Wireless Personal Communications (2024) 134:2315–2346
https://doi.org/10.1007/s11277-024-11029-y

A Strategic Approach for Robust Dysarthric Speech Recognition

A. Revathi[1] · N. Sasikaladevi[2] · D. Arunprasanth[3] · Rengarajan Amirtharajan[1]

Accepted: 2 April 2024 / Published online: 20 April 2024

Abstract
The development of a system to recognize the speeches of standard speakers has been in practice for many decades. Research development is still progressing to implement a strategy to identify the speeches uttered by people with hearing impairment/Autism spectrum disorder/dysarthria. This work includes various speech enhancement techniques to increase the intelligibility of spoken utterances. This system uses perceptual features and different modelling techniques for developing a dysarthric speech recognition system. Perceptual features are extracted from raw speeches, and intelligibility-enhanced spoken utterances and models are created. The design features extracted from the test utterances are given to the models, and based on the classifier used, the test utterance is identified to be associated with the model. An Implementation of speech enhancement techniques would facilitate better accuracy. Decision-level fusion classification on integrating features, models, and speech enhancement techniques has provided overall accuracy of 81% for recognizing isolated digits spoken by a few dysarthric speakers. Better accuracy can be ensured for the database containing more utterances from many dysarthric speakers. This system would help caretakers understand the speeches uttered by persons affected with dysarthria to provide the necessary assistance.

Keywords Perceptual features · Dysarthric speech recognition · Speech enhancement techniques

1 Introduction

Speech production is a mechanism in which speech is considered an output of a vocal tract system excited by vocal cords' vibration due to airflow from the lungs. Articulators in the vocal tract and muscles move in response to the neural signals for producing speech. Muscles in the vocal tract are weak in delivering speech for persons with dysarthria. Dysarthria

✉ Rengarajan Amirtharajan
amir@ece.sastra.edu

[1] School of Electrical and Electronics Engineering, Thanjavur 613401, India

[2] School of Computing, SASTRA Deemed University, Thanjavur 613401, India

[3] Thanjavur Medical College, Thanjavur 613004, India

refers to multiple neurogenic disorders with irregularities in speech. It is measured in the articulator movements' strength, speed, range, tone, and precision. Dysarthric patients may struggle to control the articulatory mechanisms in producing normal sounds. Dysarthric persons may be easily identified by listening to their way of creating speech. They may utter unclear, slurred, random, rapid, slow, and weak speeches. They may have difficulty controlling the facial muscles and articulators for speech production. Stroke and injury or tumour in the brain influence dysarthric patients' neurological defects. It may also be caused by facial paralysis or tongue and weakness in throat muscles.

Speech-language pathologists treat dysarthric patients and administer remedial measures to improve their speaking abilities. It is suggested that dysarthric patients' communication abilities could improve by increasing lip and tongue movement, strengthening the speech muscles, reducing the rate of uttering speech, and doing regular voice-related exercises. Converting thoughts into speech sounds includes articulators in the vocal tract system [1]. Speech is an output of the linear time-varying vocal tract system excited by quasi-periodic air pulses due to vocal cords' vibration for voiced sounds and noise for unvoiced sounds. Dysarthria is a speech disorder caused due to the lack of ability to control articulators. Speech impairment affects every activity and makes these people's lives miserable.

Dysarthria makes coordinating the nerves and articulators used for speech production difficult. Articulation of various speech sounds will get perturbed by the uncontrolled behaviour of the nerves used for speech production. The speed at which dysarthric people utter speeches is relatively slow compared to standard speakers. Due to the lack of control of muscular activity, it is difficult for patients with dysarthria to control speech parameters such as loudness, speed, pacing, breath, rhythm, and voice quality. Due to cerebral palsy, the system is proposed to recognize persons' speeches with an articulation disorder [2]. The acoustic and language models are constituent components of the speech recognition system.

An acoustic model may be specific to persons with dysarthria, but a language model may be universal irrespective of any category of speakers. A speech recognition system [3] is developed to assess the severity of dysarthric people's problems. The Partial Least Square based Voice conversion (VC) method [4] is used for dysarthric people. Healthy speeches are transformed into dysarthric utterances for data augmentation [5], and large-scale machine-learning models are used for classification.

Convolutional bottleneck networks [6] are used for speech recognition. Two hybrid speech recognition systems (DNN-HMM and GMM-HMM) [7] have been developed for speech recognition, with a 13% improvement in word error rate. The system's efficiency for dysarthric persons has been improved using rhythm knowledge [8]. The connectionist approach assesses the severity of the dysarthria, and the Hidden Markov Model is used to recognize speaker-dependent dysarthric speech. System [9] is developed to convert physically disabled persons' spoken utterances into intelligible utterances to understand better. The transformations are based on the movement of articulators for speech production. Non-negative matrix factorization [10] is used for voice conversion, which is better than GMM-based voice conversion.

Augmentation of acoustic models with articulatory information [11] shows improved recognition of the speech of dysarthric speakers. This integration is done by suiting the dysarthric speakers. Deep learning neural networks [12] are used to predict the severity of the problem concerned with dysarthric speakers. Dysarthric speeches are modified based on temporal and spectral factors [13] to improve the intelligibility of the speeches uttered by dysarthric speakers. A speech recognition system [14] is developed for dysarthric speakers using Hidden Markov Models, and the severity

of the problems associated with their speeches being uttered is evaluated. Speech enhancement techniques [15–21] are described. A description of the UA- speech database [22] is given. A speech recognition system [23] is developed for hearing impaired. The unsupervised learning method has been developed [24] to assess the auditory systems for speech recognition, which do not need a specific transcription of training data.

The dysarthric speech classification from coded telephone speech was developed [25, 26]. This feature is extracted using the deep neural network-based glottal inverse filtering method. Furthermore, an algorithm is proposed for syllable boundary and repletion of syllable detection [27] in dysarthric speech. Acoustic speech parameters [28] are analyzed for patients with Parkinson's disease. Speech patterns are analyzed to study the speaking characteristics of dysarthric speakers, and speech recognition systems [29–34] are developed. Variational mode decomposition with wavelet thresholding is used for speech enhancement. CNNs [35] classify dysarthric speeches on the UA-speech database. Speaker-independent dysarthric speech assessment [36] systems are developed. Deep neural network architectures [37] are used for analyzing the speeches of a dysarthric speaker. Empirical mode decomposition and Hurst-based mode selection (EMDH), along with deep learning architecture using a convolutional neural network (CNN) [38], are used to improve the recognition of dysarthric speech. The diversity of the speech patterns [39] of dysarthric speakers is characterized using clinical perspective and speech analytics. Dysarthric speeches are synthesized using text–into–speech (TTS) conversion systems [40] to improve the accuracy of dysarthric speech recognition.

Deep-belief-neural networks [41] are used for dysarthric speech recognition. Dysarthric speeches are augmented [42] using more training data to improve accuracy. The TORGO dataset uses transfer learning-based convolutional neural networks (CNN) [43] for dysarthric speech recognition. Variational mode decomposition with wavelet thresholding is used for speech enhancement. CNNs [35] classify dysarthric speeches on the UA-speech database. Dysarthric speech recognition [44] uses features and models on the UA-speech database. Speech emotion recognition [45] is done using CNNs. Detection of dysarthric speech [46] is done using CNNs. Automatic assessment of dysarthric speech intelligibility [47] is done using deep learning techniques. Deep-learning-based acoustic feature representation [48] is done for dysarthric speech recognition [49–52]. Audio-visual features are considered [49] for dysarthric speech recognition. A dysarthric isolated digit recognition system with speech enhancement techniques has been developed [50]. This work emphasizes using speech enhancement techniques to improve the intelligibility of the speeches uttered by dysarthric speakers to establish a robust speech recognition system for dysarthria. It also emphasizes using different spectral features and machine learning techniques to produce the speech recognition system.

In this work on speaker-independent dysarthric speech recognition, Sect. 2 describes the database used, analysis of dysarthric speeches in time and frequency domains, implementation of speech enhancement techniques, Feature extraction procedures, modelling techniques and testing procedures. Section 3 depicts the system's experimental, subjective comparison between experimental and emotional, and statistical validation results. Finally, Sect. 4 summarises the dysarthric speech recognition system's outcome by applying speech enhancement techniques, features, models and testing procedures of different modelling techniques.

2 Preliminaries

2.1 Dysarthric Speech Database

The dysarthric dataset [22] considered in this work contains speech utterances from 6 speakers (M01, M04, M07, M09, F03, F05) in the age group 18–51 for each isolated digit. As per the database description, M01, M04, M07, M09, and F03 are low speech intelligibility. F05 is a speaker with high speech intelligibility. Subjective analysis done on the speeches of F05 by hearing would also indicate the clarity of the spoken utterance, and her speeches are similar to that of standard speakers. The listeners can easily understand and recognize the speech recordings of F05. However, these speakers are spastic and athetoid and persons who use wheelchairs. Speech intelligibility is measured in the average score in word transcription tasks. Few utterances are recorded from these speakers because it is difficult to understand and reciprocate the word transcriptions correctly. So, it isn't easy to increase the robustness of the dysarthric speech recognition system.

2.2 Analysis – Dysarthric Speech

It is fascinating to characterize the speech uttered by dysarthric speakers. There are many differences between dysarthric speakers in pronouncing words/sentences. This fact necessitates the provision of an extensive database for recognizing the words uttered by them. It is understood that their speeches are highly distorted, and subjective identification of utterances becomes difficult. On average, dysarthric and regular speakers' speeches are different in style, slang, and place of articulation. Figure 1 indicates the characteristics of the speech uttered by the dysarthric person in terms of signal variations and its spectrogram for the isolated digit "one".

Signal representation and spectrogram are shown in Fig. 2, which depict another dysarthric person's characteristics for uttering the word "one". The same word they spoke at different

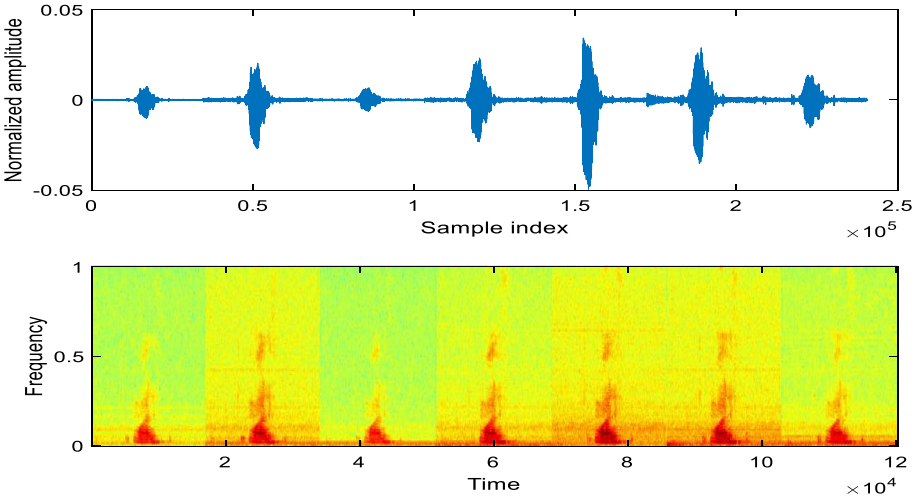

Fig. 1 Speech signal and spectrogram – Dysarthricspeaker (M09)—Digit "One."

