A Mathematical Approach to Speech Enhancement for Speech Recognition and Speaker Identification Systems

Rohun Nisa, Haweez Showkat, and Asifa Baba

Abstract In order to cope with acoustic degradation where clean sample of speech, free of interference and noise, prior to recognition stage, and identification–verification system, an efficient recognition and authentication of a particular speaker are necessary. In this paper, an approach for enhancement of speech is implemented using Fourier transform followed by spectral subtractive principle in upgrading speech signal contaminated due to noise. This methodology is employed in efficient recognition system for speech and identification–verification system of speaker as the degraded signal complicates hearing and understanding of speech signal. A Fourier transform approximates and derives spectrum of corrupted speech, and the spectral subtractive algorithm suppresses the amount of noise from noise spectrum to achieve clean signal.

Keywords Additive white Gaussian noise · Fourier transform · Spectral subtraction · Speaker identification · Speech enhancement · Speech recognition · Temporal convolutional neural network · Machine learning

1 Introduction

As we reside in native surroundings filled with noise and disturbance, there is generally unwanted noise associated with the signals particularly speech that hinders the processing of signals in original form. Noise and unwanted interference affect human–human and human–machine communications among varied fields which include degrading the properties of speech involving intelligibility together with

R. Nisa (✉) · H. Showkat · A. Baba
Department of Electronics and Communication Engineering, Islamic University of Science and Technology, Awantipora, Kashmir 192122, India
e-mail: rohunnisa@islamicuniversity.edu.in

H. Showkat
e-mail: haweezsk@gmail.com

A. Baba
e-mail: asibababa@gmail.com

A. Iqbal et al. (eds.), *Renewable Power for Sustainable Growth*, Lecture Notes in Electrical Engineering 723, https://doi.org/10.1007/978-981-33-4080-0_62

quality, identification corresponding to particular speaker, and recognition of speech [1–3]. Noise is generated everywhere, characteristics of which are either known or unknown. The field involving removal of noise and interference out of disrupted speech by incorporating variants of signal processing methodology constitutes speech processing. The different categories that comprise processing of signals include coding, enhancement, recognition, and synthesis particularly of speech. To frame the voice communication comfortable, natural, and practical, digital signal processing techniques are required [4]. Applications of speech communication requiring the noise reduction algorithms include answering machines, freehand communication, hard-of-hearing aids, localized and remote distance telecommunications, mobile and car phones, multiparty conferencing, noisy manufacturing and cockpits, teleconferencing systems, and voice over Internet Protocol (VoIP).

Normally, the word noise describes the undesirable signal that hinders and disrupts the analysis, processing, transmission, and reception of required informative acoustic signal. In order to achieve desirable representation and suppression of impact of noise, it becomes necessary to classify the concerning terminology of noise into respective four subclasses defined as follows: *additive noise* is the interference that gets associated with the signal due to varied sources when transmitted via communication channel, *interfering signals* that arise when multiple speakers are communicating at a time, *reverberation* is the effect of sound that remains after the sound is produced and is particularly due to multipath propagation, and *echo* is the sound reflection that reaches the listener after delay and arises mainly because of mixed link among microphones and loudspeakers. To take into account the corresponding problems mentioned, numerous speech signal processing techniques are employed including *reduction in noise or enhancement of speech*, *separation of source and speaker*, *de-reverberation of speech*, and *cancelation and suppression of echo* [5].

The signal analyzed through microphone is usually a representation of pure signal of speech with undesirable noise effect, resulting in corrupted signal and main challenge being to deal with background noise that causes degradation of signal of interest. The foremost consideration of suppression algorithms of noise is thus to recover and restore clean speech in original form given the superimposed signal to achieve the following essential goals: enhancing perceptual speech quality corrupted due to noise, improving objective performance criteria including intelligibility and signal-to-noise ratio (S/N or SNR) and enhancing the robustness of remaining applications of speech processing techniques comprising echo suppression and cancelation, coding of speech, recognition and synthesis of speech, particularly to noise [6].

