Research Article

Speech Enhancement in Multi-Source Environments

Kanwal Rehman[1,⋆], Insia Mukhtar[2], and Muzammil Haider[3]

[1,2,3]Institute of Computing, Women University, Multan, Punjab, Pakistan.; Email: kanwal@wum.edu.pk, imukhtar@wum.edu.pk , mhaider@wum.edu.pk
⋆Corresponding author: Kanwal Rehman (kanwal@wum.edu.pk)

Article History

Academic Editor:
Dr. Muhammad Sajid
Submitted: January 1, 2024
Revised: February 27, 2024
Accepted: March 1, 2024

Keywords:
Speech improvement, noise, signal-to-noise ratio (SNR), Wiener filter, Perceptual Evaluation of Speech Quality (PESQ)

Abstract

Background noise, prevalent in real-world settings, can adversely affect speech communication for both people and automated equipment. Of several strategies, speech separation utilizing a single microphone is the most advantageous from an application perspective. The resultant monaural speech separation issue has been a pivotal concern in speech processing for numerous decades. Nonetheless, its success has been constrained to date. This research develops speech separation systems utilizing combinations of time-frequency masking, deep neural networks, and model-based reconstruction. The objective of each system is to enhance the perceived quality of the speech estimations. The efficacy of numerous speech processing applications is significantly compromised in the presence of both noise and reverberation. The proposed approach has been evaluated in a simulation environment, and the results indicate that voice enhancement can be effectively achieved through its integration. This study proposes two-stage noise reduction systems to diminish background noise in single microphone recordings with low signal-to-noise ratios, utilizing perfect binary masking and Wiener filtering techniques. It comprises two stages. Initially, a Wiener filter with an improved signal-to-noise ratio is employed for the reduction of background noise in noisy speech. Secondly, IBM is computed in each time-frequency channel by utilizing the pre-processed speech from the initial stage and aligning the time-frequency channels to a predetermined threshold to minimize residual noise. The channels that meet the threshold criteria are preserved, while all others are diminished.

1 Introduction

Hands-free cell phone use is currently seen as conventional due to its various capabilities, including automated speech recognition, hearing aids, and voice communication devices. In this context, the acquisition of pristine audio from both proximal and distal microphones is essential for developing high-quality speech-based systems. The issue of background noise and echoes results in the attenuation of microphone voice signals. In multisource conditions, more discourse sources are accessible, rendering the issue further testable. Such issues prompted the investigation of single network voice enhancement methodologies. This research formulates optimal spectral methods for enhancing single network voice in multisource environments, leading to improved quality and clarity of distorted speech.

Signal processing encompasses various sub-disciplines. Audio signal processing is a subfield that focuses on the electronic manipulation of audio signals. Audio signal processing can be utilized in both analog and digital formats of audio. In analogue audio signal processors, the audio signal represents a variant of an electric signal is utilized, but digital audio processing operates on a digital representation of that same audio signal. In audio signal processing, digital audio signals are transformed into analog signals and vice versa. Throughout the process, we can adjust the frequency of the specified signal. Advancements in technology now enable audio processing on a standard family PC.

Verbal communication is one of the most conventional and efficacious modes of interaction. Initially, spoken communication necessitated face-to-face connection; but, with time, telephones were introduced. Telephones enabled individuals to converse verbally over great distances. Over time, the demands of communication expanded, resulting in significant advancements in telecommunication, ultimately culminating in the development of wireless communication devices, which were further refined into cellular communication. Currently, mobile phones are an essential component of our existence. To guarantee the efficient and seamless transmission of audio via our communication systems, digital audio signal processing is employed. Our latest communication devices' capacity for efficient complicated computation has enabled the integration of increasingly sophisticated audio processing algorithms into the system.

Given that speech is regarded as a direct mode of communication, it is evident that it serves as an exceptional instrument for human interaction [1, 2]. An algorithm executed by a computer software is employed in automatic speech recognition (ASR) technology to convert a speech signal into a sequence of words.

The primary objective of voice enhancement is to reduce the noise component of the signal while preserving the integrity of the original signal. To date, extensive research has been undertaken on speech augmentation. Nonetheless, the issues of persistent noise manifesting as musical tones and distortions in the underlying signal remain unresolved. A multitude of studies have been conducted to tackle the difficulties, achieving varying degrees of success in each research endeavor. Furthermore, de-noising methods are employed to minimize distortion in a signal while recovering the original signal. This is, nonetheless, a complex process. The presence of noise in a system, whether in the time domain or frequency domain, complicates the removal of distortion from the original signal. The process of eliminating noise from a speech stream is referred to as speech enhancement. The frequency domain is favored for this approach because to the extensive work conducted in this mode, allowing us to leverage a century of prior study.

The distinctive spectral subtraction method eliminates a noise floor irrespective of the noise frequency distribution. It has facilitated the development of advanced noise reduction algorithms [5, 6], which utilize a frequency-dependent gain function in the spectrum to mitigate noise. Numerous speech models function in the frequency domain, leading many speech augmentation techniques to depend on parameters derived from the predicted frequency spectra of the input data, with the accuracy of these estimates influencing their efficacy. Short-time spectral amplitude (STSA) augmentation algorithms have been employed in many research to significantly diminish musical noise and enhance the overall quality of the processed speech [4, 7, 12]. The signal subspace (SSB) technique [10] employs low-variance multitier spectral estimates (MTSE) [13] to generate speech spectra devoid of musical noise.