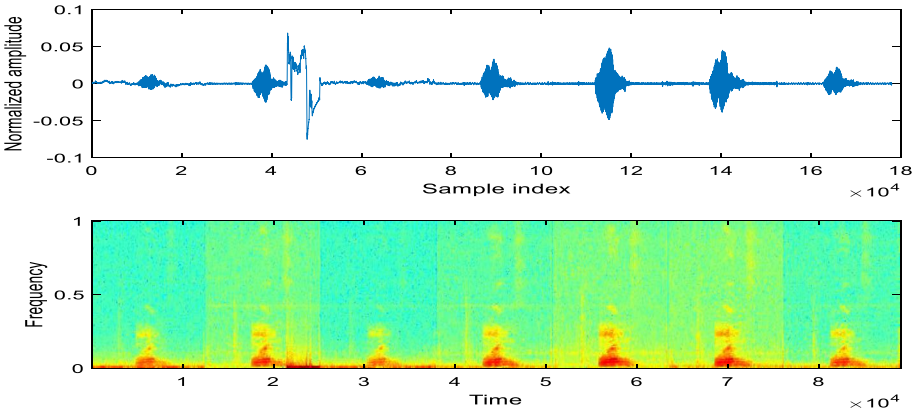

Fig. 2 Speech signal and spectrogram – Dysarthric speaker (M07)—Digit "One"

instants may have differences in amplitude and spectral energy. Since dysarthric speakers' speeches are indistinct, it is more imperative for better accuracy in recognizing dysarthric speakers' speeches.

2.3 Implementation of Speaker Independent Speech Recognition System

The speaker-dependent and independent speech recognition systems are implemented to recognize the isolated words /isolated digits/continuous sentences/spontaneous sentences uttered by speakers. A speaker-dependent system is developed using a set of speech utterances spoken by all the speakers for training and the remaining set of phrases spoken by the same speakers for testing. Speaker independent speech recognition system facilitates the use of utterances spoken by some of the enrolled speakers for training and other enrolled subjects' utterances for testing.

Feature extraction and modelling are the two stages constituting the training phase. The feature extraction stage facilitates the extraction of speech-specific robust features. Then, these features are applied to the modelling techniques for creating templates specific to speeches. The testing phase dwells upon using test feature vectors in the models designed to recognize the spoken utterance. The word is finally recognized as associated with the pertinent model based on matching. So, it is imperative to have a proper notion of robust feature selection, modelling techniques, and implementing the appropriate testing procedure. It is also essential to use techniques to improve the lucidity of dysarthric speakers' distorted speeches so that the system's accuracy can be reasonably enhanced.

2.4 Speech Enhancement Techniques

In noisy practical environments, background noise sources often degrade speech clarity. So, it is required to use efficient speech enhancement techniques to improve the clarity of the speech. The noisy speech is represented by the Eq. (1)

$$y_n(m) = x_n(m) + d_n(m) \tag{1}$$

$x_n(m)$ – Clean speech signal.

$d_n(m)$ – Noise signal.

2.4.1 *Single-channel online enhancement of speech* [15]

Background noises and reverberation affect the voice-based interaction between people. Speech enhancement techniques are used to improve the quality of the speech for better speech recognition. Online speech enhancement technique based on the all-pole model enhances speech quality. It is implemented using reverberation power and a hidden Markov model for removing noise superimposed with speech. Statistical parameters are estimated from the speech and noise, and analysis is performed by taking a short-time Fourier transform (STFT) with filters spaced in the MEL scale; spectral gain is derived.

Figure 3 indicates the speech enhancement process in the STFT domain. System parameters and signal powers are estimated using the MEL-spaced sub-bands. Then, the transformation of the power spectrum is done by using filters spaced in the Mel scale.

Consider noisy speech signals as $Y_n(m)$. STFT is taken on $Y_n(m)$ and coefficients are computed as in (2)

$$Y_n(k) = \sum_{m=0}^{k-1} Y_n(m)w(m)e^{\frac{-i2\pi mk}{N}} \tag{2}$$

k – STFT frequency bin. n – Time frame index. w(n) – Hamming Window sequence.

A power-domain filter bank is applied to compute the power in k Mel-spaced sub-bands as in (3)

$$\widehat{Y}_n(F) = \sum_{k=0}^{F-1} a_{F,k}\left|Y_n(k)\right|^2 \tag{3}$$

$a_{F,k}$ – Frequency response of the triangular filters.

HMM model is used to define the clean speech with an input probability distribution, state transition probabilities, and output observation probability distribution as in (4)

$$H_m = \left(x_m, d_m, R_m\right)^T \tag{4}$$

H_m is an HMM model including the reverberation R_m and noise parameters d_m for all subbands in a single state vector. Noise removal and improvement in intelligibility from noisy speech are made in the STFT domain. Figure 4 indicates the signals before and after applying the online speech enhancement algorithm on dysarthric speech.

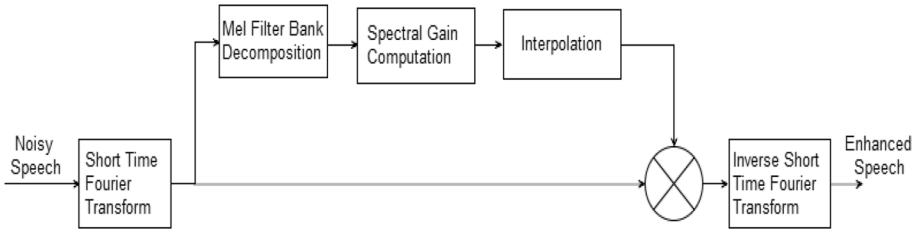

Fig. 3 Speech enhancement using a single-channel online enhancement technique

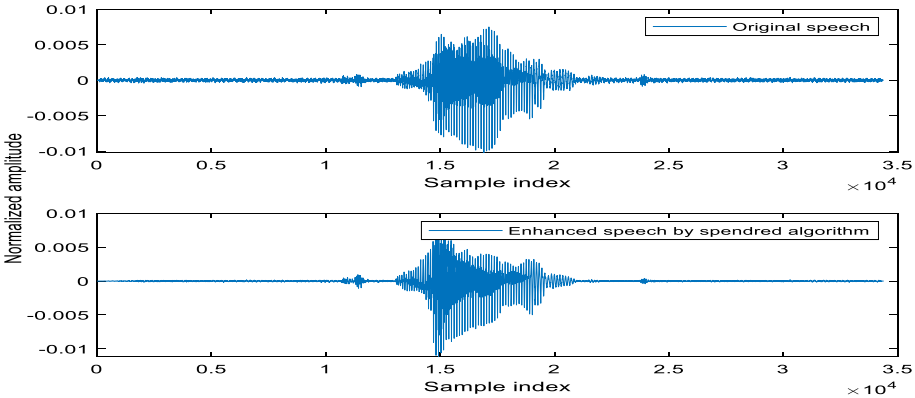

Fig. 4 Illustration of online speech enhancement algorithm

2.4.2 *Speech Enhancement Using a Minimum Mean Square Error (MSE) LogSpectral Amplitude Estimator* [16]

Minimizing MSE of the log spectra of the difference between the original signal's short-time spectral amplitude and the estimated signal is performed by the short-time spectral amplitude (STSA) estimator. The magnitude and phase response of the noisy, noise and clean speech signal is expressed in the frequency domain.

As $Y_k = R_k e^{j\gamma_k}$, $D_k = B_k e^{j\beta_k}$ and $X_k = A_k e^{j\alpha_k}$

The short-time spectral amplitude estimator $\widehat{\mathrm{A_k}}$, for minimizing the distortion measure is defined as in (5)

$$\overline{A_k} = E[\left(\log A_k - \log \widehat{A_k}\right)^2 \tag{5}$$

The expected value of $\widehat{A_k}$ given Y_k equal to the expected value of A_k given Y_k as in (6)

$$\overline{A_k} = \exp\left\{E\left[\ln A_k | Y_k\right]\right\} \tag{6}$$

MSE of log power spectra is calculated as in (7)

$$E\left\{\left(\log A_k^2 - \log \widehat{A_k}^2\right)^2\right\} \tag{7}$$

$\widehat{A_k}^2$ Denote the estimator of $\overline{A_k}$ as in (8)

$$\overline{A_k} = \sqrt{\widehat{A_k}^2} \tag{8}$$

$E[\ln A_k | Y_k]$ is computed by utilizing the moment-generating function of $\ln A_k$ given Y_k.

Let $Z_k = \ln A_k$, and $\varphi_{Z_{k|Y_k}(\mu)}$ of Z_k given Y_k Be the moment generating function, and it is defined as in (9)

$$\varphi_{Z_{k|Y_k}(\mu)} = E\{\exp(\mu Z_k | Y_k)\} \tag{9}$$

$E\{[\ln A_k | Y_k]\}$ is obtained from $\varphi_{Z_{k|Y_k}(\mu)}$ by using (10)

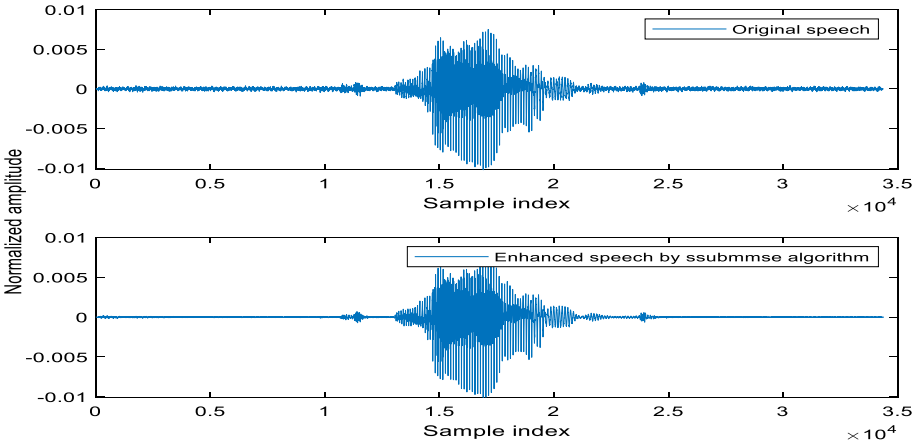

Fig. 5 Illustration of speech enhancement by log spectral amplitude estimator

$$E\{[\ln A_k|Y_k]\} = \frac{d}{dy}\varphi_{Z_{k|Y_k}(\mu)} at\mu = 0 \tag{10}$$

The STSA estimator is as in (11)

$$\overline{A_k} = \frac{\varepsilon_k}{1+\varepsilon_k}\exp\left\{\frac{1}{2}\int_{v_k}^{\alpha}\frac{e^{-t}}{t}dt\right\}R_k \tag{11}$$

Figure 5 indicates the speech enhancement process using a log spectral amplitude estimator.