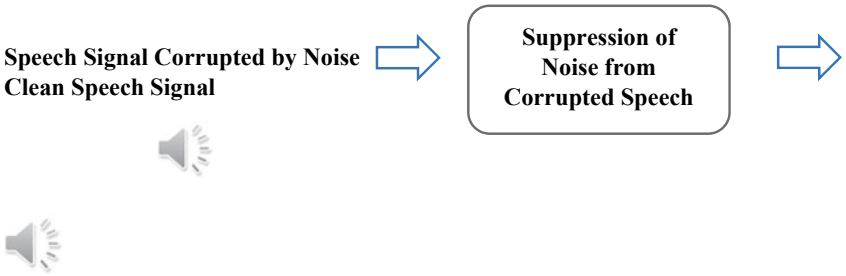

Fig. 1 Speech enhancement system

The presence of unwanted background noise in the acoustic signal severely impacts the functionality and execution involving speaker identification–verification (SIV) process that results in the reduction of recognition rate. Such systems are usually employed and incorporated before any SIV systems for enhancing the working of such systems to achieve better results, as depicted in Fig. 1.

2 Methodology Employed

Various speech enhancement methods are employed for reducing the noise in speech signal, among which spectral subtractive method is popular and commonly used method in real-world applications [1, 5]. Other traditional speech enhancement algorithms comprise statistical model-based methodologies, subspacing procedures, and binary masking principles. The spectral subtraction method including Fourier transform domain relies on eliminating the spectrum related to noise from noise corrupted speech in magnitude form obtained via Fourier transform, giving the enhanced clean speech signal as output [6]. The work on noise reduction techniques started with the novel contribution including two patents by Schroeder [7, 8] who put forward an application of analog method for spectral magnitude subtraction algorithm. After that, Boll [9] through his explanatory work specified the digital domain representation of spectral subtraction algorithm. Lim and Oppenheim [10] in form of their milestone effort represented the noise suppression problem by considering the already existing algorithms and forming a comparison. Their work explained the usefulness of reduction and suppression in noise from noise corrupted signal to upgrade the signal intelligibility and quality.

The noise reduction challenges are numerous in quantity. Pertaining to single channel where signal is recorded by one microphone, and of multichannel where signal is recorded by more than one microphone, there is the necessity to derive an optimal solution for removing as much undesirable noise as possible without degrading the standards including quality of speech signal and its intelligibility for purpose of communication. The proposed work presents a combination of Fourier

transform decomposition of noise corrupted signal together with spectral subtraction to enhance speech signal for improvement of speaker identification–verification process.

2.1 Segmentation and Framing of Speech Signal

A speech signal is usually not stationary in real sense, but is typically considered quasi-stationary for short period of time. The main rationale being the glottal system and the features of such system do not change instantly [11]. Particularly for definite units of sound in a language called as Phonemes, the characteristics of speech usually stay unchangeable and are short approx. 5–100 ms time period. As such, application of traditional signal processing techniques becomes practical to be incorporated during short time span. Normally, speech processing is applied by considering very short windows including overlapping followed by analyzing and processing of such windows, referred to as frame. Thus, a speech signal, typically stationary in windows of suppose 20 ms, is partitioned and segmented into frames of 20 ms, corresponding to N samples given as

$$N = t_{fs} f_s \tag{1}$$

where t_{fs} forms the time frame step and f_s comprises frequency of sampling of signal.

Figure 2 depicts the segmentation of speech signal into short window frames. The overlapping of frames is shown with the corresponding first part of frame overlapped with the previous frame and remaining part with the next frame. The time frame step t_{fs} specifies the time duration among the start time of corresponding frame.

