A robust speech enhancement method in noisy environments

Original Scientific Paper

Nesrine Abajaddi
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
n.abajaddi@uhp.ac.ma

Badia Mounir
LAPSSII Laboratory,
High School of Technology,
Cadi Ayyad University, Safi, Morocco
mounirbadia2014@gmail.com

Youssef Elfahm
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
y.elfahm@uhp.ac.ma

Abdelmajid Farchi
IMMII Laboratory,
Faculty of Sciences & Technics,
Hassan First University, Settat, Morocco
abdelmajid.farchi1@gmail.com

Abstract *– Speech enhancement aims to eliminate or reduce undesirable noises and distortions, this processing should keep features of the speech to enhance the quality and intelligibility of degraded speech signals. In this study, we investigated a combined approach using single-frequency filtering (SFF) and a modified spectral subtraction method to enhance single-channel speech. The SFF method involves dividing the speech signal into uniform subband envelopes, and then performing spectral over-subtraction on each envelope. A smoothing parameter, determined by the a-posteriori signal-to-noise ratio (SNR), is used to estimate and update the noise without the need for explicitly detecting silence. To evaluate the performance of our algorithm, we employed objective measures such as segmental SNR (segSNR), extended short-term objective intelligibility (ESTOI), and perceptual evaluation of speech quality (PESQ). We tested our algorithm with various types of noise at different SNR levels and achieved results ranging from 4.24 to 15.41 for segSNR, 0.57 to 0.97 for ESTOI, and 2.18 to 4.45 for PESQ. Compared to other standard and existing speech enhancement methods, our algorithm produces better results and performs well in reducing undesirable noises.*

Keywords*: speech enhancement, single frequency filtering, spectral subtraction, envelopes*

1. INTRODUCTION

Speech enhancement is an active area of research that aims to improve the quality of degraded speech and preferably its intelligibility. One of the most significant tasks of speech enhancement is to reduce or remove noise that has degraded speech quality, and this is an active area of research [1-3]. Noise reduction techniques are used in many applications such as mobile phones [4], speech recognition [5], teleconferencing systems [6], voice over internet protocol (VoIP) [7], and hearing aids. Most speech enhancement systems use a single microphone for economic reasons, even though better results can be obtained using multiple microphones [8]. The field of single-channel speech enhancement continues to be a significant area of research due to its simplicity and computational efficiency. These systems, which are based on a single microphone, employ adaptive filtering techniques to reduce the impact of noisy regions in speech signals that have a low SNR while preserving those with a high SNR. Within this context, the short-term spectral amplitude plays a crucial role in preserving speech quality and intelligibility, compared to phase information [9]. Furthermore, the speech enhancement algorithms can be classified as follows [10]: i) spectral subtraction algorithms, ii) statistical model algorithms, iii) subspace algorithms, and iv) machine learning algorithms.

Spectral subtraction algorithms, developed in the late 1970s, are effective and widely used in the spectral domain [11]. This technique involves subtracting the estimated noise from the noisy speech, assuming that the noise is additive and uncorrelated with the clean speech signal. However, this approach is susceptible to generating musical noise, which can be bothersome to listeners. To reduce this side effect, several techniques have been proposed, including applying a half-wave rectifier and setting all negative values to zero [12-14]. Another approach is to use an over-reduction factor and a spectral floor factor, but speech may be distorted and consequently lead to a loss of intelligibility [15, 16]. Due to the varying spectral distribution of the signal and noise across different frequency ranges, researchers have directed their attention toward subband processing methods. A widely recognized technique in the field of audio signal enhancement is the multiband spectral subtraction (MBSS) method [17]. This approach involves decomposing the signal into subbands to exploit the spectral information and apply different noise treatments in each subband, considering the varying impact of noise on the speech spectrum. To further enhance the performance of the multiband spectral subtraction approach, the authors in [18] propose a technique called MBSS_CBRS (multiband spectral subtraction with critical band rate scaling). This technique is designed to align with the characteristics of the auditory system, aiming to approximate the benefits of human perception and to improve the effectiveness of speech enhancement algorithms. Other studies concentrate on speech production features. In the study described in [19], the authors specifically investigate the harmonic properties of vowels (SS_HP) and utilize the sigmoid function to empirically determine the values for over-subtraction and the spectral reservation factor. Furthermore, researchers have recognized that speech can be considered as an amplitude-modulated signal, leading them to explore speech enhancement techniques in the modulation domain. In [20] the authors specifically employed the coherent harmonic demodulation technique (SE_CHD) to get the subband signals. This approach relies on a prior signal-to-noise ratio (SNR) and utilizes a gain function derived using the minimum mean square error (MMSE) approach.

Traditionally, speech enhancement algorithms commonly rely on the short-term Fourier transform (STFT) to estimate the short-term spectrum of a signal [17, 19]. This involves dividing the signal into subband signals through consecutive windowing or filter bank operations [18, 20]. A novel approach known as single-frequency filtering (SFF) has recently been introduced. [21]. This technique offers high spectral and temporal resolution and eliminates the effects of windowing through filtering. By employing filtering at the maximum frequency of $f_s/2$, SFF can capture both amplitude and phase information at each frequency. SFF has been explored in various applications, including voice activity detection [21], epoch extraction [22], hyper-nasality assessment [23], and dysarthria evaluation [24, 25]. This approach is gaining popularity for segmenting speech into multiple frequency bands due to its exceptional time-frequency resolution.

In this work, our contribution includes the development of a new algorithm combining the recent approach called SFF and the modified spectral subtraction method to enhance the quality and intelligibility of degraded speech signals. The integration of the SFF approach with noise estimation using the SNR is based on the identification of segments with high SNRs. In practice, the power of the noise tends to be lower near zero-bandwidth resonator of the single frequency, while the power of the speech signal, particularly if present, is relatively high. As a result, windows exhibiting high SNR will appear at different times for various frequencies. Therefore, we used the SFF for calculating the envelopes which help to reduce or eliminate unwanted noise concentrated around one or more specific frequencies. For each envelope, we estimated the noise from previous speech frames and applied a smoothing parameter to balance noise reduction and speech quality preservation. The experiments were conducted on various types of real-world noise, including car, train, restaurant, airport, and street noises, as well as machine-generated white Gaussian noise, to evaluate the performance of our proposed algorithm. The results demonstrate that our method outperforms the previously mentioned existing methods [17-20].