2.4.3 *Speech Enhancement by Using Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator* [17]

The signal x(t), noise n(t), and the noisy observations y(t) are expressed in the frequency domain as D_k and Y_k. The Y_k in the interval [0 T] is defined as in (12)

$$Y_k = \frac{1}{T}\int_{t=0}^{T} y(t)\exp\left(\frac{-j2\pi kt}{T}\right)\mathrm{dt} \tag{12}$$

The spectral components are uncorrelated to each other; The MMSE estimator. $\widehat{A_k}$ of A_k given Y_k is obtained as in (13) and (14)

$$\widehat{A_k} = E\{A_k|Y_k\} \tag{13}$$

$E\{.\}$ – denotes Expectation operation

$$\widehat{A_k} = \frac{\int_0^{\infty}\int_0^{2\pi} a_k p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)da_k d\alpha_k}{\int_0^{\infty}\int_0^{2\pi} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)da_k d\alpha_k} \tag{14}$$

$p(.)$- Probability density function.

$p(Y_k|a_k,\alpha_k)$ is given by (15)

$$p\left(Y_k|a_k,\alpha_k\right)=\frac{1}{\pi\lambda_d(k)}\exp\left\{\frac{-1}{\lambda_d(k)}\left|Y_k-a_k e^{-i\alpha_k}\right|^2\right\} \tag{15}$$

$p\left(a_k,\alpha_k\right)$ is given by (16)

$$p\left(a_k,\alpha_k\right)=\frac{-a_k}{\pi\lambda_x(k)}\exp\left\{\frac{-a_k^2}{\lambda_x(k)}\right\} \tag{16}$$

$\lambda_d(k)$, $\lambda_x(k)$ are the variances of the *kths* pectral component of the noise and the speech.

Substituting $p\left(a_k,\alpha_k\right)$ in Eqs. (17) and (18)

$$\widehat{A_k}=\frac{\int_0^\infty a_k^2\exp\left(\frac{-a_k^2}{\lambda_k}\right)I_0\left(2a_k\sqrt{\frac{v_k}{\lambda(k)}}\right)da_k}{\int_0^\infty a_k\exp\left(\frac{-a_k^2}{\lambda_k}\right)I_0\left(2a_k\sqrt{\frac{v_k}{\lambda(k)}}\right)da_k} \tag{17}$$

$$\widehat{A_k}=\Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k}\exp\left(\frac{-v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right)+v_k I_1\left(\frac{v_k}{2}\right)\right]R_k \tag{18}$$

$$\Gamma(1.5)=\frac{\sqrt{\pi}}{2}$$

$\Gamma(.)$ Denotes the gamma function.

$I_0(.), I_1(.)$ Denotes the modified Bessel functions of zero and first order, respectively, with parameters as in Eqs. (19–21)

$$v_k=\frac{\epsilon_k}{1+\int_k}\gamma_k \tag{19}$$

$$\epsilon_k=\frac{\lambda_x(k)}{\lambda_d(k)} \tag{20}$$

$$\gamma_k=\frac{{R_k}^2}{\lambda_d(k)} \tag{21}$$

ϵ_k and γ_k – a priori and a posteriori signal-to-noise ratios.

Figure 6 depicts the speech enhancement process by a short-time spectral amplitude estimator.

2.4.4 *Wavelet denoising for Speech Enhancement* [18]

The wavelet denoising technique suppresses noise from noisy speech to obtain clean speech. First, wavelet packet transform decomposes noisy speech into approximation and detail coefficients. Then, the threshold is fixed and applied to the final level sub-band coefficients to minimize the noise propositions. Figure 7 shows the wavelet-based enhancing quality of speech. Finally, Enhanced speech is obtained by upsampling and interpolating the modified detail and approximation coefficients.

Figure 8 demonstrates the speech enhancement process by using wavelets.

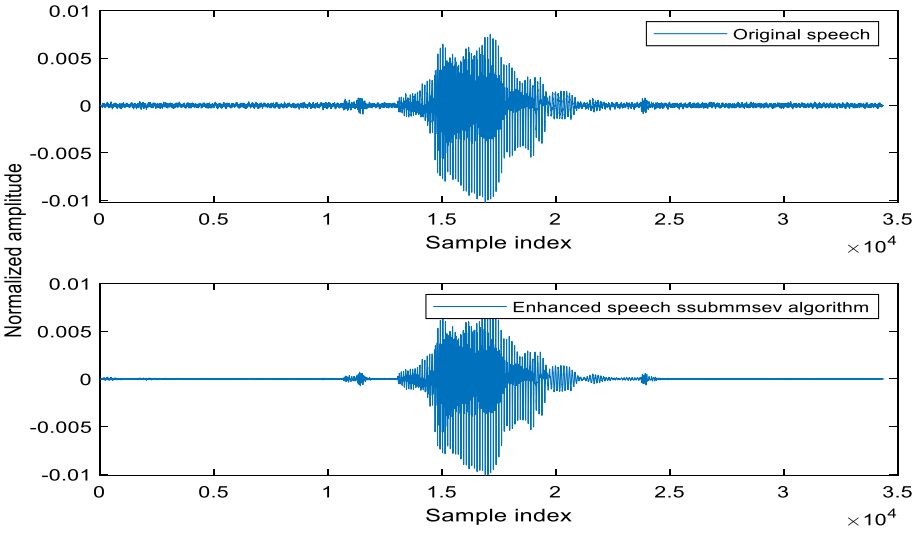

Fig. 6 Illustration of speech enhancement by short-time spectral amplitude estimator

2.4.5 *Probabilistic Geometric Approach (PGA) to spectral subtraction based Speech Enhancement* [19]

A confidence parameter of noise estimation is introduced in the gain function, which removes the noise efficiently and prevents speech distortion. The schematic shown in Fig. 9 depicts the PGA-based speech enhancement technique modules.

The equation represents the STFT of noisy speech as in (22)

$$Y_n(k) = X_n(k) + D_n(k) \tag{22}$$

n – frame number.

The STFT of $y_n(m)$ is represented as in (23)

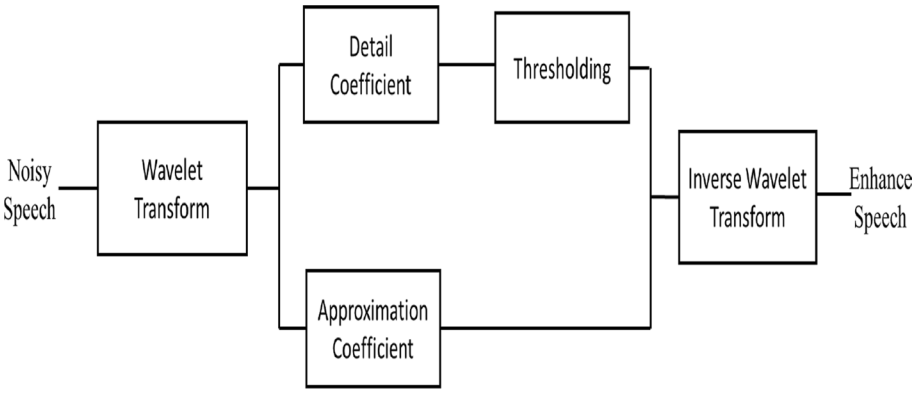

Fig. 7 Speech enhancement using Wavelet denoising technique

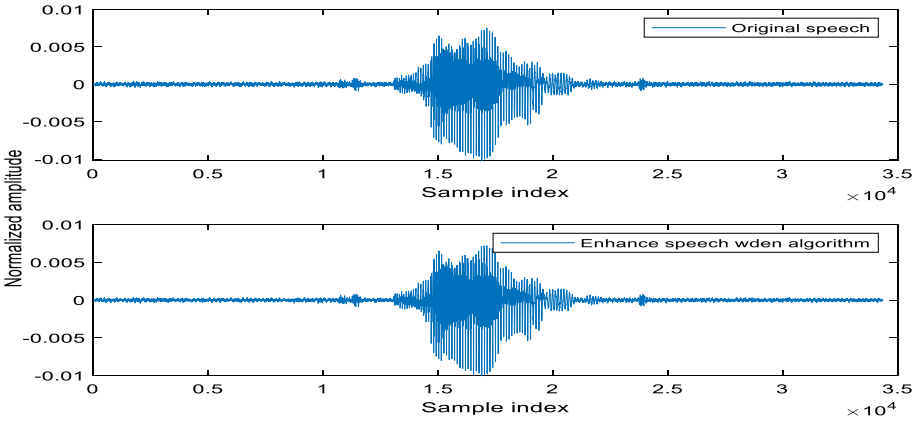

Fig. 8 Illustration of speech enhancement process by wavelets

$$Y_n(k) = \sum_{m=0}^{N-1} y_n(m) e^{-j\frac{2\pi mk}{N}} \tag{23}$$

From the basic rule of spectral subtraction, the Eq. (23) can be written as (24)

$$\left|\widetilde{X_n(k)}\right|^2 = |Y_n(k)|^2 - |D_n(k)|^2 \tag{24}$$

This equation can also be written as (25)

$$\left|\widetilde{X_n(k)}\right|^2 = \left|H_{n(PGA)}(k)\right|^2 - |D_n(k)|^2 \tag{25}$$

$H_{n(PGA)}(k)$ – gain function of the nth frame. $X_n(k)$, $Y_n(k)$ and $D_n(k)$ can be expressed in polar form as

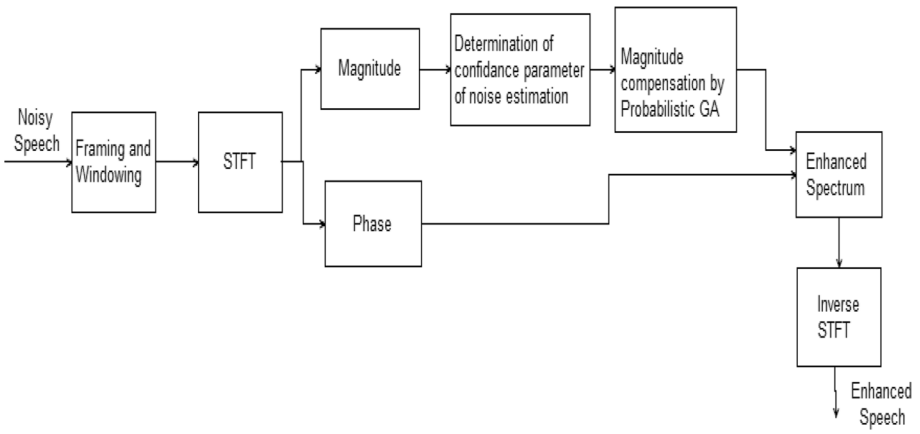

Fig. 9 Speech enhancement using Probabilistic Geometric Approach

$$X_n(k) = a_x e^{i\theta_x},\ Y_n(k) = a_y e^{i\theta_y},\ D_n(k) = \rho a_d e^{i\theta_d}$$

ρ – is a constant dependent on posterior and a priori SNRs.
a_x, a_y and a_d are the magnitude response of clean, noisy and noise signals.
θ_x, θ_y and θ_d are the phase response of clean, noisy and noise signals.
The gain function $H_{n(PGA)}(k)$ can be defined as in (26)

$$H_{n(PGA)}(k) = \sqrt{\frac{a_y^2}{a_x^2}} \tag{26}$$

The unchanged phase spectrum and compensated magnitude spectrum are combined to produce an enhanced speech by using the formula in (27)

$$\widetilde{x_n(m)} = real(Inverse\ STFT)\{\widetilde{X_n(k)}\} \tag{27}$$

Figure 10 indicates the effect of the probabilistic geometric approach in enhancing the speech uttered by the dysarthric speaker.