The duration from the beginning of new frame up to the end of current frame is referred to as overlap time t_{o}. Following from these considerations, the frame length t_{fl} is represented as

$$t_{\mathrm{fl}} = t_{\mathrm{fs}} + t_{\mathrm{o}} \tag{2}$$

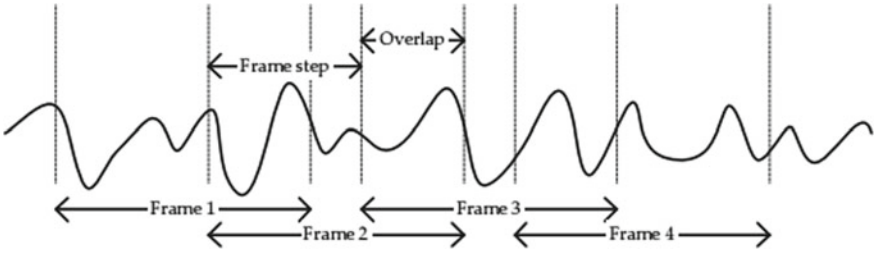

Fig. 2 Segmentation of speech into frames

Thus, the window is of length t_{fl}, which corresponds to $t_{\mathrm{fl}} f_{\mathrm{s}}$ samples.

In this method, frames are taken to be about 25 ms long, and audio file is taken to be of 16 kHz. This corresponds to 0.025 s*16,000 samples/s = 400 samples in length. We are using an overlap of 50% that constitute about 200 samples. So, the first frame will start at 0 instant, second frame will start at 200, third frame will start at 400, etc., indicated by frame1, frame2, and frame3 in the figure.

2.2 Decomposition in Fourier Transform Domain

Considering the quasi-stationary feature of speech for processing, the analysis involving speech is done taking short segmented windows referred frames and applying short time domain of Fourier transform (STFT) on respective individual short segment, yielding Fourier spectrum on corresponding frame [12]. Getting the noise corrupted signal as input is the combination of speech in clean form and corrupted due to additive noise. The model is represented as

$$y[\eta] = x[\eta] + s[\eta] \tag{3}$$

where $y[\eta]$, $x[\eta]$, and $s[\eta]$ represent the sampled noise corrupted signal, pure signal, and additive noise, with the assumption of additive noise having average time domain value of zero, not varying together with speech signal, η being the discrete index of time [13].

Now the STFT of the noise corrupted signal y (η) will thus be represented by

$$Y(\eta, \varpi) = \sum_{l=-\infty}^{\infty} y(l) w(\eta - l) e^{-j2\pi\varpi l/N} \tag{4}$$

where ϖ constitutes the discrete frequency index, N as the duration of frame (in samples), l as the frame number, and $w(\eta)$ as speech analysis function referred to window function. While considering the processing of speech signal, the Hamming window is usually used having duration range of typically 20–40 ms [14]. Windowing is required as the analysis of input signal involves processing of samples that are finite, resulting in discontinuation of respective frames. Such discontinuities among the corresponding frames are eliminated by employing windowing, resulting in smooth end of frames and getting connected accurately to the start of upcoming frame [15].

2.3 Reconstruction of the Signal

To construct the improved clean signal, $x(\eta)$, another transform referred as inverse STFT is applied on modified speech spectrum and continuing with the incorporation of least-squares overlap-add synthesis, depicted as

$$x(\eta) = \frac{1}{W_\theta(\eta)} \sum_{l=-\infty}^{\infty} \left[\left(\frac{1}{N} \sum_{\varpi=0}^{N-1} Y(l, \varpi) e^{\frac{j2\pi\eta\varpi}{N}} \right) w_\tau(l-\eta) \right] \tag{5}$$

where $w_\tau(\eta)$ represents the function referred as synthesis window, with $W_\theta(\eta)$ represented as

$$W_\theta(\eta) = \sum_{l=-\infty}^{\infty} w_\tau^2(l-\eta) \tag{6}$$

Usually, the synthesis window employed is Hanning window, depicted as

$$w_\tau(\eta) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi(\eta+0.5)}{N}\right), & 0 \le \eta \le N \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

2.4 Spectral Subtractive Principle

Spectral subtractive principle forms the practical and useful method that is employed for the suppression of ambient noise from signal. The method relies on the regeneration of spectrum involving magnitude of a signal with background noise associated with signal and subtracting the average noise spectrum approximation obtained from Fourier transform from noise corrupted signal spectrum. The scenarios involving processing of signals at receiver with communication channel are contaminated by noise, and the corrupted signal is usually encountered at the receiver end. For such circumstances, local average impact of noise is considered on spectrum of signal [16]. The addition of additive noise on signal thus raises the average value and variance of magnitude spectrum of a signal, as depicted in Fig. 3.