The paper is organized into three main sections. In Section 2, an analysis and synthesis of the single-frequency approach, envelope manipulation, modified spectral subtraction, and noise estimation are covered. Section 3 provides detailed outcomes of the conducted experiments, highlighting the performance of the proposed method. Finally, Section 4 serves as the conclusion, summarizing the key findings and implications of the research.

2. PROPOSED SPEECH ENHANCEMENT MODEL

This section aims to explore the potential of employing the single-frequency filtering (SFF) approach and modified spectral subtraction to enhance the degraded speech. The proposed method involves three main steps (Fig.1). First, the SFF analysis is performed to generate spectral envelopes at specific frequencies of interest. Then, the modified spectral subtraction is applied at each frame for each envelope. Finally, the SFF synthesis combines the processed envelopes to reconstruct the original speech. Our objective is to develop an algorithm that utilizes validated blocks to ensure the high quality and intelligibility of the speech signal. In order to achieve this, we have verified that the SFF analysis-synthesis method does not have any negative impact on the quality and/or intelligibility of the speech signal. This step aims to ensure that the SFF processing does not introduce any undesired effects or degradation to the processed signal. The second step involves determining whether the enhancement should be focused on the envelope or the phase of the signal. Choosing the optimal element is important for improving the quality and intelligibility of the speech signal. Finally, it is essential to define the suitable parameters for the spectral subtraction method to reduce the noise in the speech signal and achieve the best results in terms of improving its quality and intelligibility.

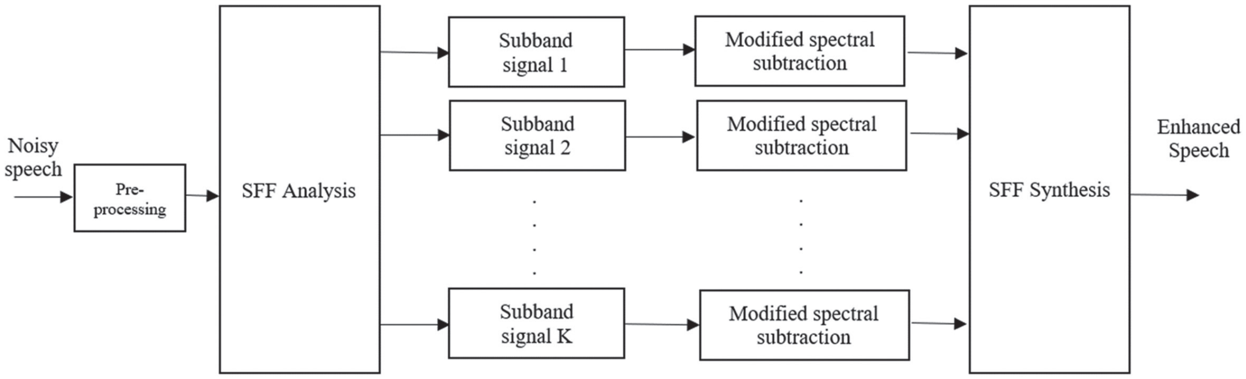

Fig. 1. Block diagram of the proposed speech enhancement algorithm

2.1. SFF ANALYSIS-SYNTHESIS APPROACH

The SFF approach is employed to achieve subband decomposition of the speech signal, allowing us to obtain the amplitude of the envelopes at a selected frequency for each instant. This technique helps to eliminate the block processing effect which can reduce the performance of the algorithms when using the Short-Time Fourier Transform (STFT) for speech enhancement. The analysis and synthesis blocks of the SFF technique are represented in Fig. 2. The amplitude envelope of the signal at any desired frequency is obtained by:

- shifting the frequency signal of the signal x[n] at frequency f_k using Eq. (1) for n=1.2...N and k=0.1…. K-1, where N and K represent the number of samples and the number of frequencies, respectively.

$$\tilde{x}[k,n] = x[n]e^{-j\frac{2\pi\bar{f}_k}{f_s}n} \qquad (1)$$

- The shifted signal is then passed through a single pole resonator at the highest frequency $f_s/2$, where f_s is the sampling frequency. The filter function transfer is given by Eq. (2) and the filtered output is given by Eq. (3) where $y[k, n]$ is a complex number that can be expressed in polar form as given by Eq. (4).

$$H[z] = \frac{1}{1+rz^{-1}} \qquad (2)$$

$$y[k,n] = -ry[k,n-1] + \tilde{x}[k,n] \qquad (3)$$

$$y[k,n] = v[k,n]e^{j\phi[k,n]} \qquad (4)$$

where the amplitude envelope $v[k, n]$ and the phase $\phi[k, n]$ of the subband signal $y[k, n]$ are defined by Eq. (5) and (6), respectively.

$$v[k,n] = \sqrt{y_r^2[k,n] + y_i^2[k,n]} \qquad (5)$$

$$\phi[k,n] = tan^{-1}\left(\frac{y_i[k,n]}{y_r[k,n]}\right) \qquad (6)$$

- The subband signal $y[k, n]$ can be reconstructed from the amplitude envelope $v[k, n]$ and phase $\phi[k, n]$ by applying Eq. (4) for the single frequency filtering synthesis. The shifted output $y[k, n]$ is shifted back to the original frequency using Eq. (7):

$$z[k,n] = y[k,n]e^{j\frac{2\pi\bar{f}_k}{f_s}n} \qquad (7)$$

- The reconstructed signal is obtained by summing the outputs $z[k, n]$ and dividing by the number of frequencies K using Eq. (8).

$$\hat{x}[k,n] = \frac{1}{K}\Re\left\{\sum_{k=0}^{K-1} z[k,n]\right\} \qquad (8)$$

- The reconstructed signal and the original signal are combined using the following equation [26]:

$$\hat{x}[k,n] = \sum_{a=0}^{\infty}(r)^{aK}x[n-aK] \qquad (9)$$

Where $K=(f_s/2)/\Delta f$, and r is less than 1 to ensure the stability of the filter.