2.4.6 *The Geometric Approach to Spectral Subtraction-Based Speech Enhancement* [20]

Noise present in speech is effectively reduced by spectral subtraction. The spectral subtraction method removes the noise based on the assumption that the noise is uncorrelated with any other system signal. Figure 11 gives a detailed description of the blocks used for geometric approach-based speech enhancement.

This equation to compute signal spectrum is as in (28)

$$\left|\widetilde{X_n(k)}\right|^2 = \left|H_{n(GA)}(k)\right|^2 - \left|D_n(k)\right|^2 \tag{28}$$

$H_{n(GA)}(k)$ – gain function of the Geometric approach of the nth frame.

The magnitude and phase response of the noisy, noise and clean speech are related as in (29)

$$a_y e^{i\theta_y} = a_x e^{i\theta_x} + a_D e^{i\theta_d} \tag{29}$$

The triangle shown in Fig. 12 depicts the phase spectra of the noisy speech and clean speech and noise signals.

In Fig. 12, Eqs. (30) give the trigonometric relations for magnitude and phase spectra of noisy speech, clean speech and noise signals.

$$\overline{AB} \perp \overline{BC}$$

$$a_y \sin(\theta_D - \theta_y) = a_x \sin(\theta_D - \theta_x)$$

Taking square of both sides

$$a_y{}^2 sin^2(\theta_D - \theta_y) = a_x{}^2 sin^2(\theta_D - \theta_x)$$

$$a_y{}^2\left[1 - cos^2(\theta_D - \theta_y)\right] = a_x{}^2\left[1 - cos^2(\theta_D - \theta_x)\right]$$

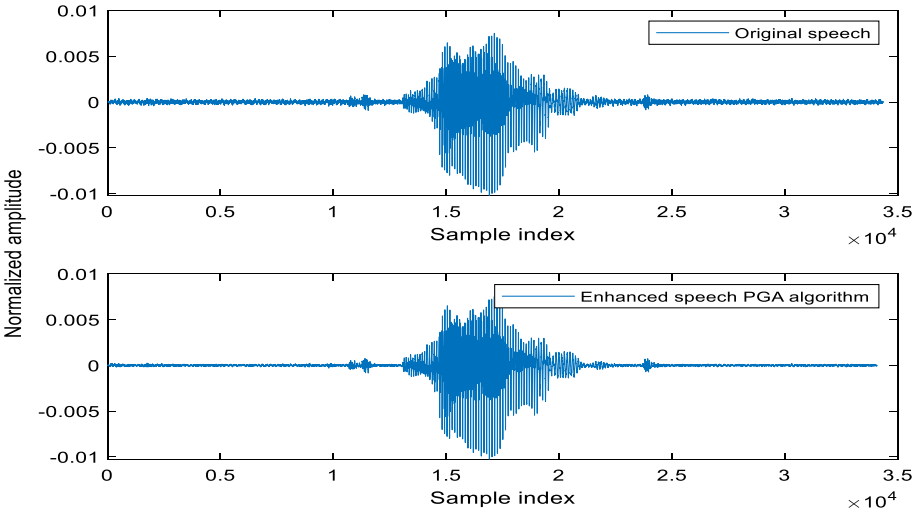

Fig. 10 Illustration of speech enhancement by a probabilistic geometric approach

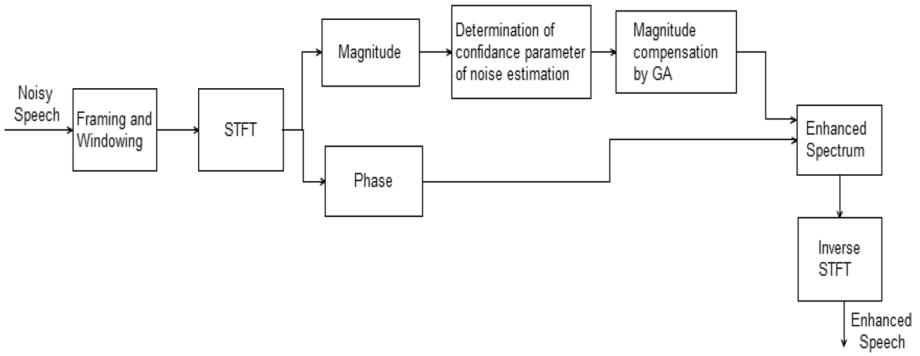

Fig. 11 Geometric approach to spectral Subtraction for speech enhancement

It can be written as in (30)

$$a_y^2\left[1 - C_{yD}^2\right] = a_x^2\left[1 - C_{xD}^2\right] \tag{30}$$

The gain function is defined as in (31)

$$H_{n(GA)} = \frac{a_x}{a_y} = \sqrt{\frac{1 - C_{yD}^2}{1 - C_{xD}^2}} \tag{31}$$

Using cosine rules in triangle ABC, Eqs. (32) and (33) are used

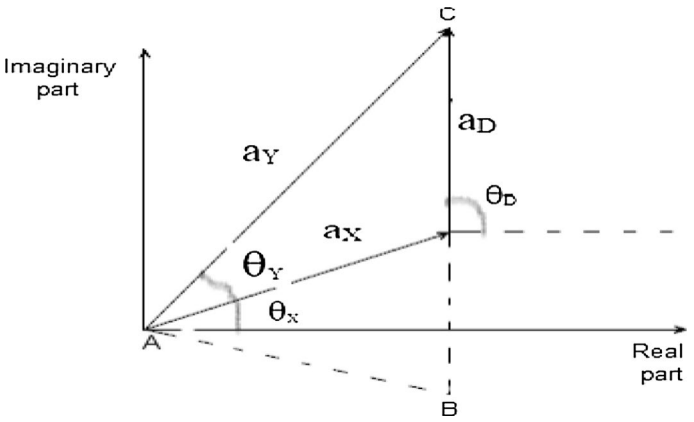

Fig. 12 Geometric representation of noisy speech, clean speech and noise spectra

$$C_{yD} = \frac{a_y^2 + a_D^2 - a_x^2}{2a_y a_D} \tag{32}$$

$$C_{xD} = \frac{a_y^2 - a_D^2 - a_x^2}{2a_x a_D} \tag{33}$$

Dividing both numerator and denominator of the equation by ${a_D}^2$ as in (34) and (35)

$$C_{yD} = \frac{\frac{a_y^2}{a_D^2} + 1 - \frac{a_x^2}{a_D^2}}{\frac{2a_y}{a_D}} \tag{34}$$

$$C_{xD} = \frac{\frac{a_y^2}{a_D^2} - 1 - \frac{a_x^2}{a_D^2}}{\frac{2a_x}{a_D}} \tag{35}$$

$$\Upsilon = \frac{a_y^2}{a_D^2},\ \varepsilon = \frac{a_x^2}{a_D^2}$$

Υ – A posteriori SNR.
ε– A priori SNR.
The gain function can be written as in (36)

$$H_{n(GA)} = \frac{a_x}{a_y} = \sqrt{\frac{1 - \frac{(\gamma+1-\varepsilon)^2}{4\gamma}}{1 - \frac{(\gamma-1-\varepsilon)^2}{4\gamma}}} \tag{36}$$

Enhanced speech is obtained by combining an unchanged phase spectrum with compensated magnitude spectrum, as in (37)

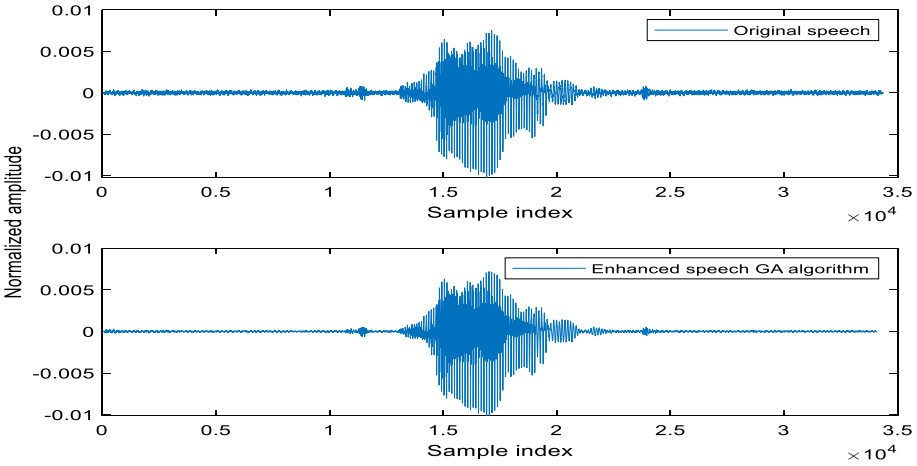

Fig. 13 Effect of speech enhancement process by a geometric approach

$$\widetilde{x_n(m)} = real(Inverse\ STFT)\{\widetilde{X_n(k)}\} \tag{37}$$

Figure 13 illustrates the speech enhancement process by the geometric approach applied to the dysarthric speaker's speech.

2.4.7 *Phase Spectrum Compensation Based Speech Enhancement* [21]

This method combines the modified phase response with a magnitude response to get the changed frequency response for the noisy speech. Analyzing the relation between the spectral and time domains during the synthesis process makes it possible to cancel out the high-frequency components, thus producing a signal with a reduced noise component. The STFT of the noisy signal is computed as in (38)

$$Y_n(k) = \left|Y_n(k)\right| e^{j\angle Y_n(k)} \tag{38}$$

The compensated short-time phase spectrum is computed by using the Eqs. (39) and (40)

The process obtains phase spectrum compensation function as in Eq. (39)

$$\wedge_n(k) = \lambda\psi(k)\left|D_n(k)\right| \tag{39}$$

$\left|D_n(k)\right|$ Defines magnitude response of the noise signal.

λ– Constant.

The anti-symmetry function $\psi(k)$ is defined as in (40)

$$\psi(k) = \begin{cases} 1 & if\ 0 < \frac{k}{N} < 0.5 \\ -1 & if\ 0.5 < \frac{k}{N} < 1 \end{cases} \tag{40}$$

Multiplication of symmetric magnitude spectra of the noise signal with anti-symmetric function ψ(k) produces an anti-symmetric $\wedge_n$(k). Noise cancellation is made during the synthesis process by utilization of the anti-symmetry property of the phase spectrum compensation function. The complex spectrum of noisy speech is computed as in Eq. (41)

$$Y_n(k) = X_n(k) + \wedge_n(k) \tag{41}$$

The compensated phase spectrum of the noisy signal is derived as in Eq. (42)

$$\angle Y_n(k) = ARG\left[Y_n(k)\right] \tag{42}$$

Recombination of the compensated phase response with the magnitude response of the noisy signal is done to get the modified spectrum, from which enhanced speech is derived by performing inverse transform as in (44) on the modified spectral response given in (43).

$$S_n(k) = \left|Y_n(k)\right| e^{j\angle Y_n(k)} \tag{43}$$

$$s(n) = real\left[inverse\ STFT\ \left(S_n(k)\right)\right] \tag{44}$$

Figure 14 indicates the performance of the speech enhancement technique by phase compensation. Figure 15 depicts the variation in the distribution of samples for each speech enhancement technique by performing histogram equalization.