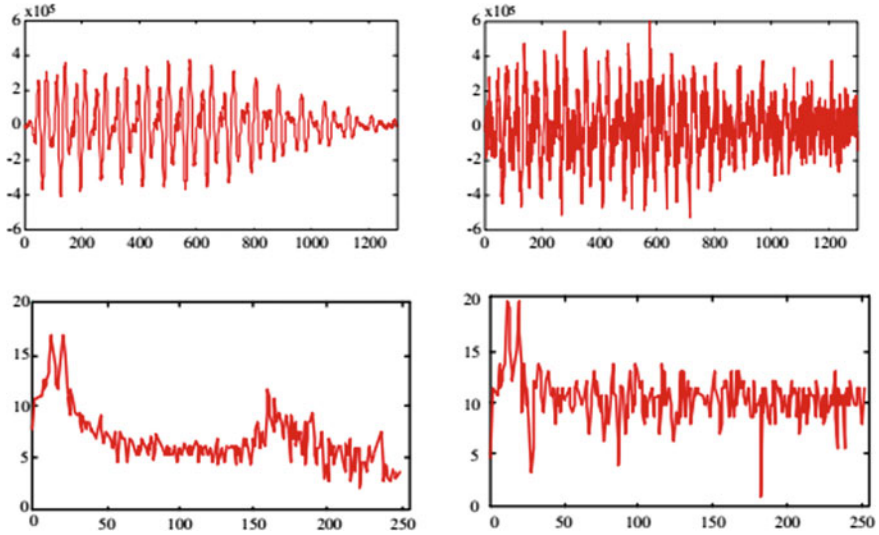

Fig. 3 Impact of noise on signal pertaining to time and frequency domain

Due to variant time characteristics of speech, the signal analysis is achieved and done using frame-by-frame analysis by incorporating short time domain of Fourier transform (STFT) on signal depicted by Eq. 4, illustrated as

$$Y(\eta, \varpi) = X(\eta, \varpi) + S(\eta, \varpi). \tag{8}$$

With the assumption of independent relation among speech signal and background noise, the corresponding magnitude spectrum of corrupted signal $y[\eta]$ is represented without cross terms and depicted as

$$|Y(\varpi)|^2 = |X(\varpi)|^2 + |S(\varpi)|^2 \tag{9}$$

To obtain the spectrum involving improved clean signal, an approximate of corrupted signal spectrum is eliminated out of input signal spectrum, represented as

$$\left|\dot{X}(\varpi)\right|^2 = |Y(\varpi)|^2 - \left|\dot{S}(\varpi)\right|^2 \tag{10}$$

The other application of spectral subtractive principle involves the realization as filter referred as spectral subtractive filter, mathematically represented as product of corrupted spectrum pertaining to speech by noise and the spectral subtractive filter (SSF), depicted as

$$\left|\dot{X}(\varpi)\right|^2 = \left(1 - \frac{\left|\dot{S}(\varpi)\right|^2}{|Y(\varpi)|^2}\right)|Y(\varpi)|^2 \tag{11}$$

$$\left|\dot{X}(\varpi)\right|^2 = \hat{H}^2(\varpi)|Y(\varpi)|^2 \tag{12}$$

where $\hat{H}(\varpi)$ represents the function referred as gain function, related to spectral subtractive filter (SSF) which is considered as filter with zero phase, having the representation of magnitude varying among the range $0 \leq \hat{H}(\varpi) \leq 1$, given as,

$$\hat{H}(\varpi) = \left\{\max\left(0, 1 - \frac{\left|\dot{S}(\varpi)\right|^2}{|Y(\varpi)|^2}\right)\right\}^{1/2} \tag{13}$$