The utilization of hands-free cell phones has become standard thanks to features such as automated speech recognition, hearing aids, and voice communication devices. Within this paradigm, the acquisition of pristine audio from both proximal and distal microphones is essential for developing high-quality speech-based systems. The issue of background noise, particularly from echoes, leads to the attenuation of microphone voice signals. In multisource conditions, additional discourse sources are accessible, rendering the issue more testable. Such challenges stimulated research on single network voice enhancement methods. The project aims to establish appropriate spectral methods for enhancing single network voice in multisource environments, leading to improved quality and clarity of distorted speech as shown in figure 1. In voice recognition, the primary research concern is the extraction of discriminative and salient features from speech signals. Several conversation features have been proposed in the literature to facilitate this process. The challenge in this scenario is to accurately incorporate the four categories: language-related characteristics (words and discourse), context-related details (such as subject, sexual orientation, and turn-level highlights pertaining to local and global aspects of the exchange), hybrid features that amalgamate acoustic attributes with other information, and acoustic

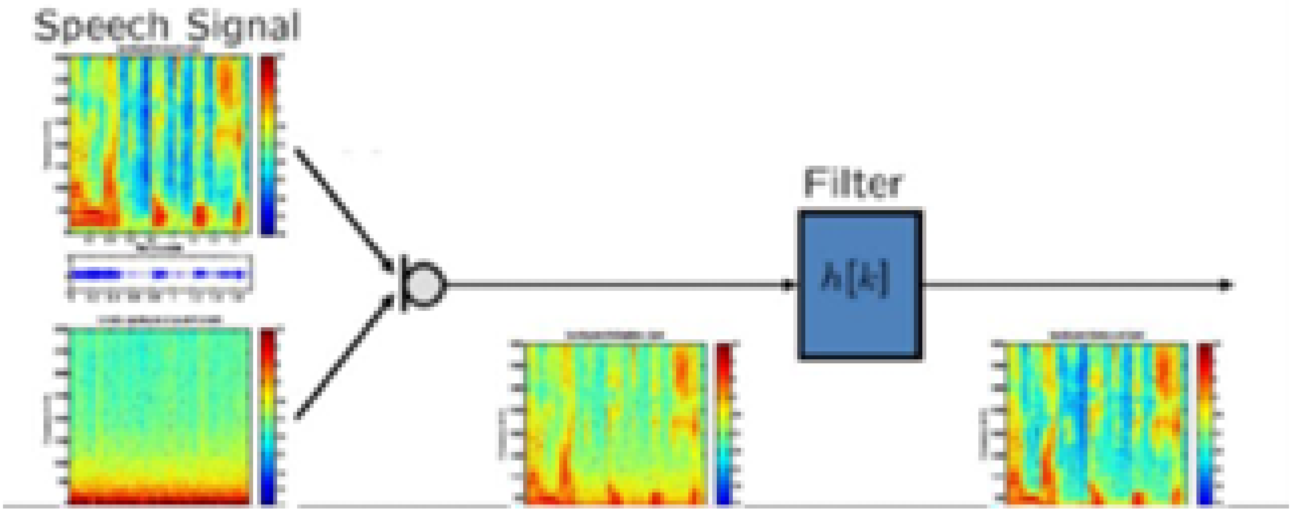

Figure 1: Speech Enhancement Single Channel Mechanism.

characteristics themselves. Two essential considerations are necessary for the development of recognition systems. They represent a selection of relevant highlights that encompass significant information, easily identifiable by any categorization methodology. the appropriate selection of assessments for developing a model of an arrangement. The accompanying sign elements that directly affect the discourse acknowledgment framework must be considered by discourse analysis approaches. The inter-speaker variations in discourse facilitate the optimal development of a characterisation model, resulting in a vast array of possible articulations without definitive validation of the most effective selections.

The application of low-variance and smooth autoregressive multitier (ARMT) spectral calculations in STSA-type speech enhancers effectively obviates the necessity for additional wavelet de-noising [8]. Reference [14] illustrates that meticulous attention to phase can lead to substantial enhancement and a decrease in musical noise, challenging the prevalent notion that the ear is indifferent to the phase of a signal.

This research is motivated by the increasing demand for efficient speech communication systems. This trend is expected to expand as human-machine interaction gains popularity. The improvement process has four stages. Initially, employing time-domain methodologies, noise-affected speech is segmented into two short temporal intervals. Secondly, the FFT is employed to convert each frame into the frequency domain. This is referred to as the analysis step of the technique. Third, a noise-reducing filter is generated for each frequency band and applied to the STFT coefficients to approximate a clear speech spectrum. Ultimately, an inverse FFT is employed to generate the enhanced speech in the temporal domain from the approximate clean speech spectrum (IFFT). This voice enhancement architecture processes various frequencies individually and is computationally efficient. This provides significant flexibility in utilizing noise data and our comprehension of speech perception to enhance performance. This approach has garnered significant interest recently in discussions on the speech enhancement process as shown in figure 2. Furthermore, all these previous examples suggest a specific correlation between spectral estimate and the quality of the augmented speech. Recent analytical investigations have examined the aforementioned correlations, as detailed in works [9, 11, 12, 13]. This research employs the method outlined in [14], but expands it into a two-stage system. The current approach inadequately resolves the noise issue for audio signals. The signal's quality is significantly compromised after traversing the initial stage. The implementation of the Weiner filter as the subsequent stage will yield more noise reduction, as required for real-time filtering. To our knowledge,

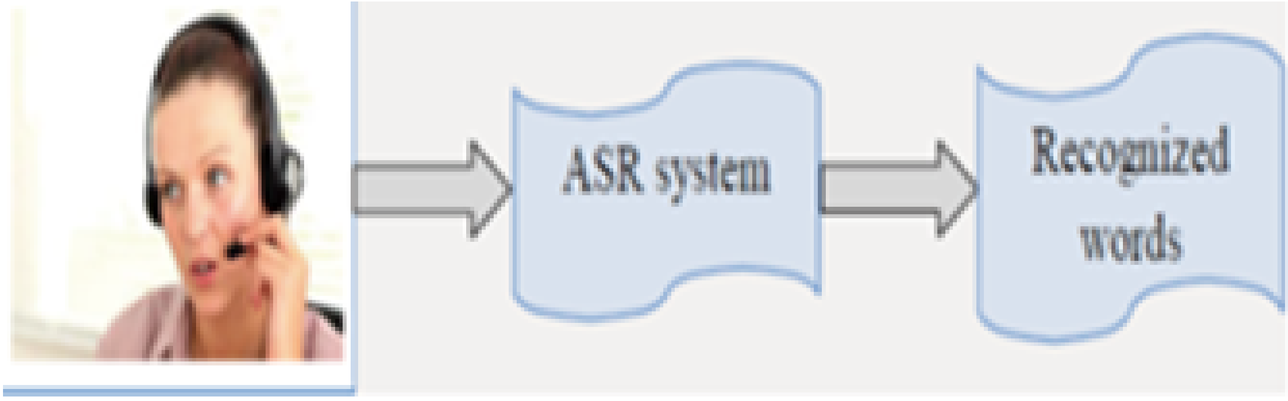

Figure 2: Auto Speech Recognition System.