2.5 Feature Extraction

PLP extraction is based on the principle of how the human ear perceives sounds [1]. The PLP extraction method is similar to the linear prediction coefficient method, except its spectral characteristics are changed based on the human auditory system. Perceptual features with filters spaced in BARK, ERB, and MEL scales are extracted from the pre-processed speech using the techniques shown in Fig. 16. The FFT technique obtains the pre-processed signal's power spectrum; the auditory spectrum is obtained by multiplying the signal's power spectrum. The squared magnitude spectrum of the filters is spaced in different frequency scales. Cube root compression and Loudness equalization simulate the human ear's power law of hearing perception. Finally, the inverse transform is performed to obtain the signal, from which cepstral coefficients are derived using LPC and Cepstral analyses.

i Procedural steps used for PLPC, MF-PLPC and ERB-PLPC extraction are summarised as follows.

1 Computation of power spectrum on pre-processed speech segment.

2 Critical band analysis uses 21, 47 and 35 critical bands in BARK, Mel, and ERB frequency scales at 16 kHz as sampling frequency. The magnitude response of the filter banks spaced in the MEL scale, BARK scale and ERB scale are shown in Figs. 17, 18 and 19. Frequency in Hz and other frequency scales, namely MEL, BARK and ERB, are related as in (45), (46) and (47).

$$f(Mel) = 2595 * \log_{10}\left(1 + \frac{f(Hz)}{700}\right) \tag{45}$$

$$(Bark) = \left[\frac{26.81 f(Hz)}{1960 + f(Hz)}\right] - 0.53 \tag{46}$$

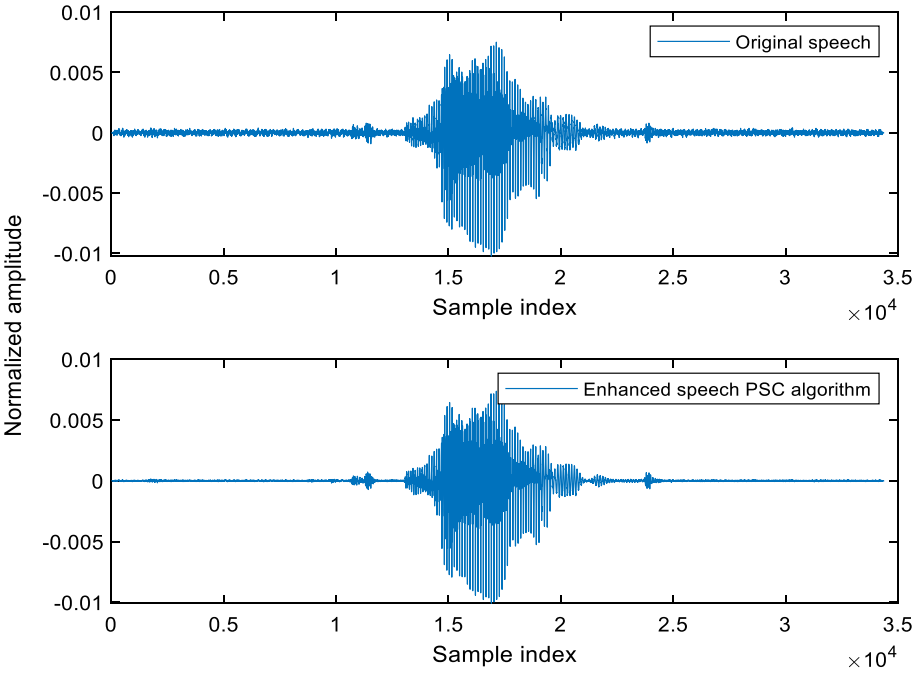

Fig. 14 Illustration of speech enhancement by phase spectrum compensation technique

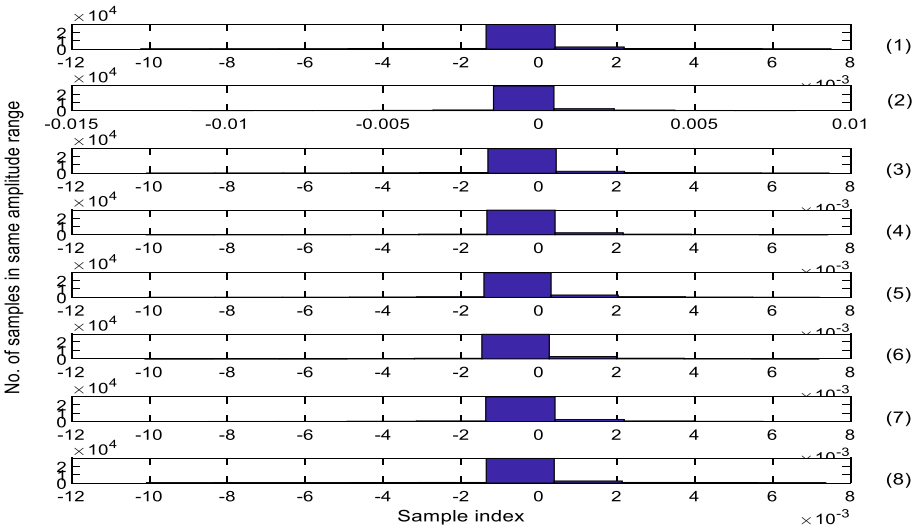

Fig. 15 Histogram equalization – (1) Original speech (2–8) Enhanced speech using speech enhancement techniques

$$f(ERB) = 24.7(4.37f(Hz) + 1) \tag{47}$$

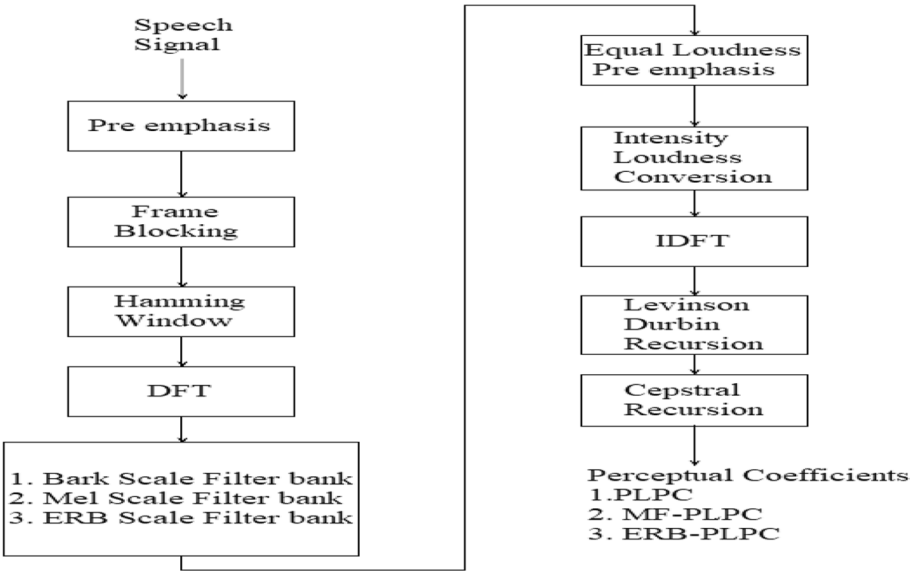

Fig. 16 Procedure—Perceptual features extraction

3 Hearing's power law is stimulated by hearing, loudness equalization, and cube root compression. Loudness equalization is done by pre-emphasis filter to weight the filter-bank outputs to simulate the sensitivity of ears to perceive sounds as in (48).

$$E(\omega) = \frac{\left(\omega^2 + 56.8 * 10^6\right)^4}{\left(\omega^2 + 6.3 * 10^6\right)\left(\omega^2 + 0.38 * 10^9\right)\left(\omega^6 + 9.58 * 10^{26}\right)} \tag{48}$$

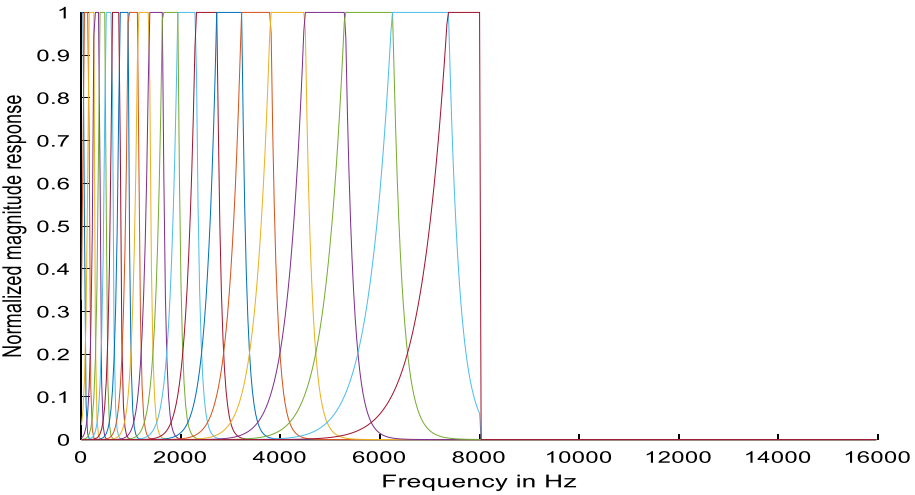

Fig. 17 Magnitude response of critical bands in the BARK scale

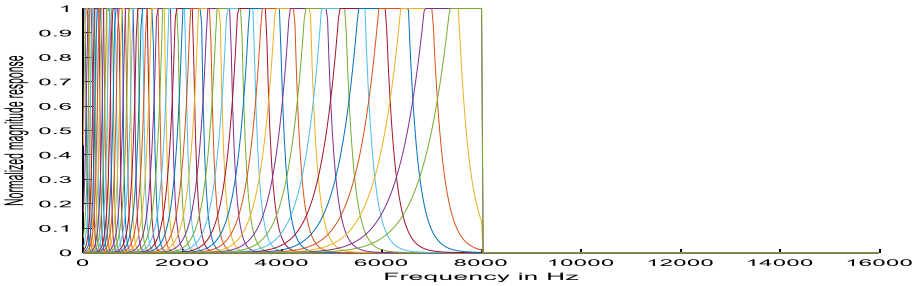

Fig. 18 Magnitude response of the critical bands in the MEL scale

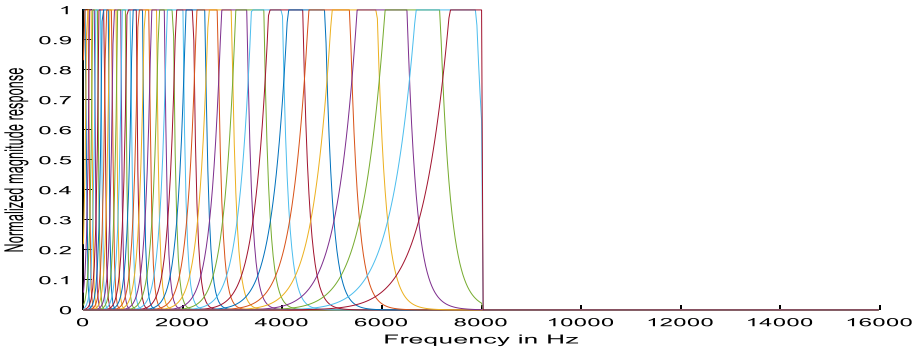

Fig. 19 Magnitude response of the critical bands in the ERB scale

Transformation of equalized values is done by a power law of hearing (i.e.) raising the power by 0.33. It is represented in (49)

$$L(\omega) = I(\omega)^{\frac{1}{3}} \tag{49}$$

4 IFFT is performed on L (ω).
5 The Levinson-Durbin procedure computes the LP coefficient.
6 LP coefficients are converted into PLP, MFPLP and ERBPLP Cepstral coefficients.