For reconstructing the signal, phase spectrum characteristics of speech are taken into account. The usual method in determining the phase or angle variation of corresponding corrupted speech is relating the angle variation of noise degraded speech to the phase of clean signal obtained after suppression. Thus, the approximation of speech regarding a short segment frame is expressed as

$$\dot{X}(\varpi) = \left|\dot{X}(\varpi)\right| e^{j<Y(\varpi)} \tag{14}$$

$$\dot{X}(\varpi) = \hat{H}(\varpi)Y(\varpi) \tag{15}$$

From this, it follows that an approximate waveform of speech in time domain can be reconstructed using inverse Fourier transform. The sequence followed in performing speech enhancement with Fourier transform and spectral subtraction approach is depicted by flow diagram in Fig. 4.

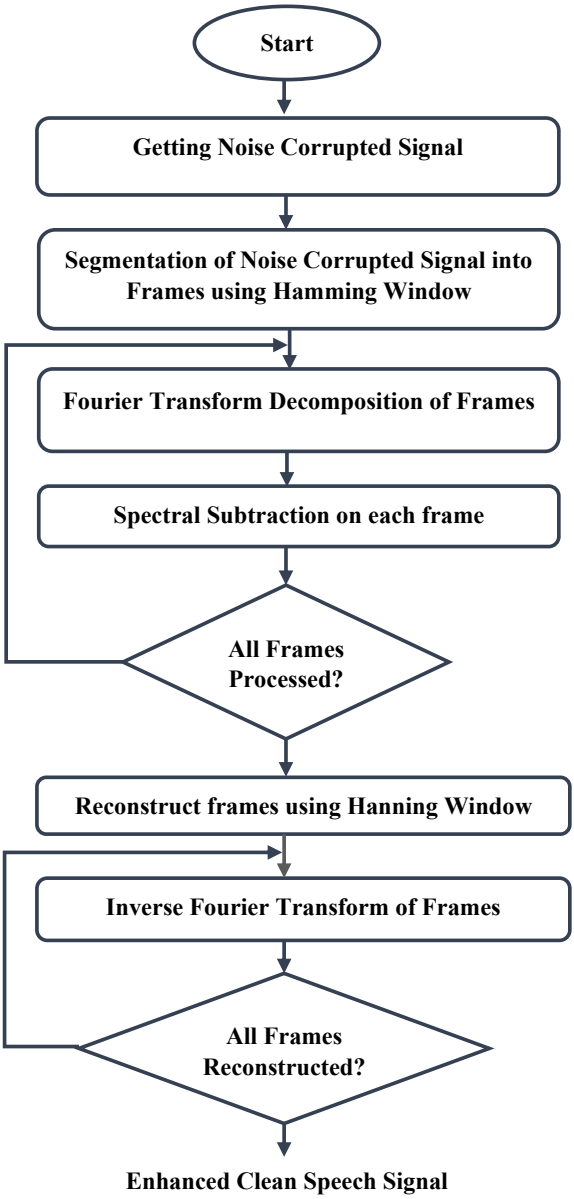

Fig. 4 Enhancement of noise corrupted speech signal approach

3 Program Code

```
clear all
[speech, fs] = wavread ( 'noisy_signal.wav' ); % read the input noisy speech signal
NFFT = 512; % N-point FFT - use of 512 point FFT
len_window = 0.025; % length of frame window (in milliseconds) taken as t_fl
overlap_window = 0.0125; % overlap time (in milliseconds) taken as t_o
frames_n = frame_signal ( speech, len_window * fs, overlap_window * fs,
@hamming ); % hamming window
cplx_spc = fft ( frames_n, NFFT, 2 );  % spectrum in complex form
mag_spc = abs ( cplx_spc ).^2; % spectrum of noisy speech signal in terms of
magnitude
phs_spc = angle ( cplx_spc ); % spectrum of noisy speech signal with phase
% spectral subtractive principle giving modified spectrum in terms of magnitude
noise_est = mean ( mag_spc (1:3, :) ); % noise estimated from first three frames
cln_spc = mag_spc – repmat ( noise_est, size ( mag_spc, 1), 1 ); % subtract
noise_est from mag_spc
cln_spc ( cln_spc < 0 ) = 0; % negative spectrum of magnitude discarded from clean
spectrum
% reconstruction of frames