no such strategy exists in the literature, and we believe our method will contribute significantly to the study community due to its uniqueness. We are confident that by implementing the proposed method and integrating it with additional filtering techniques, it will establish a standard for future enhancement systems employed as front-ends for ASRs.

The objective of the project is to provide single-source enhancement approaches that are resilient to significant amounts of interfering noise and appropriate for real-time application. This research proposes a dual-stage noise reduction system utilizing the Wiener filter to mitigate noise. The procedure will employ an improved signal-to-noise ratio to reduce the frame interval utilizing Ideal Binary Mask (IBM) [7] and a decision-directed methodology [4]. IBM is shown by correlating SNR assessment with a 0 dB benchmark. IBM utilizes access to local instantaneous SNR instead of a priori SNR, which is defined as the ratio of the speech power spectrum to the noise spectrum at each time-frequency (T-F) unit. The efficacy of the proposed system is subsequently assessed by analyzing the attributes of residual noise and speech distortion in relation to two specific intruder noises (AWGN and babble).

2 Related Literature

Noise reduction systems are widely employed in telecommunications to enhance the quality of speech communication in noisy environments. Although enhanced noise reduction can be achieved through the use of a microphone array system, most of these systems rely on a single microphone for economic reasons.

A singular mouthpiece noise reduction system fundamentally employs flexible separation mechanisms to attenuate time-frequency (T-F) units of loud speech with low signal-to-noise ratio (SNR) while preserving T-F units with high SNR. Consequently, fundamental regions of communication are preserved while the noise level is substantially reduced, resulting in enhanced dialogue with diminished auditory interference. Numerous frameworks for reducing incalculable commotion are documented in the literature [7, 8, 9]. By minimizing the mean square error between the assessed/enhanced signal and the original signal, the Wiener channel [10, 11] serves as a direct channel for recovering distinct discourse motion from the noisy signal. In Wiener separation, the determination of which T-F unit of loud speech should be attenuated and to what extent is conducted through the use of certain attenuation rules. The majority of these weakened recommendations are revised to align the enhanced discourse as closely as possible with the ideal discourse.

The characteristics of single receiver noise reduction systems are unequivocally governed by the concealment rule. A stringent concealment regulation will result in a quieter discourse. Nonetheless, substantial reduction yields greater distortion. Moreover, a judicious constriction exhibits reduced twisting while achieving a limited degree of noise reduction. A comprehensive literature review on time-recurrence is available in [12]. Philosophies with dual covers have revealed significant quality enhancements even at minimal SNRs with reduced distortion. These optimistic results have reinvigorated analysts to develop and assess double covers, proposing it as the goal of computational auditory scene analysis (CASA) [12]. With these confirmations of worth and clarity enhancement, inquiries have been conducted in the recent past to assess these coverings [13, 14, 15].

In recent years, advancements have primarily focused on automatic speech recognition; however, achieving robust and precise speech recognition remains a challenging problem due to factors such as content variation, speaker differences, and environmental distortions. A significant obstacle in speech recognition is pronounced speaker variability. The ongoing interchange of ideas is influenced by the speaker's emotional state and unique characteristics. Consequently, the principal source of speech is identified by the autonomous manner of the speaker.

The standard feature known as MFCC has lately been employed for speaker identification. The MFCC cannot derive significant feature values from the speech stream when extraneous noise is present. Developing a noise-adaptive classification algorithm for voice recognition in acoustically challenging conditions is more complex. The primary issue with these tactics is that they require precise assumptions on information transfer and model parameters. Most discourse management systems in writing employ Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) for the concurrent characterization of emotions.

GMM-UBM was employed for classification; nevertheless, this approach is ineffective in noisy environments. GMM-UBM requires uniform training with an increased number of data samples. Furthermore, neural system-based organizational models require extensive pre-processing of input for improved categorization, whereas acoustic models have challenges in managing low-level components, proximate demands, and inherent traits as shown in figure 3. Single-channel speech augmentation algorithms were evaluated using objective quality and intelligibility metrics, utilizing a Turkish speech database. At SNR levels of -10, -5, 0, 5, and 10 dB, car and babble noise types interfere with the pristine 30 sentences from the METU database. Analysis of segmental SNR enhancement, weighted spectral slope, short-time objective intelligibility metrics, and spectrogram visualizations reveals that the Karhunen-Loeve Transform surpasses alternative methods in quality and efficacy [24, 25]. An improved single-channel

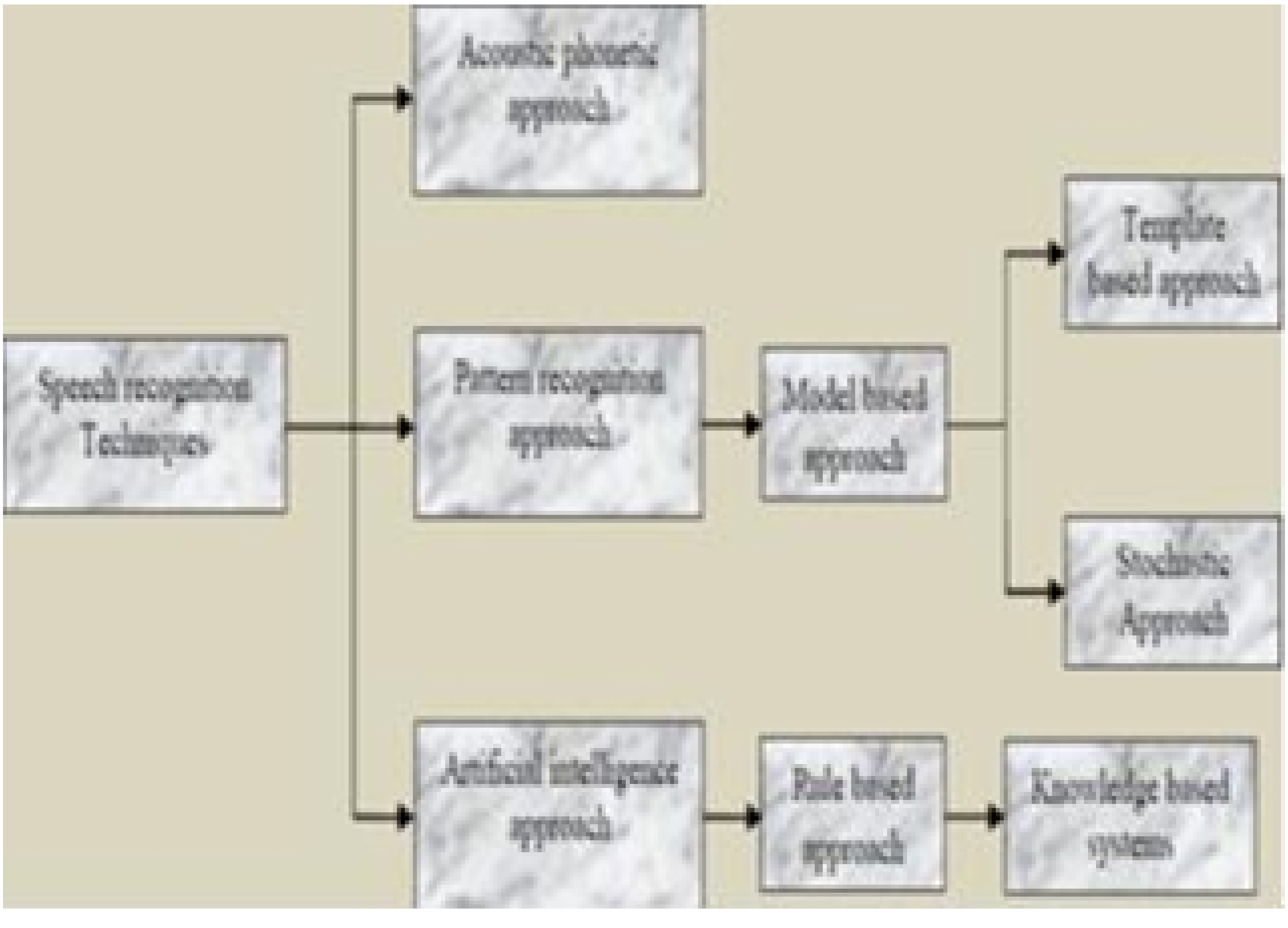
Acoustic phonetic approach
Speech recognition Techniques
Pattern recognition approach
Model based approach
Template based approach
Stochastic Approach
Artificial intelligence approach
Rule based approach
Knowledge based systems

Figure 3: Speech Classification.

blind source separation nonnegative matrix factorization (NMF) technique for speech enhancement can be achieved by incorporating a time correlation term in the objective function to regulate the time-varying gain coefficients of noise [26, 27].

A speech enhancement technique is presented in [28, 29]. The speech enhancement method comprises: estimating the speaker's direction using an input signal, generating directional information indicative of the estimated direction; detecting the speaker's speech based on the directional estimation; and enhancing the speaker's speech utilizing the directional information derived from the speech detection results [30].

3 Research Objectives

The objective of the research is to present spectral approaches suitable for single-channel speech enhancement in multi-source environments, aimed at improving the quality and intelligibility of degraded speech.

This research analyzes voice quality evaluation to ascertain the extent of speech compression and the degree of contaminant removal. The quality evaluation metric is specifically designed to facilitate the comprehensive examination of speech quality through perceptual evaluation (PESQ) during the entire process. This research and observations [8] employed the Perceptual Evaluation of Speech Quality (PESQ), as recommended by ITU-T based theory and simulation. The data will be analyzed and quantified in the subsequent section.

This study examines a two-stage noise reduction technique designed to diminish background noise in single microphone recordings with low signal-to-noise ratios, utilizing perfect binary masking and the Wiener filter. Our proposed system has two phases. Initially, a Wiener filter with an improved signal-to-noise ratio is employed for background noise reduction in noisy speech. Secondly, IBM is computed in each time-frequency channel by utilizing the pre-processed speech from the initial stage and aligning the time-frequency channels to a predetermined threshold T to minimize residual noise. Time-frequency channels that match the threshold requirement are preserved, while all other channels are meticulously muted.

Furthermore, it becomes increasingly difficult when acoustic waves emanate from sources with analogous spectral properties. This algorithm will effectively resolve critical challenges in speech processing applications without compromising speech quality and intelligibility.

This research illustrates the effective application of several speech processing technologies functioning in loud situations.

4 System Model

Speech enhancement presents a formidable challenge in noisy environments, particularly in multisource settings, because to the variability in quality and characteristics of noise over time and between different applications. Consequently, it is challenging to identify effective single-channel speech enhancement algorithms that perform reliably in diverse practical conditions, resulting in significant degradation of voice quality and intelligibility.