2.6 Implementation of Template Creation Module

For a speech recognition system, templates are created to act as a representative model pertinent to speeches to be recognized. VQ-based or fuzzy-based clustering technique forms the low dimensional cluster set from the high dimensional training set among the many modelling techniques. They include M cluster centroids for contemplating the speech model from the training data of high dimension. This process is done by computing the Euclidean distance between the training set and initial cluster centroids. These cluster centroids are updated for iterations, and finally, the cluster set formed in pertinent speech represents the training set of feature vectors. For testing, Euclidean distance is computed between test vectors and cluster set, and cluster centroid, which produces minimum distance, is restored. All the test speech features and minimum distances are calculated and

stored as a model value. This process is implemented for all models. Finally, a model is selected for the test speech to compare the minimum of model values. MHMM modelling technique facilitates the expectation–maximization procedure to generate templates containing maximum likelihood parameters. The testing procedure for MHMM enables the application of test features to the models, and log-likelihood values are computed. The model associated with the test speech has the most considerable log-likelihood value.

3 Experimental Evaluation – Results and Discussion

The dysarthric speech recognition system is evaluated based on perceptual features and various modelling techniques. Different speech enhancement techniques applied to distorted dysarthric speeches would enable the system to enhance performance. This speech recognition system encompasses training and testing phases. During training, speeches are concatenated, and conventional pre-processing techniques are applied to the speech data. After the pre-processing, extraction of perceptual features is performed, followed by using features for creating templates. Test speeches undergo pre-processing during testing, and perceptual features are extracted. These features are applied to all speech templates, and based on the classifier used; speech is identified to be associated with pertinent speech templates. Recognition accuracy/word error rate is used as a performance metric for evaluating the system. Finally, speech enhancement techniques are applied to raw training and test speeches, and the system's performance is assessed. The implementation uses the decision-level fusion of speech enhancement techniques, features, and modelling techniques to classify the pertinent dysarthric speech. Features extracted from test segments are applied to the models, and the model index based on the classifier used is derived. This process is repeated for all test segments. Finally, a decision-level fusion of correct indices about the modelling techniques is done to augment the system's performance. The decision-level fusion classifier is depicted in Fig. 20.

This decision-level fusion classifier classifies the pertinent speech based on the correct classification of features, modelling techniques, and speech enhancement techniques. Table 1 indicates the system's performance with a decision-level fusion of elements and models by taking speeches with and without speech enhancement techniques. The overall accuracy for ten digits in Fig. 21 shows the system's evaluation for recognizing dysarthric speeches against speech enhancement techniques with a decision-level fusion of results on features and models. Individual accuracy for some isolated digits is 100%, with overall accuracy for the decision-level fusion of influences of the features, models, and speech enhancement techniques at 80.2%.

Individual accuracy for some isolated digits is low because the testing is done with utterances of a dysarthric speaker with only 6% speech intelligibility. Training the models with many feature vectors can enhance the system's accuracy. The system has not provided good accuracy because it is tested for the female speaker with only 6% speech intelligibility. Decision-level fusion of results of features and models has provided a better overall accuracy of 43%, with an application of phase spectrum compensation as a speech enhancement technique. It is 12% more than the system without using a speech enhancement mechanism. So, the system's accuracy depends on features, models, speech enhancement techniques, and the test set of spoken utterances. However, the system is trained for speech utterances at all intelligibility levels. Therefore, obtaining better accuracy for speakers with very low intelligibility is difficult.

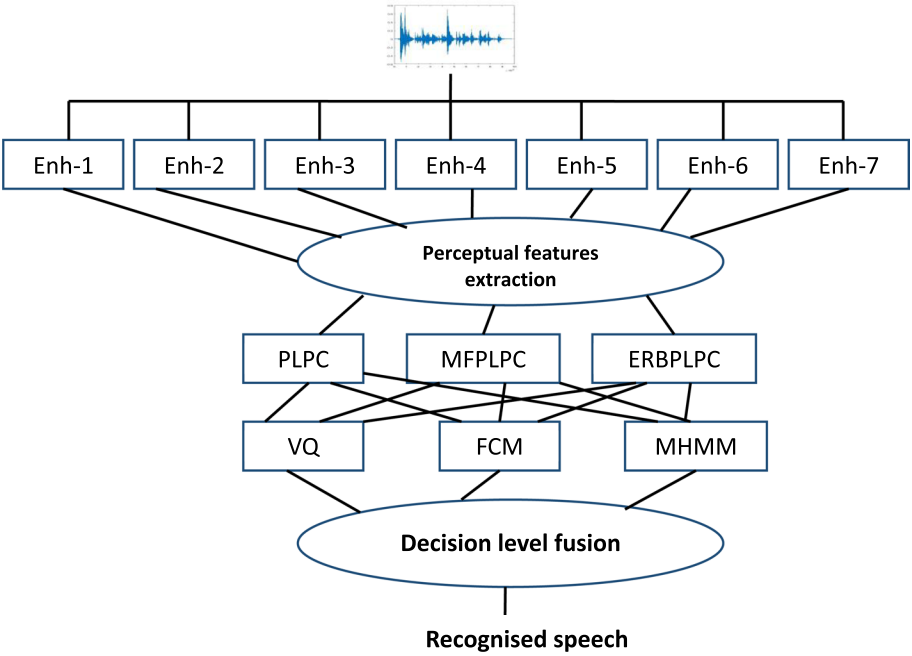

Fig. 20 Decision-level fusion classifier

Table 1 Performance – Average accuracy—decision level fusion of features and models for speech enhancement techniques

Technique	One	Two	three	four	Five	six	seven	eight	nine	Zero
Without speech enhancement	31	49	12	3	30	87	10	26	38	29
Spendred	37	28	49	1	10	62	62	13	31	28
Ssubmmse	71	61	5	30	26	59	17	25	62	19
Ssubmmsev	45	20	34	31	30	51	72	20	26	24
WDEN	20	11	17	14	33	76	100	40	45	20
GA	63	63	33	31	27	62	18	28	41	23
PGA	15	19	17	34	35	48	100	21	98	37
PSC	27	29	22	27	37	55	90	29	43	72
Decision-level fusion of all enhancement techniques	84	82	74	53	70	100	100	57	100	82

Table 2 gives the individual performance of the isolated digit recognition system for dysarthric speakers with 95% speech intelligibility by considering the perceptual features and vector quantization (VQ) models. Results show that the system provides excellent accuracy if the features and models are tested for speaker F05, diagnosed with 95% speech intelligibility. She is almost like an average speaker, and testing done using her speech utterances for all isolated digits provides exemplary accuracy. So, the speech

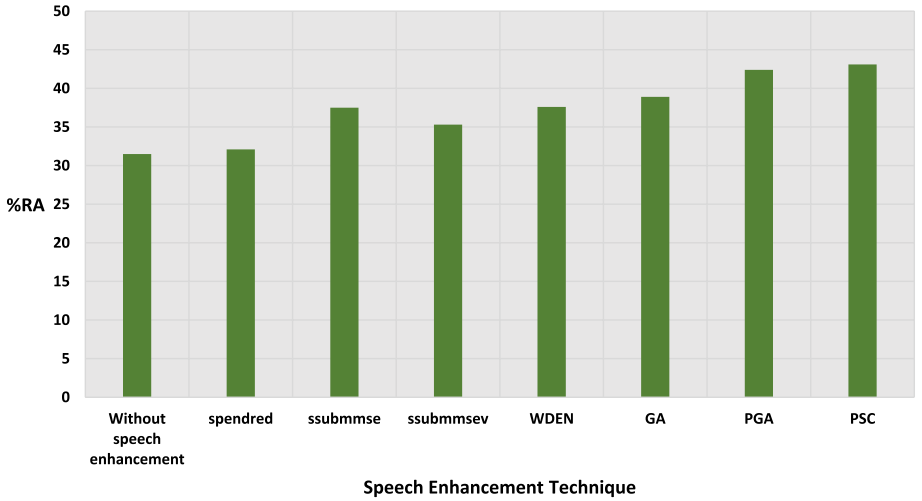

Fig. 21 Performance of the dysarthric speech recognition system –dysarthric speaker F03 (6% speech intelligibility)

utterances applied for evaluation must be acceptable, and the distortion level must be low.

3.1 Statistical Analysis and Validation of Experimental Results

The system's performance is statistically analyzed [23] to validate the perceptual features, models and speech enhancement techniques for recognizing dysarthric speakers' speeches. Table 3 indicates the usage of χ^2 a statistical distribution tool to analyze the experimental result. The number of test segments for concatenated test speech uttered by the dysarthric speaker in a pertinent digit is probable frequency. The actual frequency is the number of correctly identified test speech segments for each digit. Ten isolated digits are taken as ten attributes. Since the sample size is 100, χ^2distribution is applied to statistically analyze the choice of features, models and speech enhancement techniques. Hence, the rule of hypothesis based on χ^2distribution is framed as below:

H_0: Rejection rate is greater than or equal to 10%

H_1: Rejection rate is less than 10%

The individual χ^2test is applied at a 10% significance level, and the degree of freedom is considered nine $\chi^2 0.1 = 21.66$. Concerning the χ^2 table, the H_0 hypothesis is accepted.

Table 2 Performance of the system – Perceptual features and clustering –Female Speaker F05 (95% speech intelligibility)

Features and models	one	two	Three	four	Five	Six	Seven	eight	nine	zero	Average % RA
PLP-Kmeans	100	100	98	95	100	97	96	100	100	97	98.3
MFPLP-Kmeans	100	100	98	94	100	97	96	100	100	96	98.1
ERBPLP-Kmeans	100	100	100	94	98	97	96	100	100	96	98.1
All features – Kmeans	100	100	100	95	100	97	96	100	100	97	98.5

Table 4 indicates the system's statistical analysis for F05 speakers with perceptual features and clustering as a modelling technique with the hypothesis set below.