rcnstrctd_frames_n = ifft ( sqrt (cln_spc ).* exp ( phs_spc), NFFT, 2 );      % inverse
FFT used
rcnstrctd_frames_n = real ( rcnstrctd_frames_n (:, 1:len_window * fs) ); % with
small residuals (complex)
enhanced_signal = deframe_sig_n( rcnstrctd_frames_n, length (speech),
len_window * fs, overlap_window * fs, @hamming);
                plot ( enhanced_signal ); % wave plot of enhanced signal
                sound ( enhanced_signal, fs ); % listen to enhanced signal
```

4 Experimental Results

While enhancing the speech signal, the main rationale is suppressing the noise from corrupted speech to upgrade the signal intelligibility together with quality. Signal quality forms the subjective performance measure that evaluates to what degree the speech sounds fine and thus includes the characteristics as naturalness, roughness of noise, etc., and intelligibility forms an objective performance measure that determines how much the signal is understood.

The experiment is conducted on two speakers, taking one male voice and female voice considering the coded speech database of ITU-T P-series recommendations [17]. This coded speech database comprises the sentences with varying durations

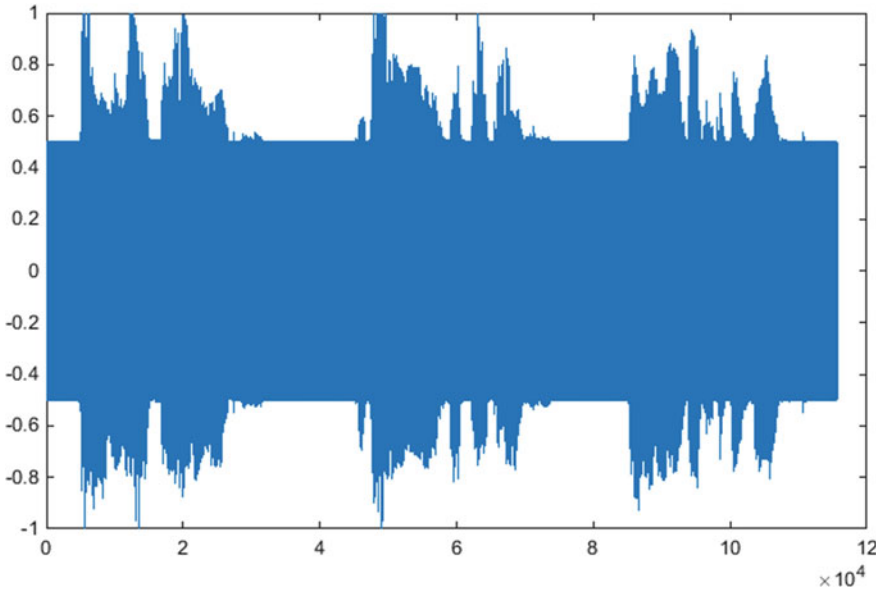

Fig. 5 Audio wave of corrupted female voice

that are uttered in diversified languages and accent. These uttered sentences are corrupted by noise particularly additive noise having contrasting signal-to-noise ratio (S/N or SNR) to authenticate and verify particular speaker and for speech recognition purpose, while incorporating this speech enhancement method as prior treatment to such systems. The experiment results of the methodology involving spectral subtractive principle on female corrupted voice and the enhanced female voice are depicted in Figs. 5, 6, 7, and 8. With the incorporation of algorithm, noise is shown to be removed from the signal, resulting in understandable speech signal.