The performance of the proposed algorithm is influenced by the two major parameters r and K. The primary objective of this research is to enhance degraded speech. To achieve this goal, a value of r=0.99 was selected, which offers higher temporal and spectral resolution. This resolution property enables a more precise and accurate analysis of the degraded speech, facilitating effective application of enhancement techniques [26]. Additionally, the impact of the number of frequencies K on speech quality and intelligibility was evaluated in this study. Clean speech signals were analyzed and synthesized at different K values while maintaining a sampling rate of 16 kHz. The resulting speech was then assessed using PESQ and ESTOI metrics. The findings presented in Table 1 show that reducing the frequency step size significantly improves speech quality and intelligibility. However, when using a frequency interval (Δf) of 50 Hz, we obtain 160 envelopes, whereas using a smaller Δf of 20 Hz results in 400 envelopes. Therefore, reducing Δf to 20 Hz can lead to improved quality and intelligibility of the processed speech. However, this reduction also increases the value of K and the number of envelopes, resulting in longer computation times. It is worth noting that utilizing the SFF method with Δf=20 Hz allows for enhancing speech without introducing any distortion.

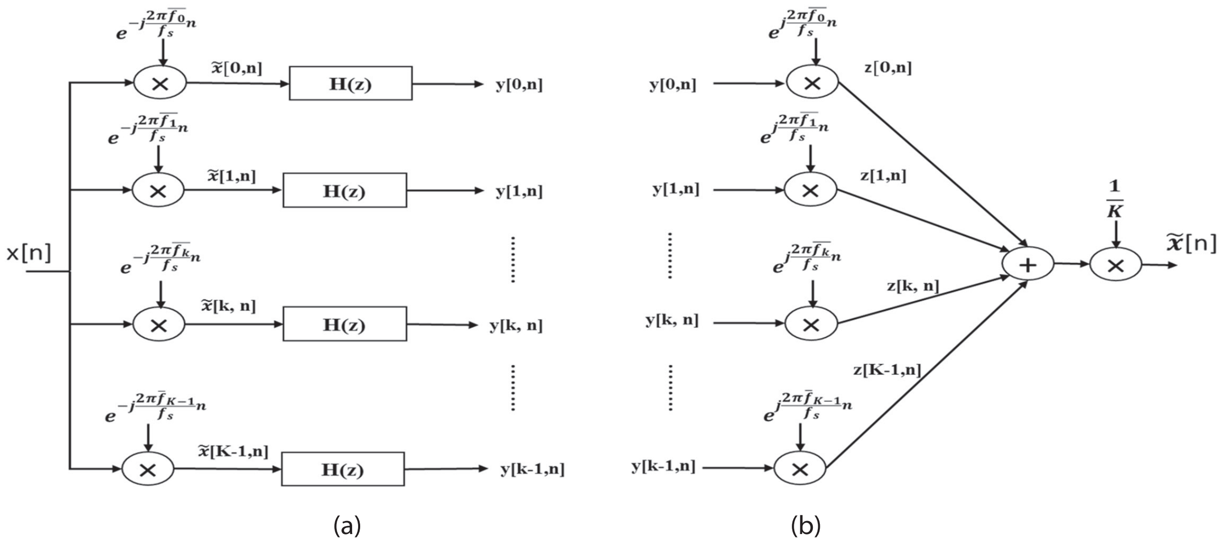
x[n]
$e^{-j\frac{2\pi \bar{f}_0}{f_s}n}$
$\tilde{x}[0,n]$
H(z)
y[0,n]
$e^{-j\frac{2\pi \bar{f}_1}{f_s}n}$
$\tilde{x}[1,n]$
y[1,n]
$e^{-j\frac{2\pi \bar{f}_k}{f_s}n}$
$\tilde{x}[k, n]$
y[k, n]
$e^{-j\frac{2\pi \bar{f}_{K-1}}{f_s}n}$
$\tilde{x}[K-1,n]$
y[k-1,n]
$e^{j\frac{2\pi \bar{f}_0}{f_s}n}$
z[0,n]
z[1,n]
z[k, n]
z[K-1,n]
$\frac{1}{K}$
$\tilde{x}[n]$
(a)
(b)

Fig. 2. (a) SFF analysis, and (b) SFF synthesis

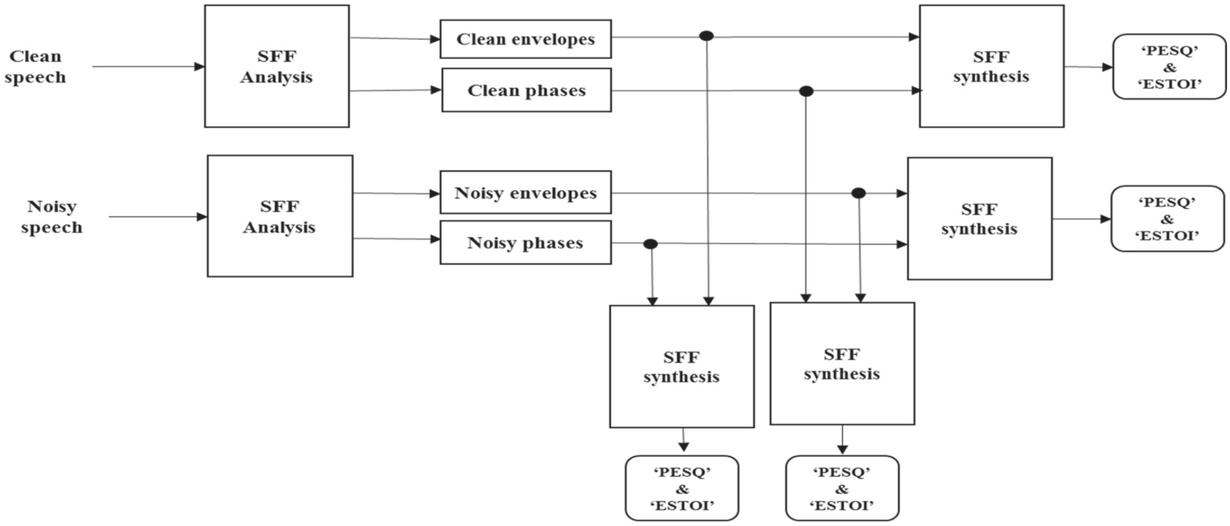

Fig. 3. Concept block diagram for the combinations of the clean and noisy temporal envelope with noisy and clean phase

Table 1. ESTOI and PESQ between original and SFF synthesized signals for different values of *Δf*.