H_1: Digit recognition rate is $\geq$ 95%

H_0: Digit recognition rate is < 95%

The individual χ^2 test is applied at a 5% significance level, and the degree of freedom is considered nine $\chi^2 0.05 = 16.919$. The χ^2 table's calculated values are much less than the table value. Hence, the H_1 hypothesis is accepted. Subjective analysis is done to supplement the experimental dysarthric speech recognition system. Four average persons are asked to recognize the speeches uttered by dysarthric speakers. They are informed to listen to the isolated digits spoken by F03 and F05 dysarthric speakers. Tables 5 and 6 indicate the subjective analysis results for recognizing the numbers uttered by dysarthric speakers. Figure 22 and 23 show the comparative analysis between the experimental and manual assessment for identifying isolated digits spoken by F03 and F05 dysarthric speakers. The practical and subjective analysis would yield low accuracy since the F03 dysarthric speaker has 6% speech intelligibility. The experimental study is better than manual recognition for all the digits except 'zero'. It reveals that ensuring better performance for dysarthric speech recognition has been challenging. The comparative analysis in Fig. 23 indicates that the subjective assessment has yielded slightly better accuracy than the experimental assessment. F05 is a dysarthric female speaker with 95% speech intelligibility, so the accuracy of the decision-level practical classification and subjective analysis is very high. It is revealed that accuracy is directly proportional to the speakers' intelligibility level of the speeches uttered. In this work, speech enhancement techniques are implemented for improvement.

Since the speeches of F03 speakers with 6% speech intelligibility are highly distorted and disordered, it is cumbersome to ensure better accuracy. There are significant variations in style, difficulty level and pronunciation of words in the speeches of these speakers. However, if the speech intelligibility is good, their speeches can be classified without ambiguity. Adapting better speech enhancement mechanisms, features, and models would be a promising solution for ensuring better accuracy for speech-impaired whose impairment level is high. Table 7 depicts the comparative analysis between the existing works and our proposed work.

Table 3 Statistical analysis of isolated digits using decision level fusion classification by X^2 distribution test – F03 speaker (6% speech intelligibility)

Isolated digits	Observed frequency (O)	Expected frequency (E)	$(O-E)^2$	$(O-E)^2/E$
One	84	100	256	2.56
Two	76	100	576	5.76
Three	67	100	1089	10.89
Four	53	100	2209	22.09
Five	76	100	576	5.76
Six	100	100	0	0
Seven	100	100	0	0
Eight	86	100	196	19.6
Nine	100	100	0	0
Zero	86	100	196	19.6
				$\sum (O-E)^2/E = 86.26$

Table 4 Statistical analysis – Performance of the Decision level Fusion system – Perceptual features and clustering – F05 speaker (95% speech intelligibility)

Isolated digits	Observed frequency (O)	Expected frequency (E)	$(O-E)^2$	$(O-E)^2/E$
One	45	45	0	0
Two	51	51	0	0
Three	47	47	0	0
Four	57	60	9	0.15
Five	57	57	0	0
Six	59	61	4	0.066
Seven	44	46	4	0.087
Eight	54	54	0	0
Nine	51	51	0	0
Zero	64	66	4	0.06
				$\sum (O-E)^2/E = 0.363$

4 Conclusion

Since the speeches uttered by dysarthric people are severely distorted and degraded, it has become essential to improve the intelligibility of dysarthric spoken utterances. Subjective analysis of recognizing dysarthric spoken words reveals that human manual recognition is complex, especially those uttered by speakers with low speech intelligibility. The system uses perceptual features, speech enhancement techniques and statistical modelling methods. The proposed decision-level fusion system comprising features, models and speech enhancement techniques could improve accuracy for recognizing isolated digits uttered by dysarthric speakers. Accuracy is 81% for the decision-level fusion classifier for identifying numbers spoken by the dysarthric speaker with 6% speech intelligibility. However, this system has provided 99% accuracy in recognizing the isolated digits uttered by the dysarthric speaker with 95% speech intelligibility. Experimental results surpass the manual recognition of numbers uttered by a speaker with deficient speech intelligibility. However, manual credit has 100% accuracy for recognizing isolated digits spoken by a dysarthric speaker with 95% speech intelligibility. This system would provide accuracy if the system is trained using the database containing a more significant number of utterances spoken by more dysarthric speakers. This system can act as a translator for caretakers to understand dysarthric speakers' speeches to provide them with the necessary assistance. A robust speech translator may be designed to convert unintelligent spoken utterances into intelligible ones and interpret speeches uttered by dysarthric speakers that can be understandable. This work emphasizes the need for more efficient speech enhancement techniques to improve speech quality. It is proposed to strengthen the selection of features, speech enhancement and modelling techniques for the system's performance improvement.

Acknowledgements The authors thank the Department of Science & Technology, New Delhi, for the FIST funding (SR/FST/ET-I/2018/221(C). Furthermore, the authors also wish to thank the Intrusion Detection Lab at the School of Electrical & Electronics Engineering, SASTRA Deemed University, for providing infrastructural support to carry out this research work.

Funding FIST funding (SR/FST/ET-I/2018/221(C).

Table 5 Performance assessment by Subjective analysis – dysarthric speaker—F03 (6% speech intelligibility)

Normal speakers in India	Results	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Person1	No. Correctly Identified	2	0	7	3	0	2	3	4	7	3
	Total no. of utterances	7	7	7	7	7	7	7	7	7	3
	Accuracy	28.57%	0%	100%	42.85%	0%	28.57%	57.14%	57.14%	100%	100%
Person2	No. Correctly Identified	4	0	7	2	0	2	4	5	7	3
	Total no. of utterances	7	7	7	7	7	7	7	7	7	3
	Accuracy	57.14%	0%	100%	28.57%	0%	28.57%	57.14%	71.43%	100%	100%
Person3	No. Correctly Identified	3	0	7	2	0	2	3	5	7	3
	Total no. of utterances	7	7	7	7	7	7	7	7	7	3
	Accuracy	42.85%	0%	100%	28.57%	0%	28.57%	57.14%	71.43%	100%	100%
Person4	No. Correctly Identified	5	0	7	3	0	3	5	5	7	3
	Total no. of utterances	7	7	7	7	7	7	7	7	7	3
	Accuracy	71.43%	0%	100%	42.85%	0%	42.85%	71.43%	71.43%	100%	100%

Table 6 Performance assessment by Subjective analysis – dysarthric speaker—F05 (95% speech intelligibility)

Normal speakers in India	Results	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Person1	No. Correctly Identified	21	21	21	21	21	21	21	21	21	21
	Total no. of utterances	21	21	21	21	21	21	21	21	21	21
	Accuracy	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Person2	No. Correctly Identified	21	21	21	21	21	21	21	21	21	21
	Total no. of utterances	21	21	21	21	21	21	21	21	21	21
	Accuracy	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Person3	No. Correctly Identified	21	21	21	21	21	21	21	21	21	21
	Total no. of utterances	21	21	21	21	21	21	21	21	21	21
	Accuracy	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Person4	No. Correctly Identified	21	21	21	21	21	21	21	21	21	21
	Total no. of utterances	21	21	21	21	21	21	21	21	21	21
	Accuracy	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

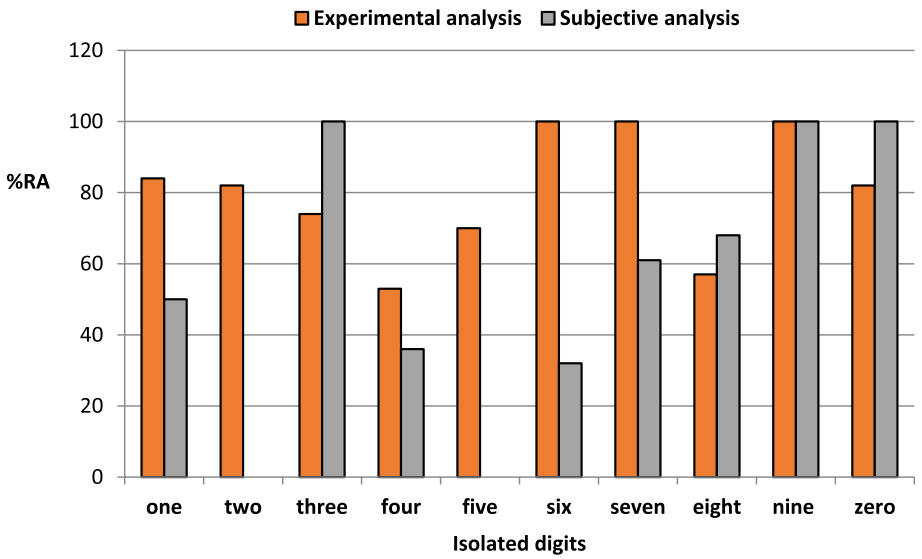

Fig. 22 Comparative analysis – Experimental and subjective assessment – F03 dysarthric speaker (6% speech intelligibility)

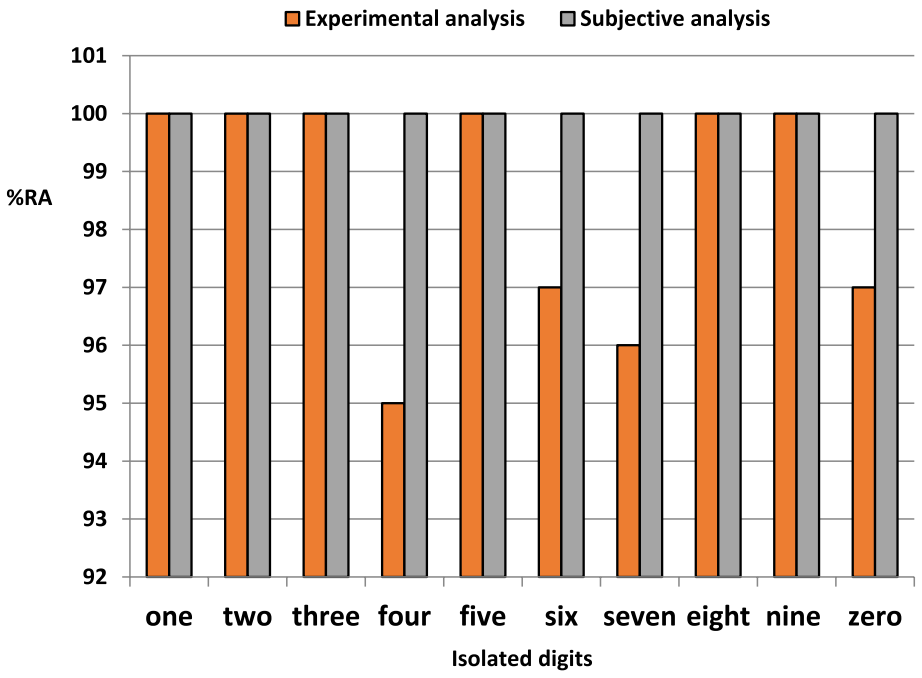

Fig. 23 Comparative analysis – Experimental and subjective assessment – F05 dysarthric speaker (95% speech intelligibility)

Table 7 Comparative analysis – Existing works and the proposed work

Reference	Database used	Technique used	% Accuracy
[35]	UA-speech database	Feature-MFCC Variational mode decomposition- speech enhancement CNN-classification	Common words-93.39 Computer commands-97.08 Digits- 98.33
[38]	Nemours corpus	Feature – MFCC Method- Empirical mode decomposition and CNN	64.86
Our proposed method	UA-speech database	Features –perceptual feature with filters in MEL, BARK and ERB scales Models – VQ, FCM and MHMM Method-Decision level fusion of features and models	For isolated digit recognition of F03 speaker with 6% speech intelligibility Without speech enhancement – 31.5 Fusion of speech enhancement techniques (spendred, ssubmmse, ssubmmsev, WDEN, GA, PGA, and PSC) -81

Data Availability The datasets generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant conflicts of interest to disclose.

Ethical Approval This article contains no studies with human participants or animals performed by authors.

References

1. Cespedes-Simangas, L., Uribe-Obregon, C., & Cabanillas-Carbonell, M. (2021). Analysis of speech therapy systems for children with physical disabilities and speech disorders: A systematic review. *European Journal of Molecular & Clinical Medicine, 8*(3), 2287–2301.
2. Takashima, Y., Takiguchi, T., & Ariki, Y. (2019). End-to-end dysarthric speech recognition using multiple databases. In ICASSP 2019–2019 IIEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6395–6399
3. Thoppil, M. G., Kumar, C. S., Kumar, A., & Amose, J. (2017). Speech signal analysis and pattern recognition in diagnosis of dysarthria. *Annals of Indian Academy of Neurology, 20*(4), 352.
4. Aihara, R., Takiguchi, T., & Ariki, Y. (2017). Phoneme-discriminative features for dysarthric speech conversion. In Interspeech, pp 3374–3378
5. Jiao, Y., Tu, M., Berisha, V., & Liss, J. (2018). Simulating dysarthric speech for training data augmentation in clinical speech applications. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6009–6013
6. Takashima, Y., Nakashika, T., Takiguchi, T., & Ariki, Y. (2015). Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition. In 2015 23rd European Signal Processing Conference (EUSIPCO), IEEE, pp 1411–1415
7. Espana-Bonet, C., & Fonollosa, J. A. (2016). Automatic speech recognition with deep neural networks for impaired speech. In Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23–25, 2016, Proceedings 3, Springer International Publishing, pp 97–107
8. Selouani, S. A., Dahmani, H., Amami, R., & Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. *International Journal of Speech Technology, 15*, 57–64.
9. Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language, 27*(6), 1163–1177.
10. Aihara, R., Takashima, R., Takiguchi, T., & Ariki, Y. (2014). A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(1), 1–10.
11. Rudzicz, F. (2010). Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(4), 947–960.
12. Tu, M., Berisha, V., & Liss, J. (2017). Interpretable objective assessment of dysarthric speech based on deep neural networks. In Interspeech, pp 1849–1853
13. Rudzicz, F. (2011). Acoustic transformations to improve the intelligibility of dysarthric speech. In Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pp 11–21
14. Lee, S. H., Kim, M., Seo, H. G., Oh, B. M., Lee, G., & Leigh, J. H. (2019). Assessment of dysarthria using one-word speech recognition with hidden markov models. *Journal of Korean Medical Science, 34*(13), 108.
15. Doire, C. S., Brookes, M., Naylor, P. A., Hicks, C. M., Betts, D., Dmour, M. A., & Jensen, S. H. (2016). Single-channel online enhancement of speech corrupted by reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(3), 572–587.
16. Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 33*(2), 443–445.

17. Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121.
18. Lallouani, A., Gabrea, M., & Gargour, C. S. (2004). Wavelet based speech enhancement using two different threshold-based denoising algorithms. In Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No. 04CH37513), IEEE, vol. 1, pp 315–318
19. Islam, M. T., Shahnaz, C., Zhu, W. P., & Ahmad, M. O. (2018). Enhancement of noisy speech with low speech distortion based on probabilistic geometric spectral subtraction. arXiv preprint arXiv:1802.05125.
20. Lu, Y., & Loizou, P. C. (2008). A geometric approach to spectral subtraction. *Speech communication, 50*(6), 453–466.
21. Stark, A. P., Wójcicki, K. K., Lyons, J. G., & Paliwal, K. K. (2008). Noise driven short-time phase spectrum compensation procedure for speech enhancement. In Ninth Annual Conference of the International Speech Communication Association
22. Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In Ninth Annual Conference of the International Speech Communication Association
23. Arunachalam, R. (2019). A strategic approach to recognize the speech of the children with hearing impairment: Different sets of features and models. *Multimedia Tools and Applications, 78*, 20787–20808.
24. Despotovic, V., Walter, O., & Haeb-Umbach, R. (2018). Machine learning techniques for semantic analysis of dysarthric speech: An experimental study. *Speech Communication, 99*, 242–251.
25. Narendra, N. P., & Alku, P. (2019). Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication, 110*, 47–55.
26. Narendra, N. P., & Alku, P. (2021). Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Computer Speech & Language, 65*, 101117.
27. Diwakar, G., & Karjigi, V. (2020). Improving speech to text alignment based on repetition detection for dysarthric speech. *Circuits, Systems, and Signal Processing, 39*, 5543–5567.
28. Cavallieri, F., Budriesi, C., Gessani, A., Contardi, S., Fioravanti, V., Menozzi, E., & Antonelli, F. (2021). Dopaminergic treatment effects on dysarthric speech: Acoustic analysis in a cohort of patients with advanced Parkinson's disease. *Frontiers in Neurology, 11*, 616062.
29. Hirsch, M. E., Lansford, K. L., Barrett, T. S., & Borrie, S. A. (2021). Generalized learning of dysarthric speech between male and female talkers. *Journal of Speech, Language, and Hearing Research, 64*(2), 444–451.
30. Hu, A., Phadnis, D., & Shahamiri, S. R. (2021). Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity. *Journal of Ambient Intelligence and Humanized Computing, 14*, 1–18.
31. Kodrasi, I. (2021). Temporal envelope and fine structure cues for dysarthric speech detection using CNNs. *IEEE Signal Processing Letters, 28*, 1853–1857.
32. Liu, S., Geng, M., Hu, S., Xie, X., Cui, M., Yu, J., & Meng, H. (2021). Recent progress in the CUHK dysarthric speech recognition system. *IEEE ACM Transactions on Audio, Speech, and Language Processing, 29*, 2267–2281.
33. Liu, Y., Penttilä, N., Ihalainen, T., Lintula, J., Convey, R., & Räsänen, O. (2021). Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2228–2243.
34. Dhanalakshmi, M., Nagarajan, T., & Vijayalakshmi, P. (2021). Significant sensors and parameters in assessment of dysarthric speech. *Sensor Review, 41*(3), 271–286.
35. Rajeswari, R., Devi, T., & Shalini, S. (2022). Dysarthric speech recognition using variational mode decomposition and convolutional neural networks. *Wireless Personal Communications, 122*(1), 293–307.
36. Tripathi, A., Bhosale, S., & Kopparapu, S. K. (2021). Automatic speaker independent dysarthric speech intelligibility assessment system. *Computer Speech & Language, 69*, 101213.
37. Zaidi, B. F., Selouani, S. A., Boudraa, M., & Sidi Yakoub, M. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. *Neural Computing and Applications, 33*, 9089–9108.
38. Sidi Yakoub, M., Selouani, S. A., Zaidi, B. F., & Bouchair, A. (2020). Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP Journal on Audio, Speech, and Music Processing, 2020*(1), 1–7.
39. Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in Computer Science, 4*, 770210.
40. Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2022). Synthesizing dysarthric speech using multi-talker TTS for dysarthric speech recognition. arXiv preprint arXiv:2201.11571.

41. Ren, J., & Liu, M. (2017). An automatic dysarthric speech recognition approach using deep neural networks. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2017.081207
42. Harvill, J., Issa, D., Hasegawa-Johnson, M., & Yoo, C. (2021). Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. In ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 6428–6432
43. Sekhar, S. M., Kashyap, G., Bhansali, A., & Singh, K. (2022). Dysarthric-speech detection using transfer learning with convolutional neural networks. *ICT Express, 8*(1), 61–64.
44. Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29*, 852–861.
45. Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., & Alibakhshikenari, M. (2023). Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. *Sensors, 23*(13), 6212.
46. Shih, D. H., Liao, C. H., Wu, T. W., Xu, X. Y., & Shih, M. H. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *In Healthcare, 10*(10), 1956.
47. Hall, K., Huang, A., & Shahamiri, S. R. (2023). An investigation to identify optimal setup for automated assessment of dysarthric intelligibility using deep learning technologies. *Cognitive Computation, 15*(1), 146–158.
48. Latha, M., Shivakumar, M., Manjula, G., Hemakumar, M., & Kumar, M. K. (2023). Deep learning-based acoustic feature representations for dysarthric speech recognition. *SN Computer Science, 4*(3), 272.
49. Yu, C., Su, X., & Qian, Z. (2023). Multi-stage audio-visual fusion for dysarthric speech recognition with pretrained models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31*, 1912–1921.
50. Revathi, A., Sasikaladevi, N., & Arunprasanth, D. (2022). Development of CNN-based robust dysarthric isolated digit recognition system by enhancing speech intelligibility. *Research on Biomedical Engineering, 38*(4), 1067–1079.
51. Almadhor, A., Irfan, R., Gao, J., Saleem, N., Rauf, H. T., & Kadry, S. (2023). E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications, 222*, 119797.
52. Jolad, B., & Khanai, R. (2023). An approach for speech enhancement with dysarthric speech recognition using optimization based machine learning frameworks. *International Journal of Speech Technology, 26*, 287–305.

Dr. A. Revathi has obtained B.E (ECE), M.E (Communication Systems), and Ph.D (Speech Processing) from National Institute of Technology, Tiruchirappalli, Tamilnadu, India in 1988, 1993 and 2009 respectively. She has been serving in Electronics and Communication Engineering for 33 years and currently working as a Professor in the Department of ECE, SASTRA Deemed University, Thanjavur, India. She has published 61 technical articles in Reputed International journals and presented papers in IEEE and Springer conferences. Her areas of interest include Speech processing, Signal processing, Image processing, Bio metrics & Security, Communication Systems, Embedded Systems and Computer Networks.

Dr. N. Sasikaladevi working as a faculty at SASTRA Deemed University since 2014. She authored a book titled "Programming in C#.NET" published by Prentice Hall of India. She has presented papers in IEEE, Springer and Elsevier conferences. She published 55 papers in various reputed journals. She received young professional award from CSI, India. She got young scientist award and woman scientist award from Department of Science and Technology, Government of India. Her research interest includes curve based cryptography, quantum cryptography and network security.

Dr. D. Arunprasanth has completed his bachelor degree in medicine at Government Medical College, Theni, Tamilnadu, India in 2016. Now Assistant Professor Thanjavur Medical College. He has served as an Assistant Surgeon at Primary health Centre, Tamilnadu, India, for two years. He has been selected by the Government to undergo training in Anesthesia at Government hospital, Trichy in 2017. During the training, he has been adjudged as one of the best students by the panel of experts who conducted the Viva-voce examination. He has got very good score in PG NEET examination and currently pursuing M.D (Pediatrics) at Government Medical College, Madurai. He has one paper in his credit as poster presentation in the conference. His areas of interest include Pediatric medicine, Pediatric Intensive care and Neonatology.

Dr. Rengarajan Amirtharajan received his B.E. degree from P.S.G. Tech. Bharathiyar University, Coimbatore, India in 1997. He received M.Tech. and PhD. Degrees from SASTRA Deemed University, Thanjavur, India in 2007 and 2012, respectively. He is currently working as Professor in SEEE, SASTRA Deemed University. He patented a novel embedding scheme USPTO in March 2015. He has also published more than 314 research articles in National & International journals.