words and discourse), context-related details (such as subject, sexual orientation, and turn-level highlights speaking to nearby and global parts of the exchange), half-and-half highlights that combine acoustic highlights with other data, and acoustic characteristics are the four categories. Two crucial concerns are required in order to construct recognition systems. They are a choice of the pertinent highlights that include the remarkable information to be clearly distinguishable by any model of classification. the proper selection of tests for creating a model of an arrangement. The accompanying sign features that directly influence the discourse acknowledgments framework must be addressed by the discourse analysis methodologies. Given the inter-speaker disparities of the discourse leads to the perfect preparation of a characterization model, there is a huge range of articulations that may be made and no confirmation of the best choices.

By using low-variance and smooth autoregressive multitier (ARMT) spectral calculations in STSA-type speech enhancers, a pertinent method eliminates the need for any further wavelet de-noising [8]. In [14], it is demonstrated that careful consideration of phase can result in a significant improvement and a reduction in musical noise, defying the widely held belief that the ear is insensitive to the phase of a signal.

The motivation for the work presented in this research comes from the continuously growing demand for effective speech communication systems. This trend promises to grow further in the future as human-machine interaction becomes increasingly popular. There are four steps in the improvement process. First, using time-domain techniques, noise-filled speech is divided into two brief time frames. Second, the FFT is used to translate each frame into the frequency domain. The analysis phase of the procedure is what we refer to as this. Third, a noise-reducing filter is created for each frequency band and applied to the STFT coefficients to approximate a clear speech spectrum. Finally, using an inverse FFT, the approximate clean speech spectrum is used to create the speech that has been augmented in the temporal domain (IFFT). This structure for voice enhancement handles different frequencies independently and is efficient in terms of computing. This allows for a great deal of flexibility in how we can use the noise statistics and our understanding of speech perception to better the performance. As a result, this particular framework has received the majority of attention in the past while discussing the speech enhancement process.

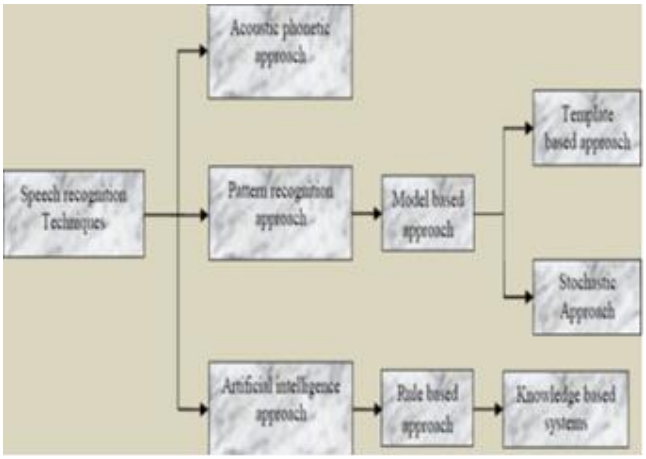

Figure 3. Systems of Recognition Speech Classification

Moreover, all these past references propose a certain relationship between the spectral estimation and the quality of the enhanced speech. Recently, analytical studies have been conducted to study the aforementioned relationships as in papers [9, 11, 12], and [13]. In this research, we use the technique proposed in [14], but extend it to a 2-stage mechanism. The existing mechanism does not address the noise issue to the extent that is desirable for audio signals. The quality of the signal is further distorted after passing through the first stage. The introduction of the Weiner filter as the second stage will result in further noise filtering as desired for real-time filtering. To the best of our knowledge, we do not find any such existing technique in literature, and it will add value to the research community as we are convinced by the uniqueness of this method. We rely on the belief that by working on the method proposed and combining it with other filtering techniques, it will set a benchmark for all the upcoming enhancement systems that are utilised as the front-ends for ASRs.

The aim of the research is to propose single-source enhancement techniques that are robust to high levels of interfering noise and suitable for real-time implementation. In this research, a double-stage noise reducing system based on the Wiener filter is suggested to reduce noise. The process will use an enhanced SNR in order to lessen the frame interval using Ideal Binary Mask (IBM) [7] and a decision-direct approach [4]. IBM is denoted by linking SNR evaluation with a 0 dB threshold. However, IBM uses access to local instantaneous SNR in place of a priori SNR, which is defined as the ratio of speech power spectrum to noise spectrum at each time-frequency (T-F) unit. The effectiveness of the proposed system is then evaluated using the characteristics of residual noise and speech distortion for two distinct intruder noises (AWGN and babbling).

2. RELATED WORK

Commotion decrease frameworks are broadly utilized media transmission frameworks to upgrade the nature of the discourse correspondence in boisterous situations. Despite the fact that, an improved clamor decrease can be acknowledged by utilizing mouthpiece exhibit framework, yet for financial reasons, the majority of these frameworks depend on single amplifier.

A single mouthpiece commotion reduction system, at its most basic level, uses flexible separating activities to weaken time recurrence (T- F) units of boisterous discourse with low SNR and hold the T- F units with high SNR. Thus, basic areas of discourse are saved though clamor level is significantly diminished, prompting an improved discourse with decreased commotion level. Incalculable commotion decrease frameworks are accessible in writing along this line [7, 8, 9]. By reducing the mean square error between the evaluated/improved flag and the original flag, the Wiener channel [10, 11] is a direct channel used to recover unique discourse motion from the boisterous flag. In Wiener separating, it is decided which T-F unit of noisy talk should be limited and by how much by using specific weakening rules. The majority of these weakening recommendations are upgraded to make the improved discourse as close to the ideal discourse as is practicable.

Unmistakably, the nature of single receiver clamour decrease frameworks is dictated by the concealment rule. As a rule, a concealment rule with solid constriction will prompt a less boisterous discourse. Be that as it may, solid lessening results in more contortion. Besides, a reasonable constriction presents less twisting however accomplished constrained measure of clamor decrease. A boundless writing audit on time–recurrence can be found in [12]. Philosophies utilising twofold covers have uncovered liberal quality upgrades even at very low SNRs with less contortion. These hopeful outcomes have revived the analysts to create/gauge double covers and proposed it as the objective of computational sound-related scene examination (CASA) [12]. With these confirmations of value and clarity improvement, inquire about is done in the ongoing past in attempting to appraise these covers [13, 14, 15].

Though in the past few years, most of the enhancements have been made in the recognition of automatic speech, vigorous as well as accurate speech recognition is considered a stimulating issue on behalf of so-called complicated issues like change in the contents as well as the speaker itself, along with environmental falsification. A major challenge in the recognition of speech is overwhelming speaker inconsistency. Continuation in the exchange of ideas that is based on the speaker's emotional state as well as affected by the individual speaker's features. Therefore, the primary source of speech is recognised through the independent mode of the speaker.

A conventional feature called MFCC has recently been used for speaker identification. The MFCC is unable to extract meaningful feature values from the speech signal if the noisier information is presented. Creating a noise-adaptive classification algorithm for speech recognition in noisy environments is more difficult. The main problem with those strategies is that they necessitate minute assumptions about the information transmission and model parameters. The majority of discourse handling techniques in writing use the HMM and GMM for the simultaneous characterisation of feelings.

GMM-UBM was used for classification, although this method does not work well in noisy environments. GMM-UBM also needs consistent training with more data samples. Additionally, neural system-based organisation models need a lot of pre-processing data for better grouping, while acoustic models struggle to handle low-level components, nearby requests, and inborn characteristics.

Using a Turkish speech database, single-channel speech enhancement algorithms were assessed using objective quality and objective intelligibility measures. At SNR values of -10, -5, 0, 5, and 10 dB, automobile and babbling noise types contaminate the clean 30 sentences from the METU database. Given the degree of segmental SNR improvement, weighted spectral slope, short-time objective intelligibility values, and spectrogram representations, it has been discovered that the Karhunen-Loeve Transform outperforms other approaches in terms of both quality and intelligence. [24, 25]. A more effective single-channel blind source separation nonnegative matrix factorization (NMF) technique with applications to speech enhancement It is possible to improve the effect of voice enhancement by including a time correlation item in the objective function to constrain the time-varying gain coefficients of noise [26, 27].

A speech enhancement method is provided in [28, 29]. The speech enhancement method includes: estimating the direction of a speaker by using an input signal, generating direction information indicating the estimated direction; detecting speech of a speaker based on the result of estimating the direction; and enhancing the speech of the speaker by using the direction information based on the result of detecting the speech [30].

3. OBJECTIVES OF RESEARCH

The goal of the research is to provide spectral techniques that are ideal for any single-channel speech improvement in any multi-source scenario, which can boost the degraded speech's quality and intelligibility.

1. This research is analysing the speech quality evaluation to determine how much the speech has been compressed and how much the contamination has been cleaned. The quality evaluation measure is for the specific reason that perceptual evaluation of speech quality (PESQ) can be carried out for analysis till the bottom of the whole process. This research work and observations [8] utilised the perceptual evaluation of speech quality (PESQ), which is suggested by the ITU-T based theory and the simulation. The results will be discussed in the next section after they have been carried out for analysis and to get values.

2. This research considers a two-staged noise reducing system in order to reduce the background noise through a single microphone recording in a low-SNR based on ideal binary masking and the Wiener filter. Our suggested system has two stages. Firstly, for background noise reduction, a Wiener filter with an enhanced SNR is utilised on noisy speech. Secondly, IBM is calculated in each time–frequency channel through the utilisation of the pre–processed speech from the first stage and matching the time–frequency channels to a pre-selected threshold T in order to minimise remaining noise. Time–frequency channels meeting the threshold requirement are conserved while all the other channels are carefully attenuated.
3. Moreover, it is more challenging when acoustic sounds originate from sources that have similar spectral characteristics. To address these imperative issues in speech processing applications, this algorithm will help to solve the issue without bringing loss of speech quality and intelligibility.
4. This research demonstrates effective usage of different speech processing applications that are operating in noisy environments.

4. SYSTEM MODEL

Speech enhancement is a difficult task in noisy environments especially in multisource environments because the quality along with the properties of noise motions is able to vary affectedly time to time as well as application to application. In this manner, it is relentless to discover flexible single-channel speech improvement algorithms that truly work in various down-to-earth conditions, hence, speech quality and intelligibility are severely reduced.

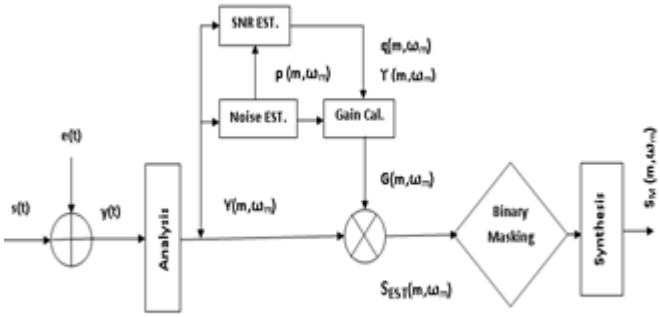

Figure 4. Proposed Model for the Scheme

In single-channel algorithms, the decrease in noise (quality improvement) is conceivable to the detriment of discourse bending (clarity decrease). In this way, it is difficult to fulfil the two measures in the meantime. In this exploration, the goal will be to build up the single-channel speech upgrade algorithm that can reduce noise (improve quality) yet not to the detriment of speech coherence (minimal distortion). A number of measures can be utilised to assess the execution of the speech improvement framework. These come in two formats: objective and abstract. There are a wide range of methodologies as to target tests, for example, the PESQ [8] measure, the IS [9,10], or SNR level addition. Abstract testing can likewise have utilized for assessment that is completed with the utilization of human helpers, and is generally observed to be best method for assessing the execution of speech improvement algorithms [22].