Δf	100	50	20	10	5
ESTOI	0.987	0.999	1	1	1
PESQ	4.17	4.359	4.64	4.64	4.64

The following step is to determine whether the SFF amplitude envelope or the SFF phase should be enhanced to improve speech quality and intelligibility.

2.2. THE TEMPORAL ENVELOPE

The significance of the temporal envelope and its application in speech enhancement has been extensively explored in various studies [27-29]. The objective of this step is to verify and validate the hypothesis that "enhancing the temporal envelope of noisy speech can considerably enhance speech quality". To achieve this, we combined the clean envelope with the noisy phase and vice versa using samples from the TIMIT database, at 10 dB for white noise, with the envelope and phase calculated using the SFF technique. Fig.3 presents the adopted scheme for constructing all the combinations, while Table 2 presents the corresponding values of ESTOI and PESQ. The results indicate that the envelope plays a significant role in improving the quality and intelligibility of speech. Therefore, we recommend modifying the amplitude of the SFF envelope to enhance speech in noisy environments.

Table 2. The average PESQ and ESTOI values were computed for all combinations of TIMIT speech degraded by 10 dB white noise with *Δf*=20 Hz.

Combinations	ESTOI	PESQ
Clean envelopes - Clean phases	1	4,64
Clean envelopes - Noisy phases	0.91	1.89
Noisy envelopes - Clean phases	0.86	1.76
Noisy envelopes - Noisy phases	0.85	1.25

2.3. MODIFIED ENVELOPE SPECTRAL SUBTRACTION

Based on the results obtained from PESQ and ESTOI, we have evaluated the effectiveness of SFF analysis-synthesis and its associated envelope in enhancing speech quality and intelligibility. The findings indicate also that enhancing the temporal envelope, which is calculated using the SFF approach, can be utilized for speech enhancement, as the phase component has a negligible impact on human intelligibility. Considering its favorable performance across different noise conditions and its low computational complexity, we propose implementing the spectral subtraction method for each frame of each envelope to improve both quality and intelligibility of the speech signal. The noise is estimated using an adaptive technique, and the over-reduction factor is adjusted for each frame of each envelope Eq. (5). The equation provided below illustrates the spectrum of the enhanced k^{th} envelope.

$$|\hat{s}_k(m,n)|^2 = \begin{cases} |v_k(m,n)|^2 - \eta(m,n)\left|\hat{d}_k(m,n)\right|^2, if & \left[\frac{\hat{d}_k(m,n)}{v_k(m,n)}\right]^2 < \frac{1}{\eta(m,n)} \\ \beta|v_k(m,n)|^2 & , else \end{cases} \quad (10)$$

Where v_k (m, n), $\hat{d}_k$ (m, n), and $\hat{s}_k$ (m, n) represent the noisy envelope, estimated noise, and estimated enhanced envelope, respectively, for the n^{th} FFT transform of the m^{th} frame in the k^{th} envelope. The parameter β is the spectral floor factor that typically ranges between 0 and 1, and it is used to prevent the estimated negative speech spectrum in each envelope. To determine the over-reduction factor $\eta(m, n)$ for the envelope k at frame m, we use the segmental signal-to-noise ratio (segSNR) as follows:

$$\eta(m,n) = \begin{cases} 5 & , segSNR(m,n) < -5dB \\ \eta_0 - \frac{3}{20} segSNR(m,n), & -5dB \le segSNR(m,n) \le 20dB \\ 1 & , segSNR(m,n) > 20dB \end{cases} \quad (11)$$

The segSNR is calculated using the formula:

$$segSNR(m,n) = \frac{\sum_{m=0}^{N_k-1} |v_k(m,n)|^2}{\sum_{m=0}^{N_k-1} \left|\hat{d}_k(m,n)\right|^2} \quad (12)$$

where N_k represents the number of frames of the k^{th} envelope. The over-reduction parameter $\eta(m, n)$ is determined based on η_0 which represents the value of $segSNR$ at 0 dB and controls the noise subtraction level for each envelope at every frame.

Accurate noise estimation is crucial for improving speech, as an incorrect estimate can result in residual noise or distorted speech. Traditional methods use voice activity detection (VAD) to estimate and adapt the noise spectra during speech silent periods, but this approach is not effective in real-time and noisy environments [30, 31]. A common technique for estimating noise is recursive averaging, where the noise spectrum is calculated by taking a weighted average of previous noise estimates and the present noisy envelope spectrum, as described in [32, 33]. The weight assigned to each estimate varies based on the a-posteriori signal-to-noise ratio (SNR) of each frequency. In our proposed technique, we independently estimate and update the noise spectrum for each frame in each envelope. Therefore, the noise spectrum estimation for each envelope is given by,

$$\left|\hat{d}_k(m,n)\right|^2 = \delta(m,n)\left|\hat{d}_k(m-1,n)\right|^2 + \{1-\delta(m,n)\}\,|v_k(m,n)|^2 \quad (13)$$

where δ is the smoothing parameter. In the recursive averaging technique, δ is chosen as a sigmoid function that depends on the a-posteriori SNR:

$$\delta(m,n) = \frac{1}{1+e^{[-a\{SNR(m,n)-T\}]}} \quad (14)$$

In this case, the variation of the noise is determined by the parameter "a" in the sigmoid function Eq. (14). The value of "a" ranges from 1 to 6 while keeping the parameter "T" constant. The parameter "T" in Eq. (14) represents the center offset of the transition curve in the sigmoid function, typically falling within the range of 3 to 5.

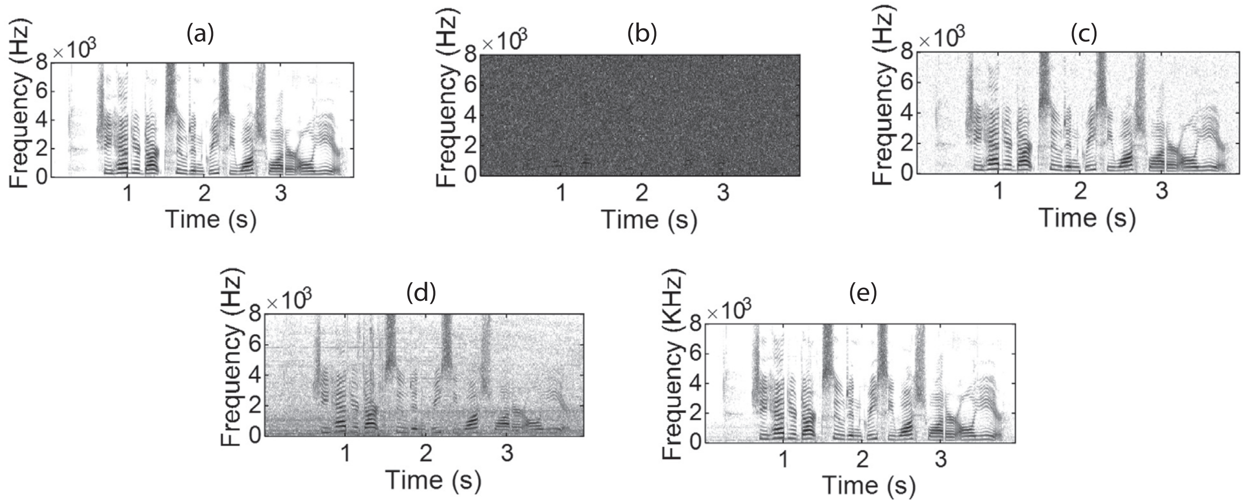

Fig. 4. Speech spectrograms for: (**a**) clean speech; (**b-d**) speech degraded by white and car noises, respectively, at 5 dB SNR; (**c-e**) the associated enhanced speech

3. EXPERIMENTAL RESULTS

In this section, we will present the results and performance evaluation of our proposed speech enhancement algorithm. Additionally, we will compare the results obtained using MATLAB software with other existing methods [17-20]. The evaluation was conducted on various types of noise, including real-world noises commonly encountered in daily life such as car, train, restaurant, airport, and street noises, as well as machine-generated white Gaussian noise. It is important to note that each type of noise exhibits a unique time-frequency distribution in speech.

To evaluate the performance of our speech enhancement algorithm, the noisy speech was obtained from the TIMIT corpus and downsampled to 8 kHz. The envelopes were calculated using the SFF with an envelope spacing of Δf=20 Hz, which provided good performance of speech quality and intelligibility as shown in Table 1. For each envelope, we applied a hamming window with a frame duration of 20 ms and a 70% overlap. The noise was estimated continuously and adaptively using Eq. (13) and the sigmoid function Eq. (14), with values of 'a' and 'T' set to 4 and 5, respectively. The over-subtraction factor $\eta(m, n)$ was computed for each envelope. Moreover, we fixed the spectral floor parameter β at a value of 0.03 [17, 18].

3.1. CORPUS

The TIMIT database, developed by the Massachusetts Institute of Technology with support from the US government, is a valuable resource for automatic speech recognition and other speech processing applications [34]. It comprises a vast collection of speech sounds and associated data. This extensive database includes recordings from 630 speakers representing different regions of the United States, each uttering 10 different phrases. It also provides transcriptions and annotations corresponding to the recorded speech. TIMIT has become a fundamental tool in the study of speech recognition and other related technologies, contributing significantly to the advancement of the field of speech processing.

3.2. PERFORMANCE EVALUATION

To assess the effectiveness of our proposed speech enhancement algorithm, we utilized objective quality and intelligibility measurement tests, including segmental signal-to-noise ratio (segSNR), extended short-term objective intelligibility (ESTOI) [35], and perceptual evaluation of speech quality (PESQ) [36]. segSNR is frequently used to detect speech distortion, as it is more precise in identifying speech distortion compared to overall SNR. Higher values for segSNR indicate lower levels of speech distortion. ESTOI is a measure of speech intelligibility that considers the accuracy and timing of phoneme recognition. PESQ is a commonly used objective measure of speech quality that compares and predicts the perceived quality of speech signals using a reference signal. Higher scores for both PESQ and ESTOI typically suggest better speech quality and intelligibility. These measures have strong correlations with subjective listening tests, making them valuable tools for assessing speech enhancement algorithms.

3.3. RESULTS AND DISCUSSIONS

We evaluated the performance of our proposed method (PM) on degraded speech corrupted by different types of noise such as white Gaussian, car, restaurant, train, street, and airport noises. We tested the PM at various SNR levels including -5, 0, 5, and 10 dB. To assess the effectiveness of the PM, we measured three parameters: segSNR, ESTOI, and PESQ. The average segSNR values obtained by the PM for different types of noise and SNR levels are presented in Table 3. For the -5 dB, 0 dB, 5 dB, and 10 dB SNR levels, the segSNR values were 5.30, 8.76, 10.07, and 13.64, respectively. It is worth noting that these segSNR values are consistently positive and exhibit an increasing trend as the SNR improves. These findings demonstrate that the proposed method performs well in terms of distortion across various noise types. In terms of intelligibility, the PM achieved average ESTOI values of 0.66, 0.76, 0.86, and 0.93 for -5 dB, 0 dB, 5 dB, and 10 dB SNR levels, respectively (Table 4). Specifically, ESTOI ranges from 0.57 to 0.74 for negative SNR values and from 0.65 to 0.97 for positive SNR values. These results provide that the PM algorithm significantly enhances the intelligibility of speech. To evaluate the speech quality, we calculated the average PESQ values, which ranged from 2.18 to 4.45 across all SNR levels (Table 5). These results confirm that our PM performs well in terms of speech quality. Fig. 4, 5, and 6 illustrate the speech spectrograms of the noisy speech at 5 dB SNR alongside the enhanced speech obtained using the PM. These figures provide a visual representation of the improvements achieved by our method.

The results of the score comparison in terms of segSNR, ESTOI, and PESQ for various types of noise at different levels of SNR, between our PM and MBSS [17], MBSS_CBRS [18], SS_HP [19], and SS_CHD [20], are presented in Tables 3, 4, and 5, respectively. The findings clearly demonstrate that our PM outperforms the other methods in terms of segSNR, indicating its effectiveness in removing background noise while preserving speech components, regardless of whether the SNR is negative or positive. The PM achieves also higher ESTOI scores, indicating improved speech intelligibility across all SNR levels. Furthermore, the PM obtains the highest PESQ score in all conditions, indicating the preservation of speech quality.

Based on these results, it can be concluded that our PM significantly enhances speech quality and intelligibility with minimal distortion compared to the methods defined previously. The effectiveness of the PM is supported by the utilization of the SFF to calculate the temporal envelopes with high-frequency resolution at

20 Hz, guided by PESQ and ESTOI scores. Moreover, our proposed method estimates the noise recursively from previous speech frames for each envelope and applies a smoothing parameter to achieve a balance between noise reduction and preservation of speech quality.

Furthermore, our PM algorithm was compared to recent algorithms that utilize deep learning techniques for tasks such as estimating parameters (e.g., tuning factor of the Wiener filter [37]) or extracting features (e.g., multi-frequency cepstral coefficients [38]).The comparative analysis revealed that our PM algorithm outperforms these approaches in terms of both speech quality and intelligibility, as depicted in Fig. 7 and 8. Moreover, an added advantage of our PM algorithm is that it does not necessitate training data. This characteristic reduces its complexity and simplifies its implementation, making it a more practical and accessible solution for noise reduction and speech enhancement tasks.

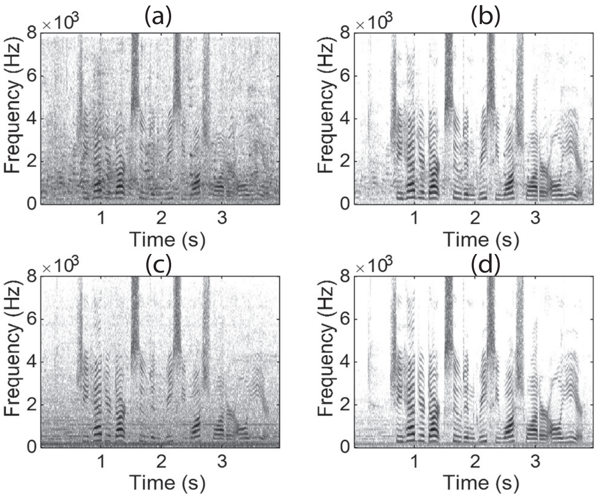

Fig. 5. Speech spectrograms for: (**a-c**) speech degraded by restaurant and train noises, respectively, at 5 dB SNR; (**b-d**) the associated enhanced speech

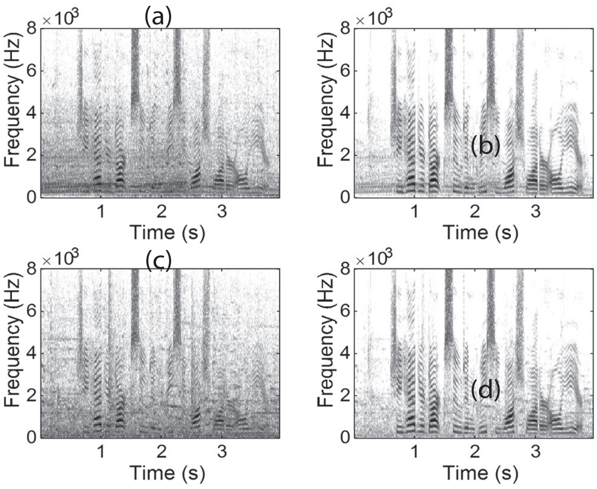

Fig. 6. Speech spectrograms for: (**a-c**) speech degraded by street and airport noises, respectively, at 5 dB SNR; (**b-d**) the associated enhanced speech

Table 3. Average segmental signal-to-noise ratio ($segSNR$) of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	segSNR			
		-5	0	5	10
White	Noisy	-7.59	-5.16	-1.92	1.72
	MBSS	1.64	4.07	10	13.40
	MBSS_CBRS	2.83	5.9	8.63	11.77
	SS_HP	4.57	7.66	5.97	4.31
	SE_CHD	0.90	1.09	1.59	1.81
	PM	**6.58**	**9.67**	**12.01**	**15.41**
Car	Noisy	-7.50	-5.02	-1.78	1.80
	MBSS	0.64	2.68	6.86	9.39
	MBSS_CBRS	2.72	4.23	9.05	12.05
	SS_HP	3.90	6.45	4.86	3.73
	SE_CHD	0.26	0.58	1.01	1.90
	PM	**5.91**	**8.46**	**11.06**	**14.06**
Restaurant	Noisy	-7.09	-4.63	-1.43	2.2
	MBSS	0.43	2.52	5.84	9.29
	MBSS_CBRS	1.79	2.14	7.25	9.96
	SS_HP	3.69	6.17	4.86	3.48
	SE_CHD	0.37	0.60	0.83	2.29
	PM	**5.70**	**8.18**	**9.26**	**11.97**
Train	Noisy	-7.46	-5.06	-1.77	1.85
	MBSS	0.21	3.88	5.53	9.35
	MBSS_CBRS	2.38	5.12	7.6	11.59
	SS_HP	1.69	3.92	2.96	2.16
	SE_CHD	0.28	0.41	0.53	2.10
	PM	**4.39**	**7.13**	**9.61**	**13.6**
Street	Noisy	-6.51	-3.9	-0.73	2.86
	MBSS	0.43	1.71	5.39	9.22
	MBSS_CBRS	2.27	9.81	7.02	11.36
	SS_HP	1.98	6.98	4.65	3.29
	SE_CHD	0.19	0.31	0.43	2.9
	PM	**4.28**	**11.82**	**9.03**	**13.37**
Airport	Noisy	-7.45	-5.02	-1.81	1.8
	MBSS	1.49	3.78	6.53	8.81
	MBSS_CBRS	2.91	5.26	7.43	11.44
	SS_HP	2.08	3.75	3.04	2.64
	SE_CHD	0.31	0.64	0.8	1.8
	PM	**4.92**	**7.27**	**9.44**	**13.45**

Table 4. Average extended short-term objective intelligibility (ESTOI) results of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	ESTOI			
		-5	0	5	10
White	Noisy	0.25	0.39	0.55	0.68
	MBSS	0.39	0.45	0.6	0.68
	MBSS_CBRS	0.4	0.58	0.69	0.71
	SS_HP	0.49	0.6	0.7	0.75
	SE_CHD	0.18	0.15	0.21	0.24
	PM	**0.57**	**0.65**	**0.79**	**0.8**
Car	Noisy	0.2	0.39	0.52	0.64
	MBSS	0.39	0.48	0.57	0.71
	MBSS_CBRS	0.43	0.51	0.62	0.78
	SS_HP	0.59	0.63	0.79	0.82
	SE_CHD	0.18	0.12	0.15	0.21
	PM	**0.69**	**0.73**	**0.89**	**0.92**
Restaurant	Noisy	0.21	0.37	0.53	0.67
	MBSS	0.38	0.53	0.69	0.71
	MBSS_CBRS	0.43	0.67	0.72	0.77
	SS_HP	0.59	0.74	0.78	0.86
	SE_CHD	0.11	0.12	0.15	0.23
	PM	**0.69**	**0.84**	**0.88**	**0.96**
Train	Noisy	0.34	0.46	0.57	0.66
	MBSS	0.49	0.53	0.61	0.78
	MBSS_CBRS	0.56	0.61	0.69	0.83
	SS_HP	0.66	0.71	0.76	0.89
	SE_CHD	0.13	0.15	0.17	0.28
	PM	**0.74**	**0.79**	**0.84**	**0.97**
Street	Noisy	0.27	0.4	0.54	0.66
	MBSS	0.35	0.49	0.67	0.75
	MBSS_CBRS	0.41	0.54	0.73	0.79
	SS_HP	0.52	0.69	0.79	0.86
	SE_CHD	0.14	0.15	0.16	0.27
	PM	**0.6**	**0.77**	**0.87**	**0.94**
Airport	Noisy	0.24	0.38	0.52	0.64
	MBSS	0.36	0.47	0.61	0.73
	MBSS_CBRS	0.41	0.59	0.75	0.81
	SS_HP	0.59	0.68	0.8	0.89
	SE_CHD	0.18	0.2	0.24	0.28
	PM	**0.67**	**0.76**	**0.88**	**0.97**

Table 5. Average perceptual evaluation of speech quality (PESQ) results of enhanced speech signals from the TIMIT database at -5,0, 5, 10 dB

Noise type	Enhancement methods	PESQ			
		-5	0	5	10
White	Noisy	1.11	1.21	1.39	1.67
	MBSS	1.31	1.53	1.82	2.22
	MBSS_CBRS	1.49	1.65	1.97	2.3
	SS_HP	1.87	2.19	2.58	2.95
	SE_CHD	0.45	0.49	0.59	0.68
	PM	**2.18**	**3.69**	**4.08**	**4.45**
Car	Noisy	1.1	1.18	1.32	1.55
	MBSS	1.31	1.49	1.98	2.25
	MBSS_CBRS	1.29	1.43	1.62	2.43
	SS_HP	2.19	2.33	2.72	3.1
	SE_CHD	0.47	0.59	0.7	0.7
	PM	**2.35**	**3.34**	**4.73**	**4.11**
Restaurant	Noisy	1.12	1.19	1.3	1.5
	MBSS	1.61	1.84	2.06	2.32
	MBSS_CBRS	1.47	1.63	1.95	2.32
	SS_HP	2.34	2.25	2.64	3.21
	SE_CHD	0.56	0.67	0.69	0.7
	PM	**2.94**	**3.85**	**4.02**	**4.18**
Train	Noisy	1.15	1.27	1.47	1.77
	MBSS	1.31	1.51	1.69	2.14
	MBSS_CBRS	1.4	1.52	1.61	2.14
	SS_HP	1.98	2.05	2.43	2.82
	SE_CHD	0.51	0.5	0.65	0.42
	PM	**2.58**	**3.65**	**4.03**	**4.42**
Street	Noisy	1.17	1.27	1.45	1.72
	MBSS	1.321	1.59	1.93	2.24
	MBSS_CBRS	1.227	1.4	2.01	2.3
	SS_HP	2.314	2.56	2.46	3.62
	SE_CHD	0.47	0.48	0.52	0.61
	PM	**3.09**	**3.76**	**4.06**	**4.22**
Airport	Noisy	1.12	1.2	1.34	1.57
	MBSS	1.31	1.79	2.1	2.42
	MBSS_CBRS	1.21	1.46	2.11	2.42
	SS_HP	2.09	2.34	2.83	3.68
	SE_CHD	0.4	0.35	0.5	0.63
	PM	**2.89**	**3.94**	**3.43**	**4.28**

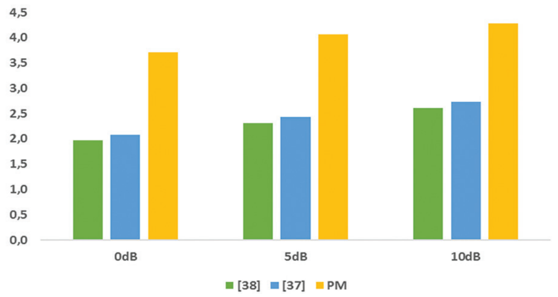

Fig. 7. Average performance comparison between the proposed method and methods used in [37] and [38] for all noises using PESQ.

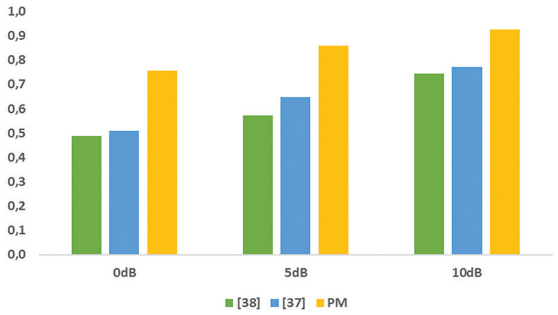

Fig. 8. Average performance comparison between the proposed method and methods used in [37] and [38] for all noises using ESTOI.

4. CONCLUSION

The proposed method aims to enhance the quality and intelligibility of speech degraded by noises. It utilizes the single-frequency filtering approach and modified spectral subtraction to effectively eliminate unwanted noise while minimizing distortion and preserving essential speech characteristics. The research demonstrates the effectiveness of this approach in improving speech quality and intelligibility under various noise types and different signal-to-noise ratio (SNR) levels. The performance of our algorithm is encouraging, and it can be suitable to meet the requirements and challenges of complex environments, by adjusting only the over-subtraction factor. It is important to note that our proposed method has a limitation related to the over-subtraction process. This aspect highlights an area for improvement in our approach. In our future work, we will primarily concentrate on addressing this limitation by incorporating voice characteristics into the algorithm. By considering the specific features of the speech segments (voiced or unvoiced), we aim to enhance the noise reduction's accuracy and effectiveness, thereby improving our method's overall performance.

5. REFERENCES:

[1] R. C. Hendriks, T. Gerkmann, J. Jensen, “DFt-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art”, Synthesis Lectures on Speech and Audio Processing, Vol. 11, Springer, 2013.

[2] K. K. Wójcicki, P. C. Loizou, “Channel selection in the modulation domain for improved speech intelligibility in noise”, The Journal of the Acoustical Society of America, Vol. 131, No. 4, 2012, pp. 2904-2913.

[3] M. I. Khattak, N. Saleem, J. Gao, E. Verdu, J. P. Fuente, “Regularized sparse features for noisy speech enhancement using deep neural networks”, Computers and Electrical Engineering, Vol. 100, 2022.

[4] E. Mabande, F. Kuech, A. Niederleitner, A. Lombard, “Towards robust close-talking microphone arrays for noise reduction in mobile phones”, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20-25 March 2016.

[5] D. Li, Y. Gao, C. Zhu, Q. Wang, R. Wang, “Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy”, Sensors, Vol. 23, No. 4, 2023.

[6] R. Bendoumia, M. T. Betina, A. Oulahcene, A. Guessoum, “Extended subband decorrelation version of feedback normalized adaptive filtering algorithm for acoustic noise reduction”, Applied Acoustics, Vol. 179, 2021.

[7] S. H. Han, S. Jeong, H. Yang, J. Kim, W. Ryu, M. Hahn, “Noise reduction for VoIP speech codecs using modified wiener filter”, Advances and Innovations in Systems, Computing Sciences and Software Engineering, Springer, 2007, pp. 393-397.

[8] D. O’Shaughnessy, "Speech communications: Human and machine", Second Edition, 1999.

[9] J. S. Lim, A. V. Oppenheim, “Enhancement and Bandwidth Compression of Noisy Speech”, Proceedings of IEEE, Vol. 67, No. 12, 1979, pp. 1586-1604.

[10] P. C. Loizou, "Speech Enhancement: Theory and Practice", CRC Press, 2013.

[11] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 27, No. 2, 1979, pp. 113-120.

[12] M. M. Sondhi, C. E. Schmidt, L. R. Rabiner, “Improving the Quality of a Noisy Speech Signal”, The Bell System Technical Journal, Vol. 60, No. 8, 1981, pp. 1847-1859.

[13] J. H. L. Hansen, M. A. Clements, "Use of objective speech quality measures in selecting effective spectral estimation techniques for speech enhancement", Proceedings of the Midwest Symposium on Circuits and Systems, Champaign, IL, USA, 14-16 August 1989.

[14] H. Xu, Z. H. Tan, P. Dalsgaard, B. Lindberg, "Spectral subtraction with full-wave rectification and likelihood controlled instantaneous noise estimation for Robust speech recognition", Proceedings of the 8th International Conference on Spoken Language Processing, 2004.

[15] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, USA, 2-4 April 1979.

[16] N. Abajaddi, B. Mounir, L. Elmaazouzi, I. Mounir, A. Farchi, "Speech Spectral Subtraction in Modulation Domain", Lecture Notes in Networks and Systems, Springer, 2022.

[17] S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13-17 May 2002.

[18] N. Upadhyay, A. Karmakar, "Single-Channel Speech Enhancement Using Critical-Band Rate Scale Based Improved Multi-Band Spectral Subtraction", Journal of Signal and Information Processing, Vol. 04, No. 03, 2013.

[19] C. T. Lu, K. F. Tseng, Y. Y. Chen, L. L. Wang, C. L. Lei, "Speech enhancement using spectral subtraction algorithm with over-subtraction and reservation factors adapted by harmonic properties", Proceedings of the International Conference on Applied System Innovation, Okinawa, Japan, 26-30 May 2016.

[20] S. Samui, I. Chakrabarti, S. K. Ghosh, "Speech enhancement based on modulation domain processing using coherent harmonic demodulation technique", Electronics Letters, Vol. 53, No. 24, 2017.

[21] G. Aneeja, B. Yegnanarayana, "Single Frequency Filtering Approach for Discriminating Speech and Nonspeech", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 4, 2015, pp. 705–717.

[22] S. R. Kadiri, B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach", Speech Communication, Vol. 86, 2017, pp. 52-63.

[23] M. H. Javid, K. Gurugubelli, A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4-8 May 2020.

[24] K. Gurugubelli, A. K. Vuppala, "Perceptually Enhanced Single Frequency Filtering for Dysarthric Speech Detection and Intelligibility Assessment", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12-17 May 2019.

[25] K. Gurugubelli, A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment", Speech Communication, Vol. 121, 2020. pp. 1-15.

[26] N. Chennupati, S. R. Kadiri, B. Yegnanarayana, "Spectral and temporal manipulations of SFF envelopes for enhancement of speech intelligibility in noise", Computer Speech & Language, Vol. 54, 2019, pp. 86-105.

[27] A. Wiinberg, J. Zaar, T. Dau, "Effects of Expanding Envelope Fluctuations on Consonant Perception in Hearing-Impaired Listeners", Trends in Hearing, Vol. 22, 2018.

[28] T. Langhans, H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, France, 3-5 May 1982.

[29] M. Koutsogiannaki, H. Francois, K. Choo, E. Oh, "Real-time modulation enhancement of temporal envelopes for increasing speech intelligibility", Proceedings of the Annual Conference of the International Speech Communication Association, 2017.

[30] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5, 2001, pp. 504-512.

[31] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, 2003, pp. 466-475.

[32] L. Lin, W. H. Holmes, E. Ambikairajah, "Speech denoising using perceptual modification of Wiener filtering", Electronics Letters, Vol. 38, No. 23, 2002.

[33] L. Lin, W. H. Holmes, E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", Electronics Letters, Vol. 39, No. 9, 2003.

[34] J. Garofolo et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download", Philadelphia Linguistic Data Consortium, 1993.

[35] J. Jensen, C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, No. 11, 2016, pp. 2009-2022.

[36] ITU, "ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", 2000.

[37] A. Garg, "Speech enhancement using long short term memory with trained speech features and adaptive wiener filter", Multimedia Tools and Applications, Vol. 82, No. 3, 2023, pp. 3647-3675.

[38] A. Garg, O. P. Sahu, "Deep Convolutional Neural Network-based Speech Signal Enhancement Using Extensive Speech Features", International Journal of Computational Methods, Vol. 19, No. 8, 2022.