5 Conclusion

In this paper, the procedure of enhancing the speech of interest incorporating Fourier transform domain and spectral subtractive principle is shown that suppresses the noise associated with speech signal. Further, this method is employed before the recognizer system for speech and speaker identification process to lessen the undesirable impact of noise and interference on speech, resulting in the improvement of speech quality and speech intelligibility. The experimental results show the audio wave of speech signal with and without the effect of noise together with the spectrum of both the signals. The waveform shows the removal of noise from female voice and deriving the clean voice free from ambient noise.

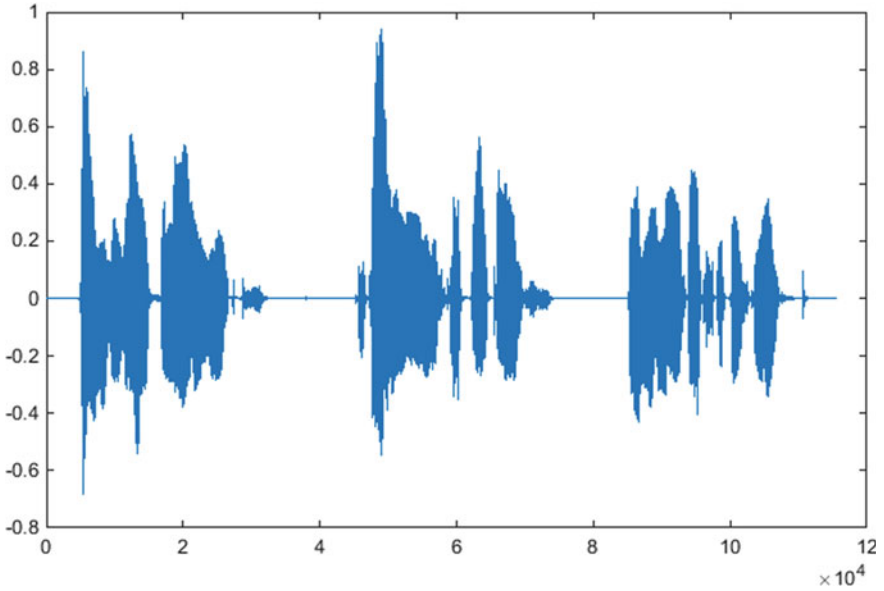

Fig. 6 Audio wave of enhanced female voice

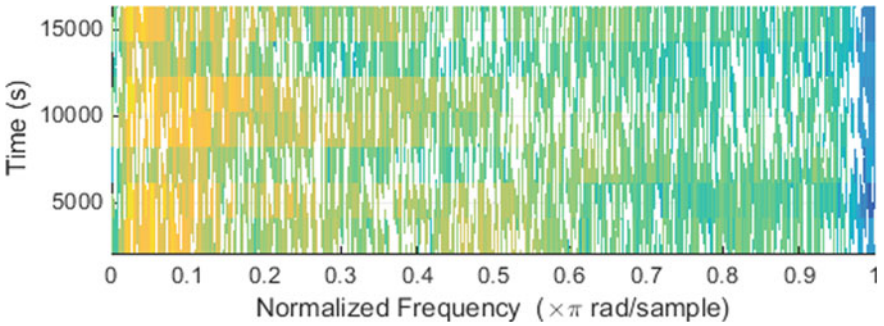

Fig. 7 Spectrum representation of corrupted female voice

6 Future Considerations

In the future work, we will implement speech enhancement in real-time systems involving time domain by employing fully convolutional neural network known as temporal convolutional neural network (TCNN), a hybrid deep learning approach. This method will involve the training of model in a speaker and noise unconstrained procedure and will include few parameters to train the model. This will explore deep neural network architecture pertaining to time domain analysis of speech enhancement. This research will further incorporate the analysis of additional speech