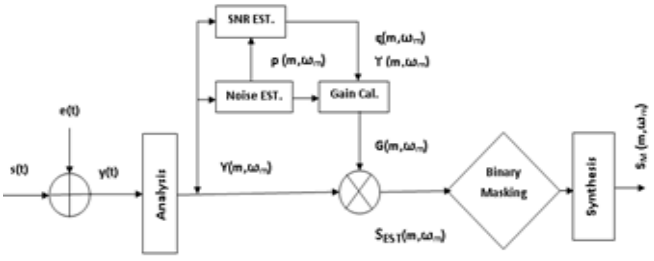

Figure 5. Mechanism for Noise Reduction

Figure 6 shows the flowchart of the presented technique. The mechanism starts with the framing mechanism. Every 10 ms, a Hann-windowed frame of 30 ms is extracted to process the audio signal. Each audio frame's discrete Fourier transform (DFT) is first converted into a block of autocorrelation's logarithmic magnitude spectrum or truncated spectrum. The logarithmic spectrum is a frequency conversion of the real spectrum, specifically the inverse Fourier transform. Thus, a truncated spectrum is a smoothed-out version of the original logarithmic spectrum in which only the lower-order coefficients are kept. We decided to keep the first J = 50 spectral coefficients with the 44100 Hz sampling rate we employed for this investigation in accordance with the general rule of selecting the length of the truncated spectrum to be shorter than an expected pitch period. To evaluate the performance, three SNR levels—5, 0, and 5 dB—were used.

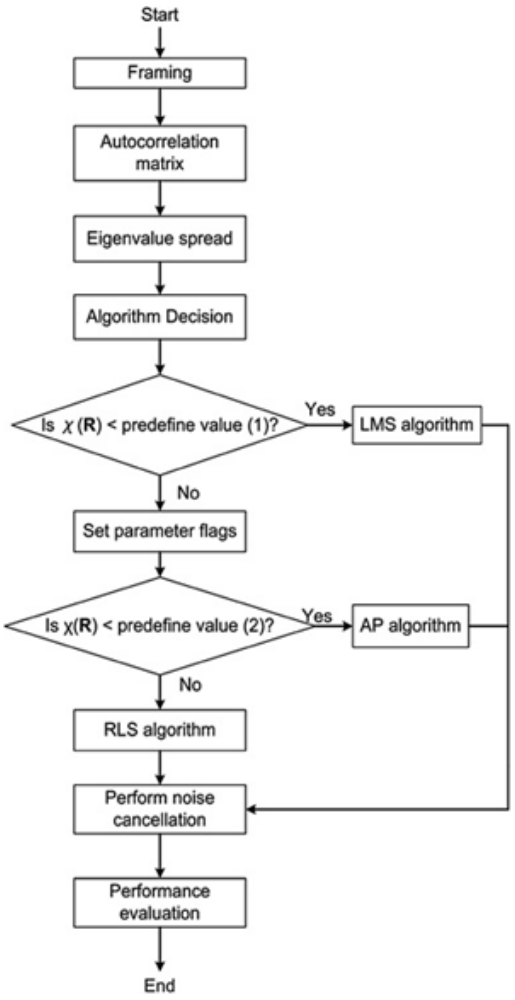

Figure 6. Flowchart for the Technique

5. MATHEMATICAL MODELING FOR NOISE REDUCTION SYSTEM

In classical terms, the noisy speech in any scenario is given by the equation:

$$y(t) = s(t) + e(t) \tag{1}$$

Where s(t) and e(t) represent the original noise-free speech and noise itself respectively. Y(m, ω_m), S(w, ω_m) and E(w, ω_m) be identified as ω_m spectral component of frame m of noisy speech y(t). Both noise and speech are non-stationary in nature but, in small-intervals (9–32ms), are assumed to be stationary. Hence, it is assumed here that both of them are quasi-stationary in nature. The spectral gain is convoluted in the computing of two basic SNR assessments, posteriori and a priori, given by

$$\gamma(m,\omega_m) = \frac{|Y(m,\omega_m)|^2}{E\left\{|E(m,\omega_m)|^2\right\}} = \frac{|Y(m,\omega_m)|^2}{\sigma_e{}^2(m,\omega_m)} \tag{2}$$

$$\xi(m,\omega_m)=\frac{\left|S(m,\omega_m)^2\right|}{E\left\{\left|E(m,\omega_m)\right|^2\right\}}=\frac{\sigma_S{}^2(m,\omega_m)}{\sigma_e{}^2(m,\omega_m)} \tag{3}$$

where E{.} is expectation operator, $\sigma_s(m, \omega_m)$ and $\sigma_\theta(m, \omega_m)$ is a posteriori and a priori SNR respectively. In real-time applications, the Power Spectral Density (PSD) of clean speech $|S(m, \omega_m)|^2$ and the noise $|E(m, \omega_m)|^2$ are unidentified as simply the noisy speech is accessible. PSD of noise is calculated through speech gaps exploiting the standard recursive relation:

$$\hat{\sigma}_e{}^2(m,\omega_m)=\xi\hat{\sigma}_e{}^2(m-1,\omega_m)+(1-\xi)\hat{\sigma}_Y{}^2(m-1,\omega_m) \tag{4}$$

where, ξ is the smoothing factor and $\hat{\sigma}_y^2(m-1,\omega_m)$ is the estimate from existing frame. Both the SNRs are computed as:

$$SNR_{INSTANT}(m,\omega_m)=\frac{\left|Y(m,\omega_m)\right|^2}{\sigma_e{}^2(m,\omega_m)}-1 \tag{5}$$

$$\xi_{PRIO}{}^{DD}(m,\omega_m)=\beta\frac{\left|G(m-1,\omega_m)*Y(m,\omega_m)\right|^2}{\hat{\sigma}_e{}^2(m,\omega_m-1)}+(1-\beta)F\left\{SNR_{INSTANT}(m,\omega_m)\right\} \tag{6}$$

Where $\xi_{PRIO}{}^{DD}(m, \omega_m)$ represents the computing of a priori SNR using decision-direct (DD) approach. DD is computationally effective and performs notable in noise reduction claims. However, in this procedure, a priori SNR tails the shape of instantaneous SNR which leads to one-frame deferral. Computed from the noisy speech $Y(m, \omega_m)$ by multiplying with Wiener filter gain function we get:

$$\left|S_{EST}(m,\omega_m)\right|=\left|Y(m,\omega_m)\right|*\mathrm{G}^{DD}(m,\omega_m) \tag{7}$$

Square root of the Wiener gain function is computes as given by equation:

$$G^{DD}(m,\omega_m)=\sqrt{\frac{\xi_{PRIO}{}^{DD}(m,\omega_m)}{\xi_{PRIO}{}^{DD}(m,\omega_m)+1}} \tag{8}$$

To reduce residual noise, ratio of estimated magnitude spectrum to clean speech ($|S(m, \omega_m)|/|S_{EST}(m, \omega_m)|$) is compared against a predefined threshold T. T–F units satisfying the constraint i.e. ($|S(m, \omega_m)|/|S_{EST}(m, \omega_m)|$)[T are preserved whereas those violating the constraints are attenuated. The modified magnitude spectrum $S_M(m, \omega_m)$ is calculated as:

$$\left|S_M(m,\omega_m)\right|=\begin{cases}\left|\bar{S}_{EST}(m,\omega_m)\right| & \left|S_{EST}(m,\omega_m)\right|/\left|S(m,\omega_m)\right|\geq T\\ 0 & \left|S_{EST}(m,\omega_m)\right|/\left|S(m,\omega_m)\right|<T\end{cases} \tag{9}$$

Many objective measures derived in literature for performance of noise reduction systems include PESQ–MOS and segmental SNR (SNR_{SEG}) [25]. PESQ–MOS measure has been experimented to produce good correlation with MOS; resulting in MOS results from 1 to 5, and high score indicates better speech quality. Likewise, SNRSEG is also extensively used objective measure and finds best correlation in noise reduction. SNR_{SEG} is defined as:

$$SNR_{SEG}(m,\omega_m)=\frac{10}{M}\sum_{m-0}^{M-1}\log_{10}\left(\frac{\left|S(m,\omega_m)\right|^2}{\left\|S(m,\omega_m)\right|-\left|S_{EST}(m,\omega_m)\right\|^2}\right) \tag{10}$$

Where $S(m, \omega_m)$ and $\hat{S}(m, \omega m)$ show the frames of clean and estimated speech respectively. To discard non-speech frames, every frame was threshold by a 0dB lower bound and -35dB upper bound. Therefore, ITU-T Recommendation P.835 is used to measure the speech distortion and residual noise. P.835 measure is formulated by relating the basic objective measures to establish composite measure [10]:

Csig=3:093-1:029SL+0:603SP-0:009SWS

Cbak=1:634+0:478SP-0:007SWS
+0:063SSNRSEG (11)

6. RESULTS AND DISCUSSION

As per time differentials, multiple types of tools and techniques have been developed and it is represented that in order to test the proposed solution, proper testing must be carried out so that it can be ensured that the proposed solution is better than the existing solution. Better outcomes can be obtained if the right testing methodology is applied. Testing is a difficult process, and in this aspect of this research, we considered the new method of a binary mask. The two testing parameters that are considered in the proposed approach are speech intelligibility enhancement and the quality of the synthesised speech, to overall improve the quality and the throughput.

Objective evaluation and calculated measurement values are used to evaluate and validate the performances and to estimate the calculations of the given speech enhancement algorithms II through compression algorithms and techniques. The PESQ approach is chosen as an ITU-T recommendation, and the outdated ITU-T recommendation—which was woefully inadequate for the intended purpose and is still unacceptable to remove distortion caused by unwelcome frequency—is replaced. The [20] theory contains all of the theoretical information that was pertinent to PESQ. The PESQ scores fall within the range of -0.5 to 4.5, however the output range is a follower of the MOS, or mean opinion score, which had a starting point of 1.0 and a finishing point of 4.5. High-level data indicates progress, and vice versa. The first and most popular way for evaluating the effectiveness of speech enhancement techniques and implementations reflected by appropriate flow charts is the objective metric based on SNR. Given that the mean ratios are taken over the full signal, which is non-stationary, and that the speech signals rapidly fluctuate, the SNR based estimation not only accurately reflects the speech quality for compression but also accurately reflects the strong association with the speech quality. SNR must be determined in small packets of minor pieces, and their mean must then be estimated for segmental or fragmental SNR identification (SNRSeg) [12].

The performances and outcomes of the given algorithms' and the base of approaches' development are shown in Table 1 accordingly. With the suggested algorithm and methods in all signal to noise ratio estimate levels and noise conditions in the presence of undesired frequencies, which produce the distortion in signal and interrupt it for contamination, a noticeable and readable development in PESQ phenomena is seen and analysed. According to noisy speech, the 5 dB street noise NRPCA (PESQ=1.15) has the maximum improvement, while the 5 dB babbling noise (PESQ=0.42) has the lowest improvement. The highest PESQ scores at 0 dB, 5 dB, and 10 dB noise levels were reported in the exhibition hall for analysis ΔPESQ=2.86, ΔPESQ=3.13, and ΔPESQ=3.31, respectively. The PESQ score improved from 2.02 with LMMSE to 2.72 with the proposed algorithm (ΔPESQstreet=0.71) in street noise at 0dB. Similarly, the PESQ score improved from 2.20 with subspace to 2.69 with the proposed algorithm (ΔPESQ babble=0.49) at 5dB bubble noise. Also, the PESQ score improved from 2.39 with NRPCA to 3.31 with the proposed algorithm (ΔPESQ Exhibition hall=0.92) at 10 dB.

Table 1. PESQ scores in various noise environments

Noise Type	Methods	0dB	5dB	10dB
Airport Noise	LMMSE	2.05	2.34	2.63
	Subspace	1.91	2.33	2.61
	NRPCA	1.84	2.04	2.47
	Proposed	2.76	2.98	3.10
Car Noise	LMMSE	2.19	2.39	2.73
	Subspace	1.98	2.30	2.65
	NRPCA	1.81	2.08	2.21
	Proposed	2.72	3.00	3.22

Street Noise	LMMSE	2.02	2.24	2.61
	Subspace	1.66	2.28	2.43
	NRPCA	1.71	1.85	2.24
	Proposed	2.72	3.0	3.25
Babble Noise	LMMSE	2.04	2.31	2.76
	Subspace	1.77	2.20	2.58
	NRPCA	1.80	2.09	2.35
	Proposed	2.60	2.69	3.18
Exhibition Hall	LMMSE	1.96	2.29	2.43
	Subspace	1.92	2.32	2.51
	NRPCA	1.72	2.11	2.39
	Proposed	2.86	3.13	3.31

The PESQ improvements performed are shown in the given figures.

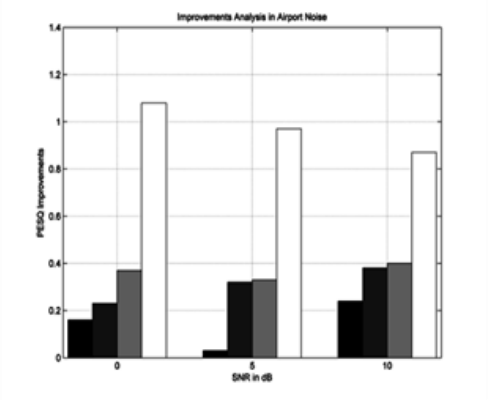

Figure 7. PESQ improvement in Airport Noise

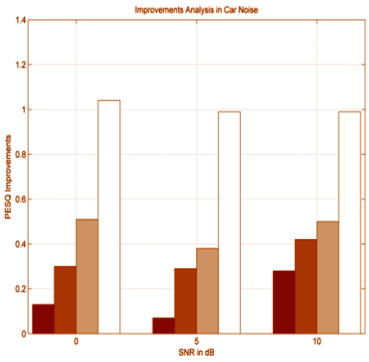

Figure 8. PESQ improvement in Car Noise

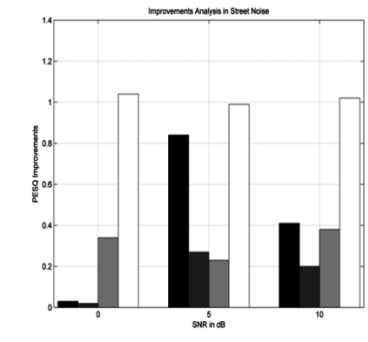

Figure 9. PESQ improvement in Street Noise

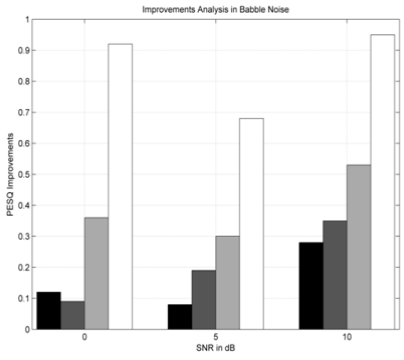

Figure 10. PESQ improvement in Babble Noise

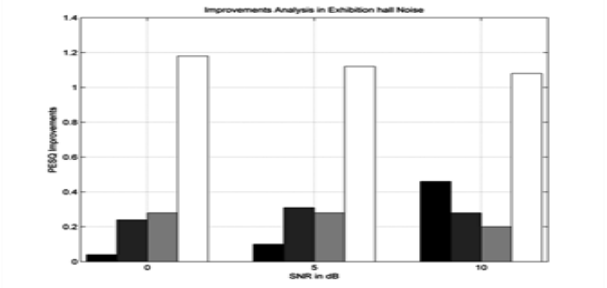

Figure 11. PESQ improvement in Exhibition hall Noise

The SNR segmented score improved from 1.66 with LMMSE to 3.35 with the proposed algorithm (ΔSNRSegstreet=1.69) in street noise at 0dB. Similarly, the SNRSeg score improved from 3.14 with NRPCA to 4.99 with the proposed algorithm (ΔSNRSegairport=1.85) at 5dB of car noise. Also, the SNRSeg score improved from 5.21 with NRPCA to 7.36 with the proposed algorithm (ΔSNRSegExibitionhall=2.15) at 10dB exhibition hall. The samples spectrograms for all processing algorithms and enhancement techniques to enhance signals. By viewing and assessing the spectrograms of the signal processing algorithms for the desired purpose, the voice utterance was interrupted by the babbling noise at 0dB SNR level. The suggested method, which is depicted in Figure 17, is superior at reducing or eliminating background noise and un-desired frequencies and cleaning speech of contaminated frequencies.

Table 2. SNRSeg scores in various noise environments

Noise Type	Methods	0dB	5dB	10dB
Airport Noise	LMMSE	2.36	3.85	5.71
	Subspace	2.02	3.89	5.96
	NRPCA	1.69	3.14	5.84
	Proposed	3.34	4.99	6.19
Car Noise	LMMSE	2.41	4.10	5.69
	Subspace	2.15	4.28	6.03
	NRPCA	1.51	3.21	5.03
	Proposed	3.0	4.45	6.73
Street Noise	LMMSE	1.66	3.93	6.08
	Subspace	1.13	4.28	5.40
	NRPCA	1.56	3.30	5.53
	Proposed	3.35	4.91	7.23
Babble Noise	LMMSE	1.84	3.56	5.13
	Subspace	1.48	3.45	5.46
	NRPCA	1.49	3.35	5.47
	Proposed	2.95	4.60	6.8
Exhibition Hall	LMMSE	2.16	4.34	5.71
	Subspace	2.44	4.78	5.71
	NRPCA	1.49	3.95	5.21
	Proposed	3.69	4.83	7.36

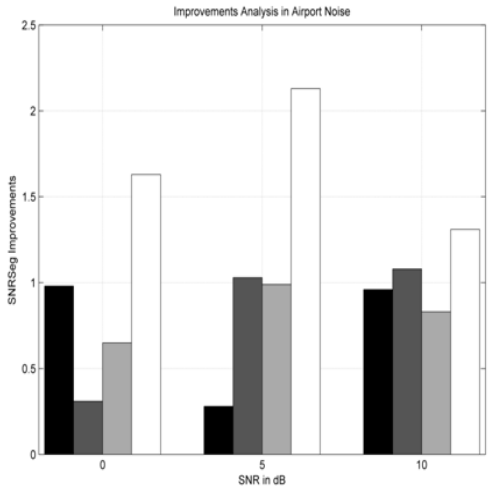

Figure 12. SNRSeg improvements in Airport Noise

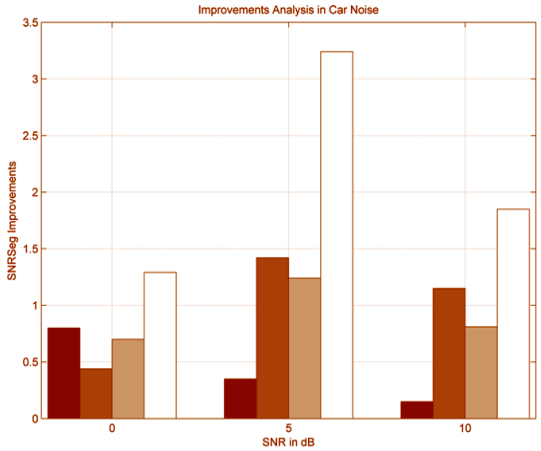

Figure 13. SNRSeg improvements in Car Noise

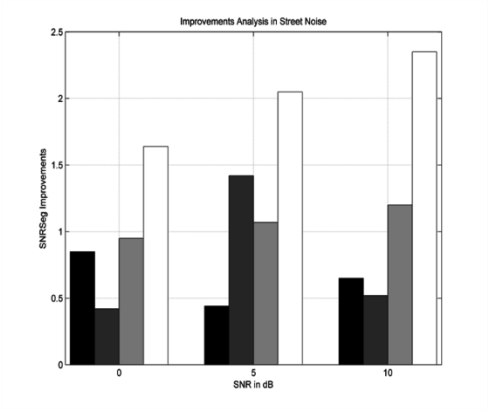

Figure 14. SNRSeg improvements in Street Noise

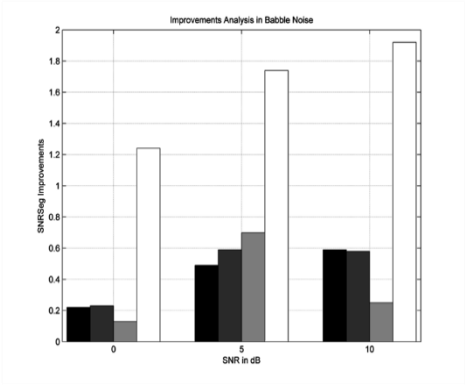

Figure 15. SNRSeg improvements in Babble Noise

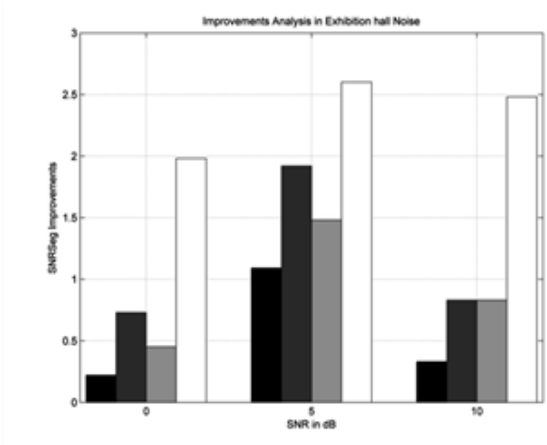

Figure 16. SNRSeg improvements in Babble Noise

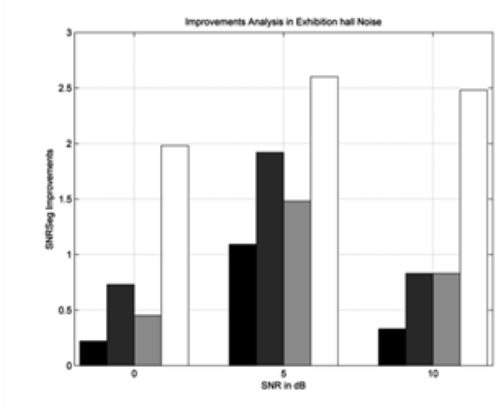

Figure 17. SNRSeg improvements in Exhibition hall Noise

Table 3. ANOVA based Statistical Analysis of Residual Noise and Speech Distortion

Residual Noise Analysis (BAK)					
Noise Type	SNR (dB)	Un-P	Wiener filter	Stage 1	Proposed BAK
Airport	0	1.58	2.08	2.21	2.75
	5	1.99	2.32	2.39	3.03
	10	2.48	2.86	2.98	3.35
Babble	0	1.58	2.06	2.23	2.68
	5	1.99	2.41	2.47	2.98
	10	2.48	2.85	2.98	3.29
Car	0	1.63	2.19	2.25	2.71
	5	1.98	2.41	2.59	3.03
	10	2.44	2.84	2.97	3.29
Exhibition Hall	0	1.58	2.09	2.18	2.86
	5	1.99	2.47	2.52	3.15
	10	2.48	2.68	2.79	3.38

Street	0	1.64	2.09	2.19	2.73
	5	2.09	2.47	2.56	3.03
	10	2.55	2.68	2.92	3.34

The highest notable improvement is 10 dB airport (ΔBAKairport=3.35) while the lowest one is 0 dB, i.e., babble noise, which means (ΔBAKbabble=2.68). The SIG segmented scores are improved from 4.28 with Exhibition Hall using the proposed algorithm (ΔSIGExhibition hall =4.28) at 10dB, while the lowest is 3.65 with street means (ΔSIGstreet=3.65) at 0dB.

Table 4. Speech Distortion Analysis(SIG)

					SIG
Airport	0	2.51	2.47	2.77	3.77
	5	2.94	2.99	3.28	4.0
	10	3.41	3.41	3.69	4.03
Babble	0	2.51	2.49	2.63	3.67
	5	2.94	3.00	3.12	3.97
	10	3.41	3.47	3.51	4.19
Car	0	2.46	2.72	2.84	3.75
	5	2.95	3.24	3.38	4.07
	10	3.43	3.86	3.92	4.23
Exhibition Hall	0	2.51	2.44	2.77	3.88
	5	2.94	2.98	3.28	4.15
	10	3.41	3.27	3.69	4.28
Street	0	2.45	2.41	2.53	3.65
	5	2.93	3.08	3.28	3.98
	10	3.46	3.61	3.63	4.23

The suggested speech improvement in Table 5 is time complex, which indicates how long the algorithm will take to run through to completion. The database's 30 combinations, with a combined run-time (RT) of 150 seconds, were compared. Table 5 demonstrates that procedures that don't entail training execute considerably more quickly than those that do (DNN). For instance, NRPCA requires 20.5 seconds to complete the processing of a single utterance at a given SNR level (let's assume 0dB). As a result, the run-time for 30 utterances is 20.5*30=615 seconds. The total time required by NRPCA to generate enhanced voice utterances is 3075 seconds for five SNR levels (0dB, 5dB, 10dB) when RT is multiplied by factor 5. For one syllable at one SNR level, the suggested algorithm requires 10 seconds to run through to completion. As a result, the suggested method's temporal complexity is 1500 seconds, indicating that it converges quickly compared to other approaches.

Table 5. Time Complexity comparison in seconds

Algorithm	Total Duration	Run-time (RT) One SNR Level	Run-time (RT) Five SNR Levels	Training time One SNR level	Training time Five SNR levels
Noisy	150	-	-	-	-
NRPCA	-	615	5*RT = 3075	-	-
Subspace	-	150	5*RT = 750	-	-
Wiener Filter		180	5*RT = 900	-	-
DNN	-	1800	5*RT = 9000	400	2000
Proposed	-	300	5*RT = 1500	-	-

Figure 18 illustrates the use of the binary time-frequency based mask, along with time domain waveform analysis of the suggested algorithm and simulation in comparison to benchmark algorithms. The presented good compression techniques based results on speech signal are applied to clean speech utterance, noisy speech degraded by airport noise at 0 dB SNR value, NRPCA, speech processed by the subspace method, speech processed by L-MMSE, and the proposed represented algorithms for the process.

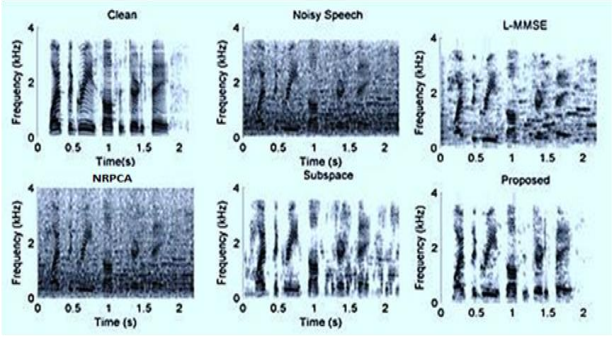

Figure 18. Time-varying spectral analysis and observations of proposed algorithm against baseline algorithms

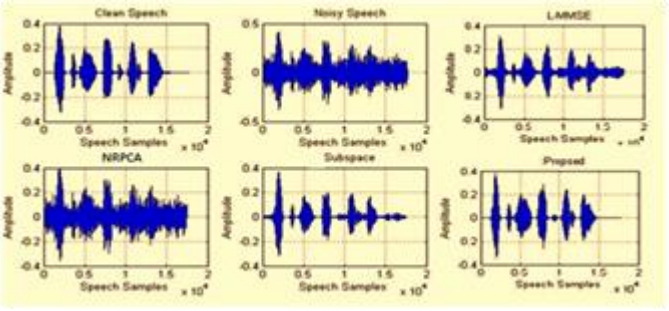

Figure 19. Time-waveform analysis and observations of proposed algorithm against baseline algorithms

7. CONCLUSION

In this study, a two-stage noise-reducing system was proposed. First, loud speech is subjected to a Wiener filter with an improved SNR for background noise reduction. In order to reduce residual noise, IBM is calculated in each time-freq channel using the pre-processed speech from the first stage and matching the time-freq channels to the pre-selected threshold T. To address these imperative issues in speech processing applications, this algorithm will help to solve the issue without bringing loss of speech quality and intelligibility. Another advantage of this research is the effective usage of different speech processing applications operating in noisy environments. Objective evaluation and calculated measurement values are used to evaluate and validate the performances and to estimate the calculations of the given speech enhancement algorithms II through compression algorithms and techniques. The PESQ method is chosen as an ITU-T recommendation, and it is replaced with the old version and previous ITU-T based recommendations, which are quite insufficient for the desired purpose and remain unacceptable to eliminate distortion due to unwanted frequency. This research work on observations utilised the PESQ, which is suggested by the ITU-T based theory and the simulation. Time-frequency channels play an important role in time-frequency channels and have discarded the constraints. The speech signals' overall help in signal-to-noise levels (0dB, 5dB, and 10dB) using babble and street noise creators are overall processes in the proposed solution. However, the improvements in speech have been reported and the intelligibility and quality, even at low signal-to-noise ratio levels, were able to produce better results and outcomes. One of the key gaps is that the project methodology has not been properly implemented and there is a high probability that if the project methodology is selected from the right approach, better results and throughput can be obtained. It is highly suggested that proactive planning is required so that the project can be completed within the specific given time with the specific duration and cost. For evaluation of the progress of speech enhancement and compression systems, multiple varieties can be used for the desired purpose.

References

1. Ding, H. (2011). Speech enhancement in transform domain (Doctoral dissertation, Nanyang Technological University).
2. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. Speech communication, 49(10-11), 763-786.
3. Loizou, P. C. (2007). Speech enhancement: theory and practice. CRC press.
4. Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on speech and audio processing, 7(2), 126-137.
5. Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE transactions on Speech and Audio Processing, 2(2), 345-349.
6. Charoenruengkit, W., Erdol, N., & Gunes, T. (2006, September). Parametric approach for speech denoising using multitapers. In 2006 14th European Signal Processing Conference (pp. 1-5). IEEE.
7. Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing, 32(6), 1109-1121.
8. Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE transactions on Speech and Audio processing, 12(1), 59-67.
9. Martin, R. (1994). Spectral subtraction based on minimum statistics. power, 6(8).
10. Sorqvist, P., Handel, P., & Ottersten, B. (1997, April). Kalman filtering for low distortion speech enhancement in mobile communication. In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 2, pp. 1219-1222). IEEE.
11. Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. Proceedings of the IEEE, 70(9), 1055-1096.
12. Vary, P., & Eurasip, M. (1985). Noise suppression by spectral magnitude estimation—mechanism and theoretical limits—. Signal processing, 8(4), 387-400.
13. Reidy, P. F. (2015). A comparison of spectral estimation methods for the analysis of sibilant fricatives. The Journal of the Acoustical Society of America, 137(4), EL248-EL254.
14. Händel, P. (2006). Power spectral density error analysis of spectral subtraction type of speech enhancement methods. EURASIP Journal on Advances in Signal Processing, 2007, 1-9.
15. Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267,.
16. Kadir, K. A. (2010). Recognition of Human Speech using q-Bernstein Polynominals. International Journal of Computer Application, 2(5), 22-28.
17. Reddy, D. R. (1976). Speech recognition by machine: A review. Proceedings of the IEEE, 64(4), 501-531.
18. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.
19. Shinde, R. B., & Pawar, V. P. (2012). A review on acoustic phonetic approach for marathi speech recognition. International Journal of Computer Applications, 59(2), 40-44.
20. Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. The Journal of the Acoustical Society of America, 110(2), 1150-1163.
21. Mohamed, A. R., Dahl, G. E., & Hinton, G. (2011). Acoustic modeling using deep belief networks. IEEE transactions on audio, speech, and language processing, 20(1), 14-22.
22. Deng, L. (2004). Switching dynamic system models for speech articulation and acoustics. In Mathematical Foundations of Speech and Language Processing (pp. 115-133). Springer, New York, NY.
23. Deng, L., & Yu, D. (2007, April). Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 4, pp. IV-445). IEEE.

24. Arslan, Ö., & Engın, E. Z. (2018, May). Evaluation of single-channel speech enhancement algorithms by using objective quality and intelligibility measures. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.
25. Tu, Y. H., Du, J., & Lee, C. H. (2019, May). DNN training based on classic gain function for single-channel speech enhancement and recognition. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 910-914). IEEE.
26. Kavalekalam, M. S., Nielsen, J. K., Christensen, M. G., & Boldt, J. B. (2018, April). A study of noise PSD estimators for single channel speech enhancement. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5464-5468). IEEE.
27. Chen, Y. (2017, May). Single channel blind source separation based on nmf and its application to speech enhancement. In 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN) (pp. 1066-1069). IEEE.
28. Wung, J., Souden, M., Pishehvar, R., & Atkins, J. D. (2020). U.S. Patent No. 10,546,593. Washington, DC: U.S. Patent and Trademark Office.
29. Bryan, Nicholas J., and Vasu Iyengar. "Speech enhancement for an electronic device." U.S. Patent No. 10,535,362. 14 Jan. 2020.
30. Cho, Jae-youn, C. U. I. Weiwei, and Seung-Yeol Lee. "Speech enhancement method and apparatus for same." U.S. Patent No. 10,529,360. 7 Jan. 2020.