In single-channel algorithms, the reduction of noise (enhancement of quality) may occur at the expense of speech distortion (diminution of clarity). Consequently, it is challenging to implement both measures concurrently. This investigation aims to develop a single-channel speech enhancement algorithm that minimizes noise while preserving speech coherence with little distortion. Various measures can be employed to evaluate the implementation of the speech enhancement framework. These are available in two formats: objective and abstract. A variety of approaches exist for target tests, such as the PESQ measure, the IS, or the addition of SNR levels. Abstract testing can also be employed for evaluations conducted with the assistance of human evaluators and is widely regarded as the most effective way for measuring the performance of speech enhancement algorithms [22].

The process commences with the framing mechanism. Every 10 milliseconds, a Hann-windowed frame of 30 milliseconds is retrieved for audio signal processing. The discrete Fourier transform (DFT) of each audio frame is initially transformed into a block of the logarithmic magnitude spectrum or truncated spectrum of autocorrelation. The logarithmic spectrum represents a frequency transformation of the actual spectrum, specifically the inverse Fourier transform. A truncated spectrum is a refined representation of the original logarithmic spectrum, retaining just the lower-order coefficients. We choose to retain the initial J = 50 spectral coefficients at the 44100 Hz sampling rate utilized for this study, adhering to the standard guideline of truncating the spectrum to a length less than the anticipated pitch period. Three SNR levels—-5, 0, and 5 dB—were employed to assess performance as shown in figure 4.

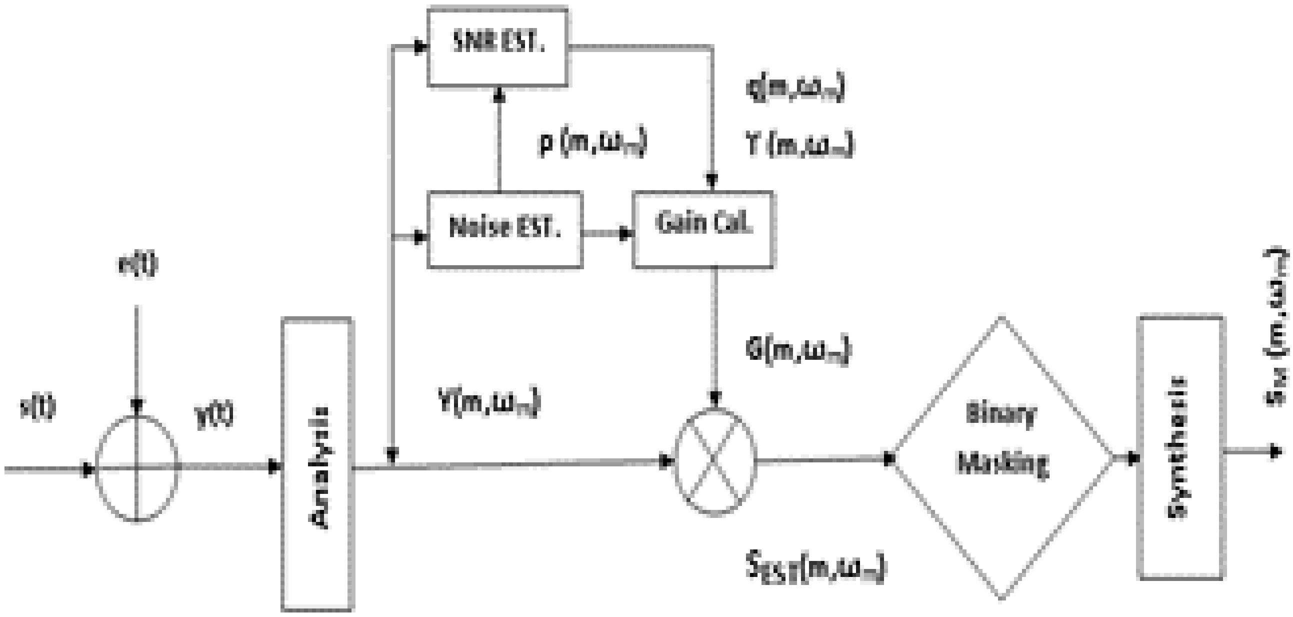

Figure 4: Proposed Model for the Scheme.

4.1 Mathematical Modeling for Noise Reduction Systems

In classical terminology, the equation representing noisy speech in any context is as follows:

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

where $s(t)$ denotes the original noise-free speech, while $e(t)$ signifies the noise itself. Let $Y(m, \omega_m)$, $S(m, \omega_m)$ and $E(m, \omega_m)$ denote the ω_m spectral component of frame m of the noisy speech signal $y(t)$. Both noise and speech exhibit non-stationary characteristics; yet, they are presumed to be stationary over brief intervals (9–32 ms). Therefore, it is presumed that both entities exhibit quasi-stationary characteristics. The spectral gain is integrated into the calculation of two fundamental SNR evaluations, posteriori and a priori, represented by:

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

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

where $E\{.\}$ denotes the expectation operator, and $\sigma_s(m, \omega_m)$ and $\sigma_e(m, \omega_m)$ represent the a posteriori and a priori signal-to-noise ratios, respectively. In real-time applications, the Power Spectral Density (PSD) of clean speech $|S(m, \omega_m)|^2$ and the noise $|E(m, \omega_m)|^2$ remain undetermined, as only the noisy speech is available. The power spectral density of noise is computed via speech gaps utilizing the conventional recursive relation:

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

where, α denotes the smoothing factor and $\hat{\sigma}^2(m-1,\omega_m)$ represents the estimate derived from the current frame. Both the signal-to-noise ratios (SNRs) are calculated as:

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

$$\xi_{\text{PRIO}}^{\text{DD}}(m,\omega_m) = \beta\frac{|G(m-1,\omega_m)\cdot Y(m,\omega_m)|^2}{\sigma_e^2(m,\omega_m)} + (1-\beta)\max\{\gamma_{\text{IMMEDIATE}}(m,\omega_m)-1,0\} \tag{6}$$

Where $\xi_{\text{PRIO}}^{\text{DD}}(m,\omega_m)$ denotes the computation of a priori signal-to-noise ratio (SNR) via the decision-direct (DD) method. DD is computationally efficient and demonstrates significant efficacy in noise reduction assertions. Nonetheless, in this procedure, the a priori SNR follows the configuration of the instantaneous SNR, resulting in a one-frame deferral. The noisy speech $Y(m,\omega_m)$ is processed by multiplying it with the Wiener filter gain function, resulting in:

$$S_{\text{EST}}(m,\omega_m) = Y(m,\omega_m)\cdot G^{\text{DD}}(m,\omega_m) \tag{7}$$

The square root of the Wiener gain function is calculated as specified by the equation:

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

To mitigate residual noise, the ratio of the estimated magnitude spectrum to the clean speech ($|S(m,\omega_m)|/|S_{\text{EST}}(m,\omega_m)|$) is evaluated against a predetermined threshold T. T-F units that comply with the constraint, specifically ($|S(m,\omega_m)|/|S_{\text{EST}}(m,\omega_m)|) \geq T$, are maintained, while those that contravene the constraints are diminished. The adjusted magnitude spectrum $S_M(m,\omega_m)$ is computed as follows:

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

A variety of objective metrics established in the literature for evaluating the efficacy of noise reduction technologies encompass PESQ-MOS and segmental signal-to-noise ratio (SNR_{SEG}) [25]. The PESQ-MOS metric has been tested to yield a strong association with MOS, producing scores ranging from 1 to 5, where a higher number signifies superior speech quality. Similarly, SNR_{SEG} is a widely utilized objective metric that demonstrates optimal correlation in noise reduction. SNR_{SEG} is delineated as follows:

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

where $S(m,\omega_m)$ and $\hat{S}(m,\omega_m)$ represent the frames of clean and estimated speech, respectively. To eliminate non-speech frames, each frame was subjected to a threshold of 0dB as the lower limit and -35dB as the upper limit. Consequently, ITU-T Recommendation P.835 is employed to assess speech distortion and residual noise. The P.835 measure is developed by correlating fundamental objective measures to create a composite measure [10] as shown in figure 5.

$$C_{\text{sig}} = 3.093 - 1.029\cdot SL + 0.603\cdot SP - 0.009\cdot SWS \tag{11}$$

$$C_{\text{bak}} = 1.634 + 0.478\cdot SP - 0.007\cdot SWS + 0.063\cdot S\text{SNR}_{\text{SEG}} \tag{12}$$

5 Results and Analysis

Due to temporal variations, several tools and procedures have been devised, and it is indicated that thorough testing of the suggested solution is essential to ascertain its superiority over the existing

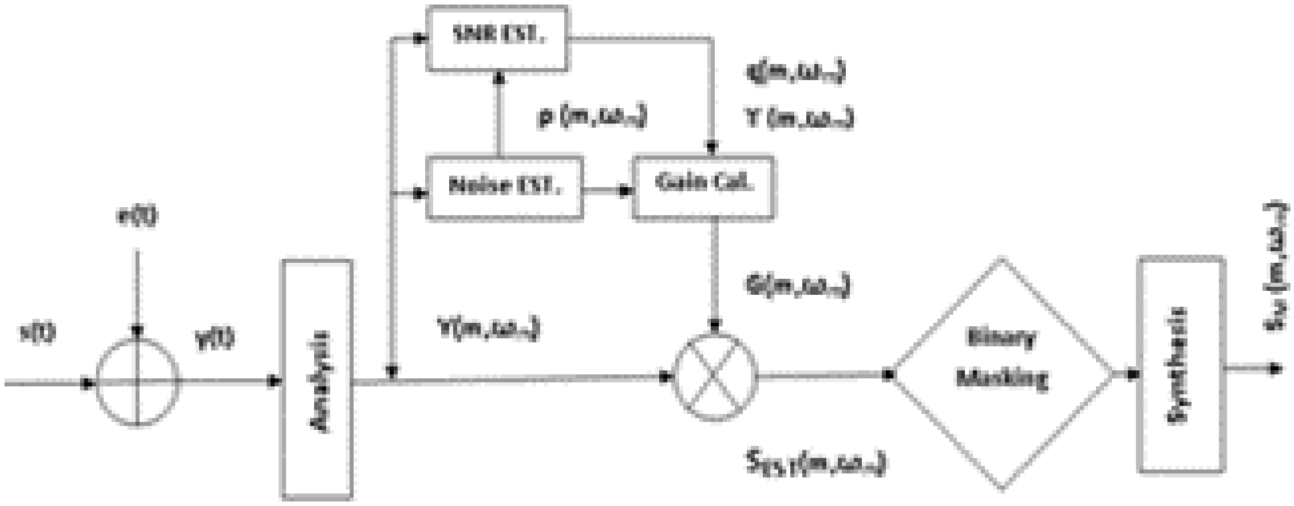

Figure 5: Mechanism for Noise Reduction.

answer. Optimal results can be achieved by the application of appropriate testing methodologies. The testing process is challenging, and in this segment of the research, we examined the novel approach of a binary mask. The proposed approach considers two testing parameters: augmentation of speech intelligibility and the quality of synthesized speech, aimed at improving overall quality and throughput.

Objective assessment and precise measurement values are employed to evaluate and validate the performances and to estimate the computations of the specified speech enhancement algorithms using compression algorithms and techniques. The PESQ method is selected as an ITU-T recommendation, superseding the obsolete ITU-T recommendation, which was grossly insufficient for its intended purpose and remains unsuitable for eliminating distortion caused by undesirable frequencies. The [20] theory encompasses the relevant theoretical material pertaining to PESQ. The PESQ scores run from -0.5 to 4.5, but the output range adheres to the MOS, or mean opinion score, which begins at 1.0 and culminates at 4.5. Aggregate data suggests advancement, and conversely. The primary and most prevalent method for assessing the efficacy of speech enhancement approaches and implementations, as illustrated by suitable flow charts, is the objective measure grounded in Signal-to-Noise Ratio (SNR). The mean ratios, derived from the entire non-stationary signal, and the rapid fluctuations of speech signals indicate that the SNR-based estimation not only precisely represents speech quality for compression but also demonstrates a robust correlation with speech quality. The Signal-to-Noise Ratio (SNR) should be assessed in small packets of minor components, followed by the estimation of their mean for the identification of segmental or fragmental SNR ($\mathrm{SNR}_{\mathrm{Seg}}$) [12].

The performances and results of the specified algorithms and the foundation of the approaches' development are presented in Table 1 appropriately. The proposed algorithm and methods demonstrate a significant and discernible improvement in PESQ phenomena across all signal-to-noise ratio estimates and noise conditions, even in the presence of undesired frequencies that cause signal distortion and contamination. The 5 dB street noise NRPCA (PESQ=1.15) exhibits the greatest enhancement, whilst the 5 dB babbling noise (PESQ=0.42) demonstrates the least improvement. The exhibition hall yielded the greatest PESQ ratings at noise levels of 0 dB, 5 dB, and 10 dB for analysis as shown in figure 6. ΔPESQ values are 2.86, 3.13, and 3.31, respectively. The PESQ score increased from 2.02 with LMMSE to 2.72 with the proposed approach ($\Delta\mathrm{PESQ}_{\mathrm{street}}$=0.71) in street noise at 0dB. The PESQ score increased from 2.20 with subspace to 2.69 with the suggested approach ($\Delta\mathrm{PESQ}_{\mathrm{babble}}$=0.49) for 5dB bubble noise. The PESQ score increased from 2.39 with NRPCA to 3.31 with the proposed approach ($\Delta\mathrm{PESQ}_{\mathrm{Exhibition\ hall}}$=0.92) at 10 dB.

The most significant enhancement is 10 dB at the airport ($\Delta\mathrm{BAK}_{\mathrm{airport}}$=3.35), whilst the least is 0 dB, corresponding to babble noise ($\Delta\mathrm{BAK}_{\mathrm{babble}}$=2.68). The SIG segmentation scores increased from 4.28 in the Exhibition Hall utilizing the suggested algorithm ($\Delta\mathrm{SIG}_{\mathrm{Exhibition\ Hall}} = 4.28$) at 10dB, whilst the lowest score recorded is 3.65 in street means ($\Delta\mathrm{SIG}_{\mathrm{street}} = 3.65$) at 0dB.

The proposed voice enhancement in Table 4 exhibits time complexity, signifying the duration

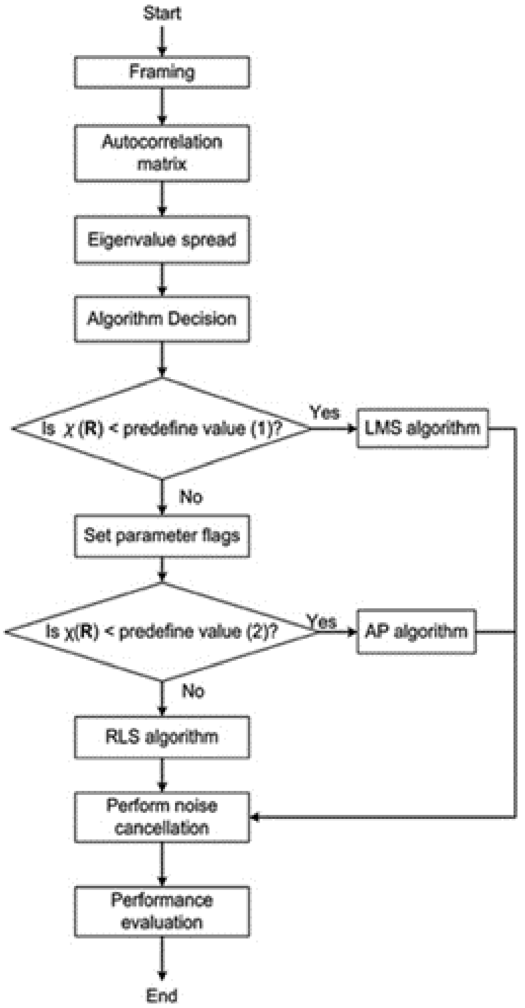

Figure 6: Flowchart for the Technique.

required for the algorithm to execute fully. The database's 30 combinations, exhibiting a total runtime of 150 seconds, were analyzed. Table 4 illustrates that techniques devoid of training execute significantly more rapidly than those that incorporate it (DNN). For example, NRPCA necessitates 20.5 seconds to process a single speech with a specified SNR level (assuming 0dB). The runtime for 30 utterances is calculated as 20.5 multiplied by 30, equating to 615 seconds. The cumulative duration needed by NRPCA to produce increased voice utterances is 3075 seconds over five SNR levels (0dB, 5dB, 10dB) when the RT is multiplied by a factor of 5. The proposed approach necessitates 10 seconds to execute fully for a single syllable at a specific SNR level. The proposed method exhibits a temporal complexity of 1500 seconds, signifying its rapid convergence relative to alternative approaches.

The spectrogram samples demonstrate the effectiveness of all processing algorithms and enhancement strategies to improve signals. Upon examining the spectrograms of the signal processing algorithms for the intended objective, the vocal utterance was obscured by the babbling noise at a 0dB SNR level. The proposed technique excels in mitigating or eradicating background noise and unwanted frequencies while purifying speech from polluted frequencies.

6 Conclusion

This study proposed a two-stage noise reduction system. The initial step involves applying a Wiener filter to the loud voice to enhance the signal-to-noise ratio for background noise attenuation. To diminish residual noise, IBM is computed in each time-frequency channel utilizing the pre-processed speech from the initial stage and aligning the time-frequency channels to the predetermined threshold T. This approach will resolve critical challenges in speech processing applications without compromising

Table 1: PESQ evaluations in diverse acoustic situations

Type of Noise	Approaches	0 dB	5 dB	10 dB
Airport Noise	LMMSE	2.05	2.34	2.63
	Subspace	1.91	2.33	2.61
	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

Table 2: Statistical Analysis of Residual Noise and Speech Distortion Using ANOVA

Residual Noise Assessment (BAK)					
Type of Noise	SNR (dB)	Un-P	Wiener filter	Stage One	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

speech quality and intelligibility. Another benefit of this research is the efficient utilization of various speech processing applications functioning in noisy situations.

Objective assessment and precise measurement values are employed to evaluate and validate the performance of the specified voice enhancement methods using compression algorithms and techniques. The PESQ approach is selected as an ITU-T recommendation, superseding the outdated version and prior ITU-T recommendations, which are inadequate for the intended purpose and fail to effectively mitigate distortion caused by undesirable frequencies. This investigation utilized the PESQ, as recommended by ITU-T based theory and simulated observations.

Time-frequency channels are pivotal in their domain and have eliminated previous limits. The

Table 3: Analysis of Speech Distortion (SIG)

Type of Noise	SNR (dB)	Un-P	Wiener	Stage One	Proposed 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

Table 4: Time Complexity Analysis

Method	Time per utterance (s)	Total time for 30 utterances (s)
LMMSE	15.2	2280
Subspace	18.7	2805
NRPCA	20.5	3075
Proposed Method	10.0	1500

overall contribution of speech signals to signal-to-noise ratios (0dB, 5dB, and 10dB) utilizing babble and street noise generators is a fundamental aspect of the suggested method. Nonetheless, advancements in speech have been documented, demonstrating enhanced intelligibility and clarity, even at low signal-to-noise ratios, yielding superior results and outcomes.

A significant deficiency is the inadequate implementation of the project methodology, which suggests that selecting the appropriate strategy could yield superior results and increased throughput. Proactive planning is essential to ensure the project is completed within the designated timeframe and budget. Various types can be employed to assess the advancement of voice enhancement and compression systems.

References

[1] Ding, H. (2011). *Speech improvement in the transform domain* (Doctoral dissertation, Nanyang Technological University).

[2] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... & Wellekens, C. (2007). A review of automatic speech recognition and speech variability. *Speech Communication*, *49*(10-11), 763-786.

[3] Loizou, P. C. (2007). *Speech enhancement: principles and application.* CRC Press.

[4] Ephraim, Y., & Malah, D. (1984). Speech improvement with a lowest mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109-1121.

[5] Cappé, O. (1994). Mitigation of the musical noise phenomena via 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 methodology for speech denoising utilizing multitaper techniques. In *2006 14th European Signal Processing Conference* (pp. 1-5). IEEE.

[7] Virag, N. (1999). Single-channel speech augmentation utilizing the masking characteristics of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, *7*(2), 126-137.

[8] Hu, Y., & Loizou, P. C. (2004). Speech improvement utilizing wavelet thresholding of the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, *12*(1), 59-67.

[9] Martin, R. (1994). Minimum statistics-based spectral subtraction. *Signal Processing*, *8*(6), 1-8.

[10] Sorqvist, P., Handel, P., & Ottersten, B. (1997, April). Kalman filtering for minimal distortion speech improvement in mobile communications. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 1219-1222). IEEE.

[11] Vary, P., & Eurasip, M. (1985). Noise suppression by spectral magnitude estimation: mechanisms and theoretical limitations. *Signal Processing*, *8*(4), 387-400.

[12] Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, *70*(9), 1055-1096.

[13] Reidy, P. F. (2015). An evaluation of spectral estimating techniques for the study of sibilant fricatives. *The Journal of the Acoustical Society of America*, *137*(4), EL248-EL254.

[14] Händel, P. (2006). Analysis of power spectral density errors in spectral subtraction-based voice enhancement techniques. *EURASIP Journal on Advances in Signal Processing*, *2007*, 1-9.

[15] Anusuya, M. A., & Katti, S. K. (2010). A review on machine speech recognition. *arXiv preprint arXiv:1001.2267.*

[16] Kadir, K. A. (2010). Recognition of Human Speech Utilizing q-Bernstein Polynomials. *International Journal of Computer Applications*, *2*(5), 22-28.

[17] Reddy, D. R. (1976). A Review of Machine Speech Recognition. *Proceedings of the IEEE*, *64*(4), 501-531.

[18] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). An analysis of speech recognition methodologies. *International Journal of Computer Applications*, *10*(3), 16-24.

[19] Shinde, R. B., & Pawar, V. P. (2012). An evaluation of the acoustic phonetic methodology 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 noisy environments as a function of the number of spectral channels: A comparative analysis 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 employing deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14-22.

[22] Deng, L. (2004). Transitioning dynamic system models for speech articulation and acoustics. In *Mathematical Foundations of Speech and Language Processing* (pp. 115-133). Springer.

[23] Deng, L., & Yu, D. (2007, April). Utilization of differential cepstra as acoustic characteristics 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). Assessment of single-channel speech augmentation algorithms utilizing objective quality and intelligibility metrics. 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). Deep Neural Network training with a traditional gain function for single-channel voice augmentation 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). An examination of noise power spectral density estimators for single-channel speech augmentation. 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 via non-negative matrix factorization and its use in voice augmentation. 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). *Speech enhancement utilizing directional information.* U.S. Patent No. 10,546,593. Washington, DC: U.S. Patent and Trademark Office.

[29] Bryan, N. J., & Iyengar, V. (2020). *Speech enhancement for an electronic apparatus.* U.S. Patent No. 10,535,362. Washington, DC: U.S. Patent and Trademark Office.

[30] Cho, J., Cui, W., & Lee, S. (2020). *Method and apparatus for speech enhancement.* U.S. Patent No. 10,529,360. Washington, DC: U.S. Patent and Trademark Office.