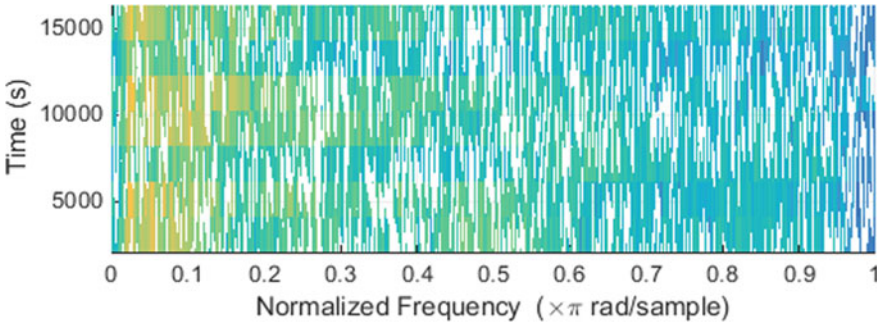

Fig. 8 Spectrum representation of enhanced female voice

processing tasks including de-reverberation of speech, echo suppression and cancelation, source separation, and speaker separation using TCNN model so as to upgrade SNR, quality, and intelligibility of speech signal.

References

1. Benesty J, Makino S, Chen J (2005) speech enhancement. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-27489-8
2. Iqbal A et al (eds) (2020) Soft computing in condition monitoring and diagnostics of electrical and mechanical systems. In: Advances in intelligent systems and computing, vol 1096. Springer, Singapore. https://doi.org/10.1007/978-981-15-1532-3
3. Iqbal A et al (eds) (2020) Meta heuristic and evolutionary computation: algorithms and applications. In: Studies in computational intelligence, vol 1096. Springer, Singapore. https://www.springer.com/gp/book/9789811575709
4. Karam M, Khazaal HF, Aglan H, Cole C (2014) Noise removal in speech processing using spectral subtraction. J Signal Inf Process 5:32–41. https://doi.org/10.4236/jsip.2014.52006
5. Benesty J, Sondhi MM, Huang Y (2008) Springer handbook of speech processing. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-49127-9
6. Loizou PC (2007) Speech enhancement theory and practice, 2nd edn. CRC Press Taylor and Francis, London
7. Schroeder MR (1965) U.S. Patent No. 3180936, filed 1 Dec 1960, issued 27 Apr (1965). https://patents.google.com/patent/US3180936
8. Schroeder MR (1968) U.S. Patent No. 3403224, filed 28 May 1965, issued 24 Sept (1968). https://patents.google.com/patent/US3403224
9. Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 27(2):113–120. https://doi.org/10.1109/TASSP.1979.1163209
10. Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. IEEE Proc 67:1586–1604. https://doi.org/10.1109/proc.1979.11540
11. Andersen BB, Dyreby J, Kjærskov FH, Mikkelsen OL, Nielsen PD, Zimmermann NH, Jensen B (2004) Bandwidth extension of narrowband speech using linear prediction. Aalborg University, Denmark, Worksheets
12. Paliwal K, Wójcicki K, Schwerin B (2010) Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 52:450–475. https://doi.org/10.1016/j.specom.2010.02.004

13. Upadhyay N, Karmakar A (2015) Speech enhancement using spectral subtraction-type algorithms: a comparison and simulation study. In: 11th International multi-conference on information processing 2015, Procedia Comput Sci 54:574–584. https://doi.org/10.1016/j.procs.2015.06.066
14. Paliwal K, Wójcicki K (2008) Effect of analysis window duration on speech intelligibility. IEEE Signal Process Lett 18(15):785–788. https://doi.org/10.1109/LSP.2008.2005755
15. Sen S, Dutta A, Dey N (2019) Audio processing and speech recognition concepts techniques and research overviews. Springer, Singapore. https://doi.org/10.1007/978-981-13-6098-5
16. Vaseghi SV (2000) Advanced digital signal processing and noise reduction, 2nd edn. Wiley & Sons Ltd, New York. https://doi.org/10.1002/0470841621
17. ITU-T Test Signals for Telecommunication Systems. https://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm