Research on Biomedical Engineering (2022) 38:1067–1079
https://doi.org/10.1007/s42600-022-00239-7

ORIGINAL ARTICLE

Development of CNN-based robust dysarthric isolated digit recognition system by enhancing speech intelligibility

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

Received: 6 June 2022 / Accepted: 29 August 2022 / Published online: 13 September 2022

Abstract

Purpose Developing a computer-assisted speech training/recognition system for recognizing the speeches of dysarthric speakers has become necessary because their speeches are highly distorted due to the motor disorder in their articulatory mechanism.

Methods In this work, two-dimensional spectrograms in BARK and MEL scale and Gammatonegram are used as features to tune the convolutional neural network (CNN) architecture designed to perform the dysarthric speech recognition.

Results Overall recognition accuracy is 88%, 97.9%, and 98% for the CNN-based dysarthric speech recognition system using Gammatonegram, spectrogram, and Melspectrogram, respectively. However, decision-level fusion of these features results has yielded 99.72% overall accuracy with 100% individual accuracy for some of the dysarthric isolated digits. This work is extended to have a phase spectrum compensation technique to improve the intelligibility of dysarthric speeches, and the decision-level fusion classifier provides relatively better accuracy of 99.92% for classifying isolated digits spoken by dysarthric speakers.

Conclusion This work can be utilized to recognize the distorted speeches of dysarthric speakers like normal speeches.

Keywords Dysarthric speech recognition · CNN · Machine learning · Spectrogram · Time–frequency representation

Introduction

Speech is the sequence of sounds being considered an output of the time-varying vocal tract system. Articulators are moving in response to the neural signals for producing regular speech. Dysarthric speeches are distorted because persons with dysarthria are affected by a motor speech disorder, and they cannot control the movement of articulators. As a result, dysarthric speakers experience impediments in speaking properly. Articulators and muscles involved in speech production mechanisms are damaged or paralyzed for dysarthric speakers, and they find difficulty in conveying information to others through speech. The speech intelligibility of dysarthric speakers (Kim et al. 2008) considered in our work ranges between 2 and 95%. A dysarthria severity level (Gupta et al. 2021) is assessed using short speech segments based on residual neural networks. Dysarthric severity classification (Joshy and Rajan 2021) uses deep neural networks with speech utterances from Torgo and UA-speech databases. Articulatory features and deep CNN (Emre et al. 2019) are used for developing speaker-independent speech recognition systems for dysarthric speakers.

The dysarthric speech recognition system (Kim et al. 2018) is implemented using Mel frequency cepstral coefficients (MFCC) and assessed using GMM-HMM, DNN-HMM, CNN-HMM, and CLASM-HMM classifiers. Dysarthric speech recognition (Albaqshi and Sagheer 2020) is done using MFCC and convolutional recurrent neural networks for the Torgo database. Listen, Attend and Spell (LAS) model (Takashima et al. 2019) is investigated for dysarthric speech recognition, and the performance metric used is character error rate (CER). The intelligibility of dysarthric speeches (Chen et al. 2020) is enhanced using a gated CNN-based voice conversion system. An automatic speech recognition system is developed for dysarthric speakers.

✉ A. Revathi
revathi@ece.sastra.edu

[1] School of EEE, SASTRA Deemed University, Thanjavur 613 401, India

[2] School of Computing, SASTRA Deemed University, Thanjavur 613 401, India

[3] Institute of Child Health & Research Centre, Madurai Medical College, Madurai 625020, India

The accuracy of dysarthric speech recognition (Sidi Yakoub et al. 2020) is improved using empirical mode decomposition and CNN. This work mainly uses the speech enhancement technique to improve the accuracy of the dysarthric isolated digit recognition system. It utilizes deep machine learning neural network models for template creation and testing for original dysarthric speeches and intelligibility-enhanced speeches using phase spectrum compensation (PSC) as a speech enhancement mechanism. This paper is organized as follows. 'Development of dysarthric speech recognition system' section describes the database used in our work and analyses normal and dysarthric speech in time, frequency, and time–frequency domains. 'Implementation of the CNN-based dysarthric speech recognition' section describes the methods for implementing the system with feature extraction, CNN-based model development, and the speech enhancement technique used. Experimental results based on the proposed features and CNN-based system are presented in the 'Results of the dysarthric speech recognition system based on experiments conducted' section. 'Discussion based on the outcome of the experiments' section illustrates the discussion on the experimental results. The conclusion of the work is summarized in the 'Conclusions' section.

Development of dysarthric speech recognition system

Dysarthric speech recognition is developed to recognize the speeches uttered by dysarthric speakers. Since a dysarthric speaker's speeches are highly distorted, developing a speech recognition system is gaining paramount importance, and it is quite challenging to develop a robust system. This section illustrates the details of the database used and analysis of dysarthric speeches in time and frequency domains.

Details of the database used—dysarthric speaker information (Kim et al. 2008)

Table 1 indicates the speaker information considered in our study. An isolated digit recognition system is developed to recognize the digits uttered by dysarthric speakers. These speakers are diagnosed as patients with spastic. Intelligibility levels vary between 2 and 95%.

Analysis of speech in time, frequency, and time–frequency domains

Speeches uttered by normal and dysarthric speakers are analysed in time, frequency, and time–frequency domains. For example, Fig. 1 presents the analysis of speech signals uttered by a normal speaker in the time, frequency, and time–frequency domain. He takes less than a second (0.84 s) to utter this word.

Table 1 Information—speakers considered for the current study

Speaker	Age	Speech intelligibility	Dysarthria diagnosis
M09	18	High (86%)	Spastic
M07	58	Low (28%)	Spastic
M04	> 18	Very low (2%)	Spastic
M01	> 18	Very low (10%)	Spastic
F05	22	High (95%)	Spastic
F03	51	Very low (6%)	Spastic

Analysis of speech uttered by dysarthric speaker M09 with 86% speech fluency in time, frequency, and time–frequency domain is indicated in Fig. 2. Since this dysarthric speaker with 86% intelligibility, there are more similarities in signal characteristics as compared to the speech uttered by the normal speaker. For example, this speaker takes 2.15 s to utter the digit 'one'.

Figure 3 indicates the analysis of speech uttered by the dysarthric speakers with 6% intelligibility in the time, frequency, and time–frequency domain. A dysarthric speaker utters this isolated digit with 6% intelligibility, and this fact is demonstrated in signal characteristics as compared to that of the normal speaker. This speaker takes 2.6 s to utter the isolated digit 'one'.

This analysis indicates that dysarthric speakers take more time to speak simple words, and the severity level of the dysarthric speakers affects their ability to speak to a larger extent. As a result, it takes longer for them to speak. So, there is a need to develop an automated system to recognize their speeches and become a translator.

Implementation of the CNN-based dysarthric speech recognition

Dysarthric speech recognition is implemented by considering two phases: training and testing. During the training phase, features are extracted from the speeches uttered by the dysarthric speakers, the application of features to the modelling techniques to create templates as representative models of speeches, and models are fine-tuned for speech recognition. During testing, features are extracted from the speeches earmarked for testing. These features are applied to the models. Depending on the classifier used, speech is recognized as associated with the model in a pertinent isolated digit.

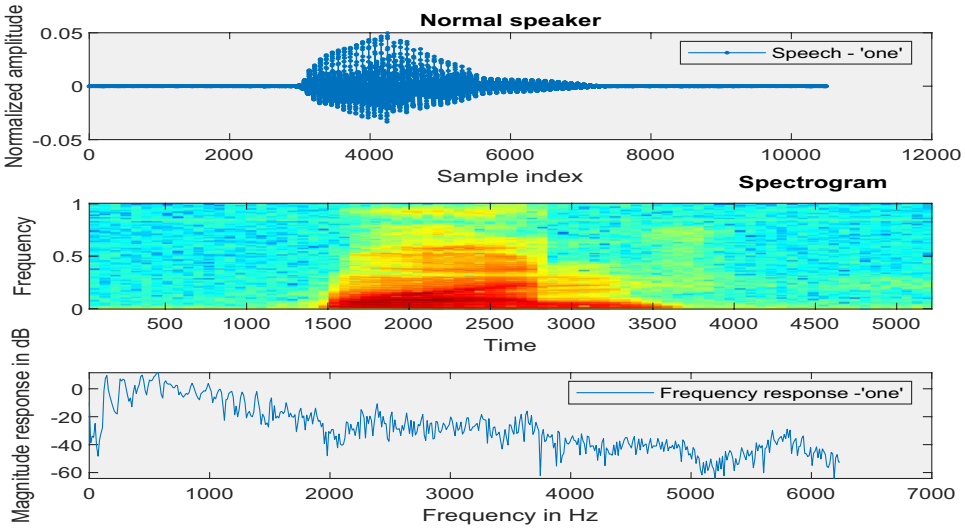

Fig. 1 Normal speaker—analysis of speech in time, frequency, and time–frequency domain

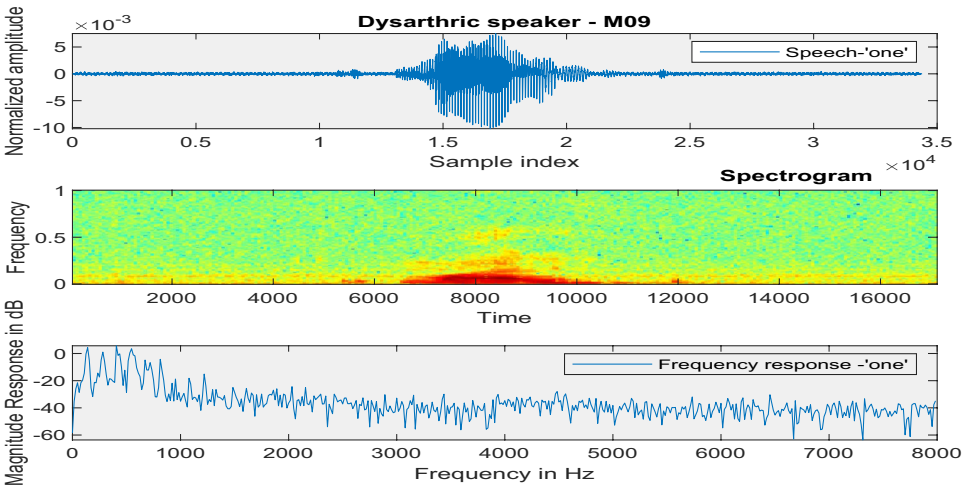

Fig. 2 Dysarthric speaker M09 with 86% intelligibility—analysis of speech in time, frequency, and time–frequency domain

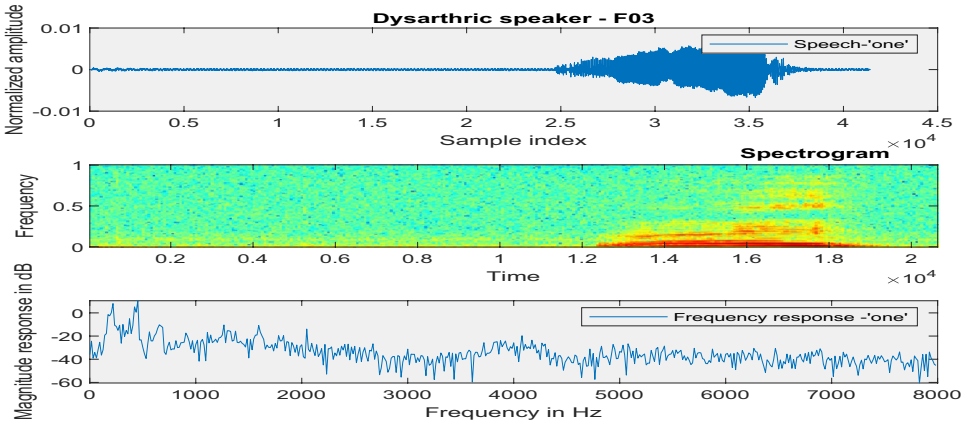

Fig. 3 Speech—dysarthric speaker F03 with 6% intelligibility—analysis in time, frequency, and time–frequency domain

Feature extraction phase

In this work on CNN-based dysarthric speech recognition, time–frequency representational features are used to fine-tune the CNN models. Features extracted should have high discriminating capability among the speeches considered. Speech utterances in a pertinent isolated digit are concatenated, and spectrogram, Melspectrogram, and Gammatonegram are extracted for the speech frames containing 8192 samples, and this process is repeated for

every 256 samples. The block schematic used for feature extraction is depicted in Fig. 4.

Eighty percent of the features are used for training and the remaining 20% for testing. Size of the spectrogram, Melspectrogram, and Gammatonegram feature is [129,127], [64,127], and [64,49], respectively. For one frame, spectrogram, Melspectrogram, and Gammatonegram are plotted as in Figs. 5, 6, and 7.

Development of CNN templates

Spectrogram, Melspectrogram, and Gammatonegram two-dimensional feature sets for each isolated digit are applied to the CNN network. Network models are fine-tuned to perform speech recognition for dysarthric speakers. Table 2 describes the CNN layered architecture (Soliman et al. 2021 J. Zhang et al. 2017, Arias-Vergara et al. 2021, Vavrek et al. 2021, Sangwan et al. 2020, Chen et al. 2020, P. H. Binh et al. 2021) for Gammatonegram-based dysarthric isolated digit recognition implemented as a work.

Similar CNN architecture is implemented with variations in the image input size [129, 127, and 1] for spectrogram and [64, 127, and 1] for Melspectrogram-based CNN networks. Figure 8 indicates the modules used for creating group CNN templates for the proposed features.

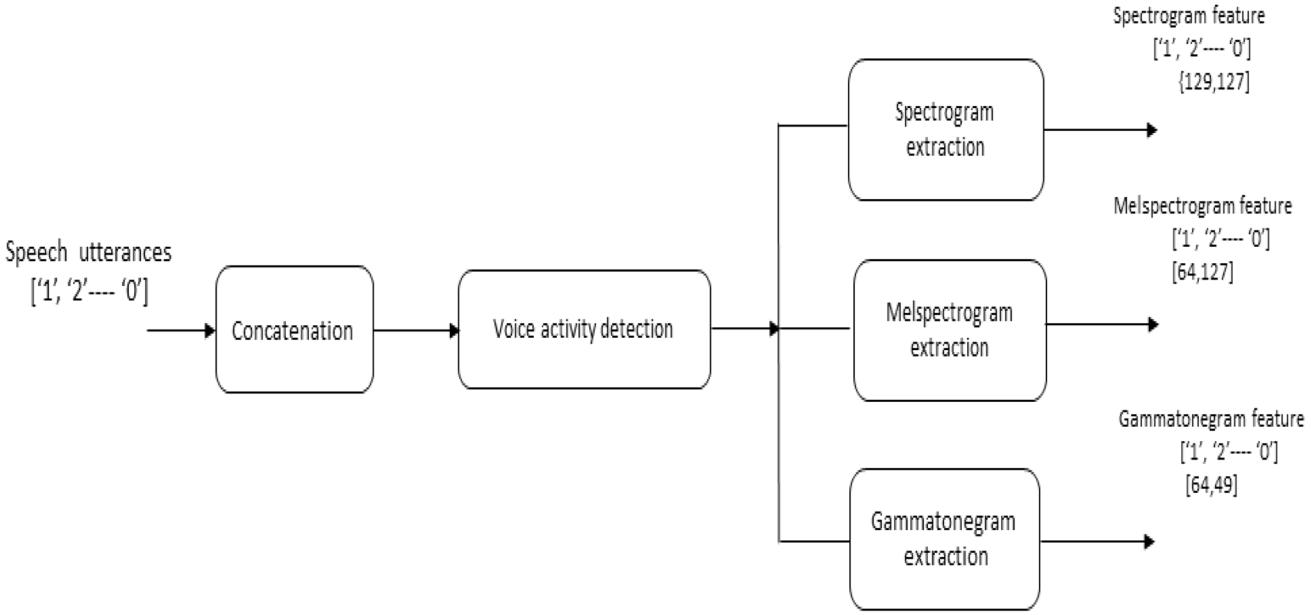

Fig. 4 Feature extraction phase

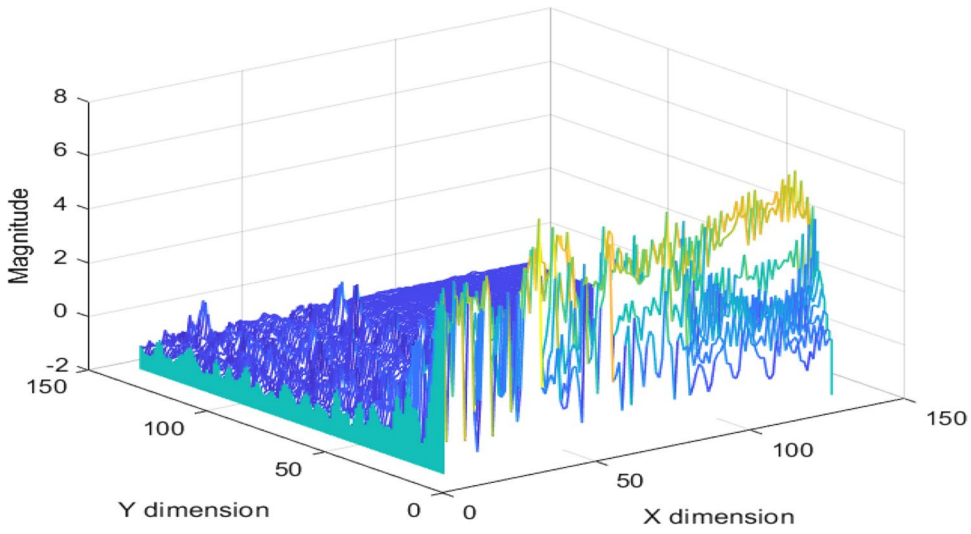

Fig. 5 Spectrogram—waterfall plot—dysarthric speech—isolated digit 'one'

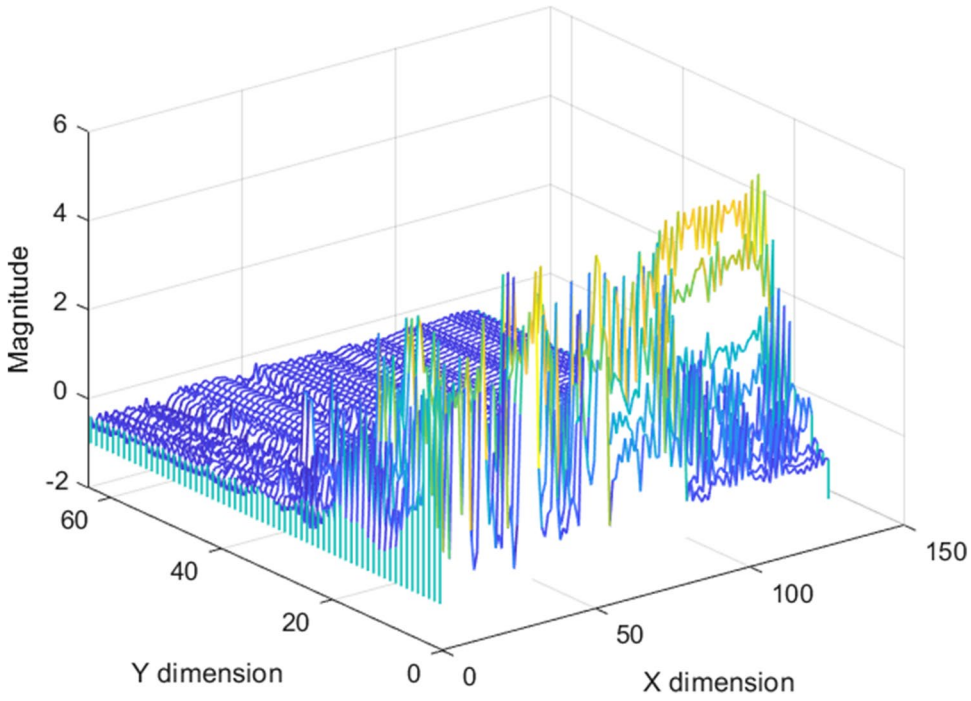

Fig. 6 Melspectrogram—waterfall plot—dysarthric speech—isolated digit 'one'

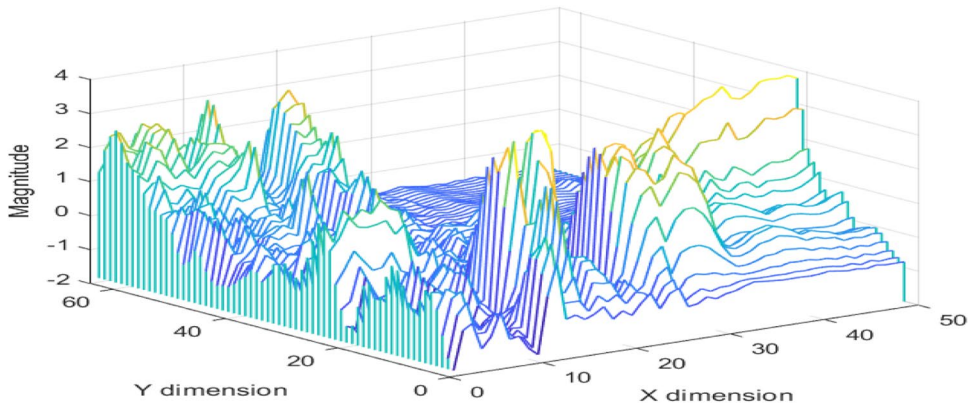

Fig. 7 Gammatonegram—waterfall plot—dysarthric speech—isolated digit 'one'

Table 2 CNN layered architecture—Gammatonegram—dysarthric isolated digit recognition

15×1 Layer array with layers:		
1	'' Image Input	64×49×1 images with 'zerocenter' normalization
2	'' Convolution	7 3×3 convolutions with stride [1 1] and padding 'same'
3	'' Batch Normalization	Batch normalization
4	'' ReLU	ReLU
5	'' Max Pooling	2×2 max pooling with stride [2 2] and padding [0 0 0 0]
6	'' Convolution	14 7×7 convolutions with stride [1 1] and padding 'same'
7	'' Batch Normalization	Batch normalization
8	'' ReLU	ReLU
9	'' Max Pooling	7×7 max pooling with stride [7 7] and padding [0 0 0 0]
10	'' Convolution	28 3×3 convolutions with stride [1 1] and padding 'same'
11	'' Batch Normalization	Batch normalization
12	'' ReLU	ReLU
13	'' Fully Connected	10 fully connected layer
14	'' Softmax	softmax
15	'' Classification Output	crossentropyex

Phase spectrum compensation-based speech enhancement (Stark et al. 2008)

In this method, the modified phase response is combined with a magnitude response to get the changed frequency response for the noisy speech. Analysing the relation between spectral-domain and time-domain 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 (1).

$$\mathrm{Y_n(k)} = \left|\mathrm{Y_n(k)}\right| \mathrm{e}^{\mathrm{j}\angle \mathrm{Y_n(k)}} \tag{1}$$

The compensated short-time phase spectrum is computed by using Eqs. (2) and (3).

The process obtains phase spectrum compensation function as in Eq. (2).

$$\wedge_n(k) = \lambda\psi(k)\left|D_n(k)\right| \tag{2}$$

$\left|D_n(k)\right|$ specifies magnitude response of the noise signalλ—constant

The anti-symmetry function $\psi(k)$ is defined as in (3).

$$\psi(k) = \begin{cases} 1 \; if \; 0 < \frac{k}{N} < 0.5 \\ -1 \; if \; 0.5 < \frac{k}{N} < 1 \end{cases} \tag{3}$$

Multiplication of symmetric magnitude spectra of the noise signal with anti-symmetric function $\psi(k)$ 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. (4).

$$Y_n(k) = X_n(k) + \wedge_n(k) \tag{4}$$

The compensated phase spectrum of the noisy signal is derived as in Eq. (5).

$$\angle Y_n(k) = ARG[Y_n(k)] \tag{5}$$

Recombination of the compensated phase response with 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 (7) on the modified spectral response given in (6).

$$S_n(k) = \left|Y_n(k)\right| e^{j\angle Y_n(k)} \tag{6}$$

$$s(n) = real[inverse\ STFT(S_n(k))] \tag{7}$$

Figure 9 indicates the performance of the speech enhancement technique by phase compensation.

Results of the dysarthric speech recognition system based on experiments conducted

Speech utterances in pertinent isolated digits are concatenated, and spectrogram, Melspectrogram, and Gammatonegram two-dimensional features are extracted after voice activity detection for original raw dysarthric speeches. Similarly, speech intelligibility is improved using phase spectrum compensation as a speech enhancement technique on raw dysarthric speeches. The proposed time–frequency representational features are extracted for the enhanced dysarthric speeches. Eighty percent of the features have been used for training. Twenty percent of the features are considered for testing. These features are given to the CNN models. Based on the matching, each row of the test vectors is associated with one of the groups in training. Group can be categorical

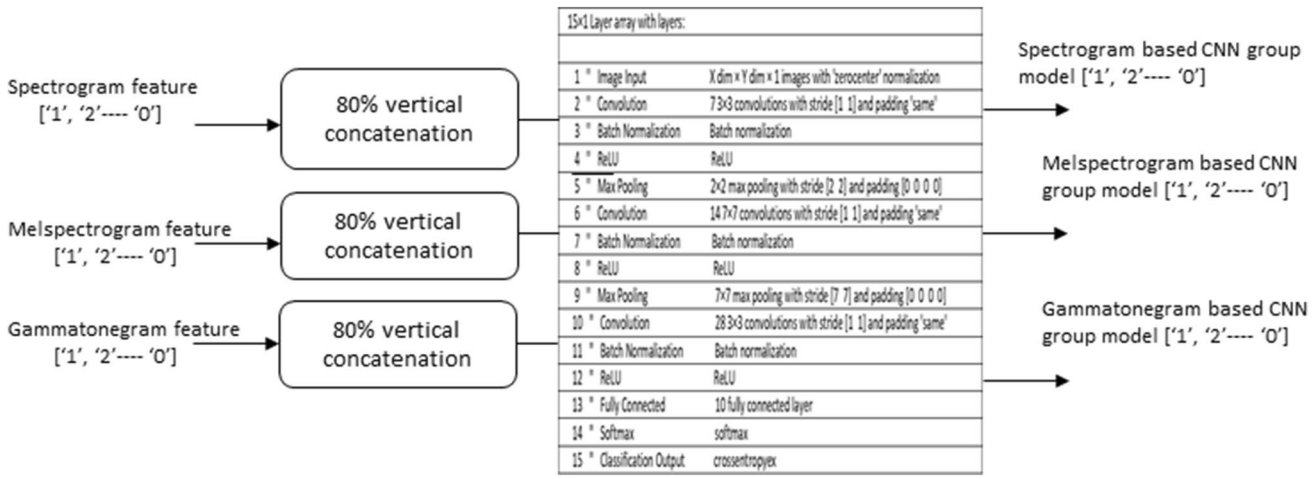

Fig. 8 CNN template creation

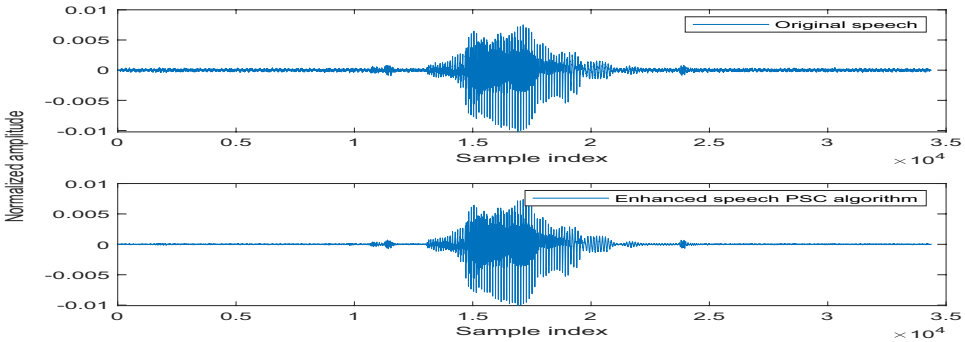

Fig. 9 Illustration of speech enhancement technique—phase spectrum compensation

string of indices of isolated digits '1', '2', '3', '4', '5', '6', '7', '8', '9', '0'. The confusion graph in Fig. 10 indicates the performance of the Gammatonegram and CNN-based dysarthric isolated digit recognition without applying PSC for speech enhancement. The overall average system accuracy is 88.3%, with a relatively low accuracy of 83% for recognizing the digit '2'.

The confusion chart shown in Fig. 11 depicts spectrogram and CNN-based dysarthric isolated digit recognition system without PSC. The average accuracy is 97.8%.

The confusion chart depicted in Fig. 12 demonstrates the performance of the dysarthric isolated digit recognition for the Melspectrogram and CNN-based system. The average accuracy is 98%.

Decision-level fusion of results of three spectrograms for the CNN-based system is done, and Table 3 reveals the performance of the decision-level fusion system. Figure 13 indicates the proposed decision-level fusion system. The overall accuracy of the decision-level fusion classifier is 99.72%.

Figures 14, 15, and 16 indicate the performance of the spectrogram, Melspectrogram and Gammatonegram, and CNN-based system by using PSC as an enhancement technique on the raw dysarthric speeches for improving Intelligibility.

The average recognition accuracy for Gammatonegram, spectrogram and Melspectrogram, and CNN-based system with PSC as speech enhancement technique is 96.67%, 99.46%, and 98.76%, respectively. Table 4 indicates the performance of the dysarthric isolated digit recognition system for PSC as a speech enhancement technique with a decision-level fusion of results corresponding to the two-dimensional features such as spectrogram, Melspectrogram, and Gammatonegram.

Overall average accuracy is 99.92%. Figure 17 indicates the comparative performance of the system with and without the speech enhancement technique.

This work on isolated digit recognition is extended to perform connected word recognition. Twenty related words spoken by dysarthric speakers are taken, and Fig. 18 indicates the confusion charts for connected word recognition using the proposed features and CNN-based systems. The performance of the decision-level fusion of correct indices on the features and CNN-based systems is indicated in Fig. 19. This system is evaluated using PSC for speech enhancement, and the accuracy is good.

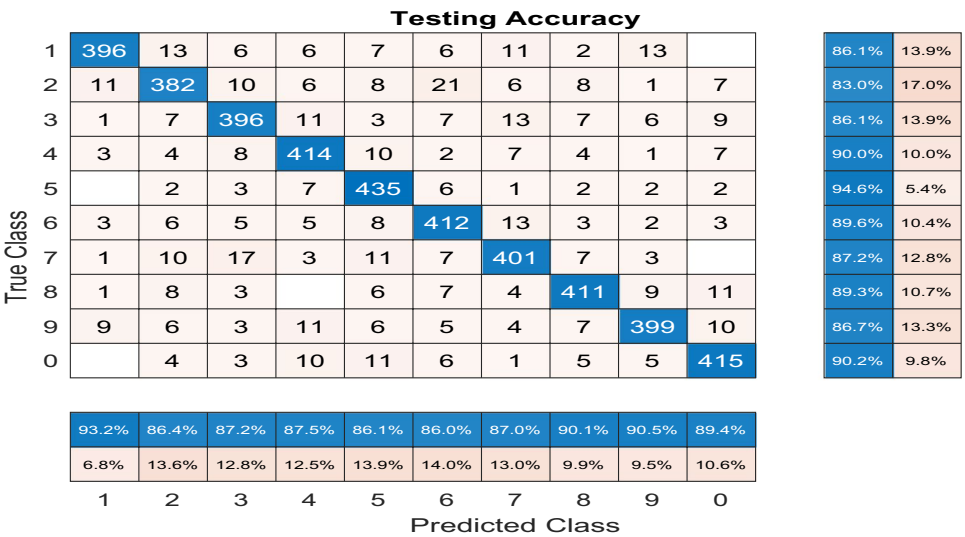

Fig. 10 Confusion chart—Gammatonegram and CNN-based system (without PSC)

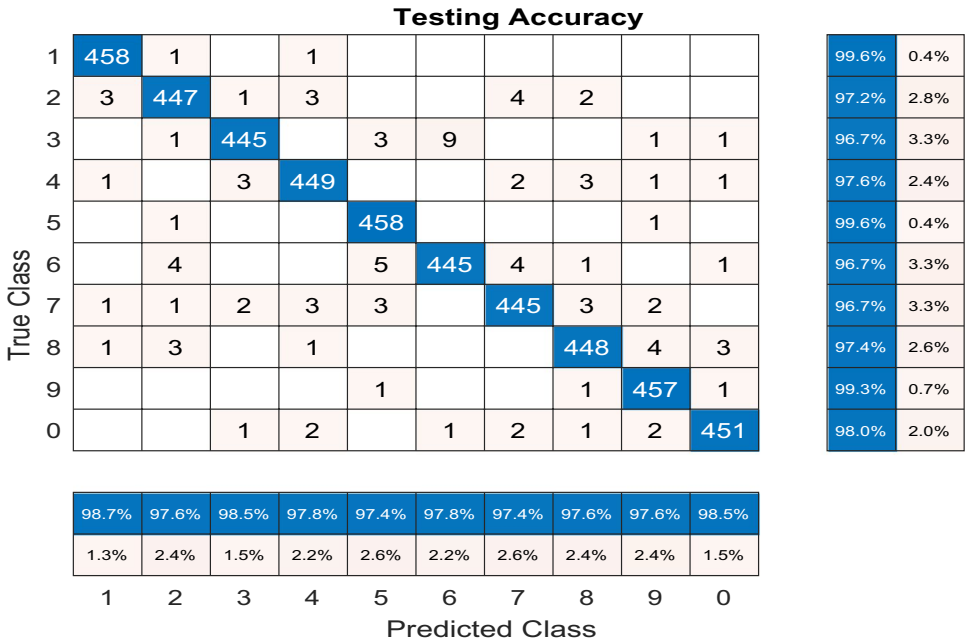

Fig. 11 Performance chart—spectrogram and CNN-based system (without PSC)

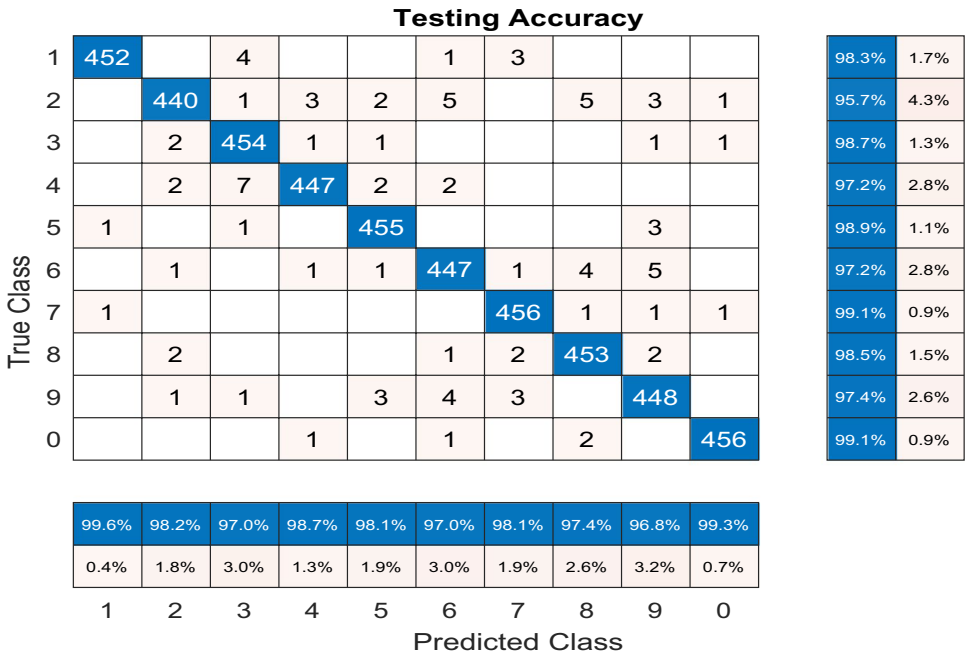

Fig. 12 Performance chart—Melspectrogram and CNN-based system (without PSC)

Table 3 Performance assessment—decision-level fusion of spectrograms and CNN-based system (without PSC)

Isolated digits	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero
%RA	100	99.2	99.6	99.4	100	99.6	100	99.8	99.8	99.8

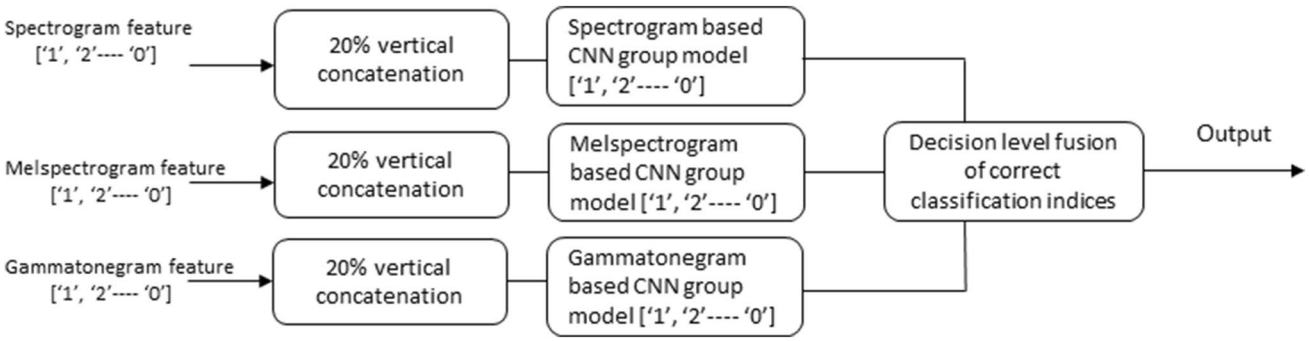

Fig. 13 Decision-level fusion classifier

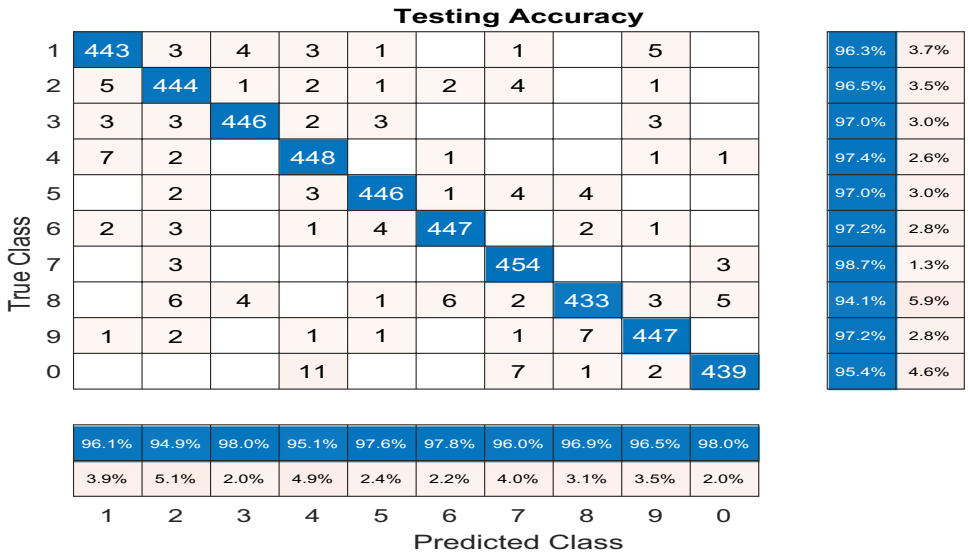

Fig. 14 Performance chart—Gammatonegram and CNN-based system with PSC for speech enhancement

Discussion based on the outcome of the experiments

In this work on dysarthric speech recognition, speech utterances of dysarthric speakers are split into two sets, each for training and testing. Spectrogram, Melspectrogram, and Gammatonegram features are extracted from basic training speeches, and CNN templates are created for each isolated digit based on the pertinent input features. Test sets of utterances in each isolated digit are tested, and the system's performance is analysed based on three different two-dimensional spectrogram features with CNN for modelling and classification. Out of the features used, the overall accuracy of the system for spectrogram and Melspectrogram is the same. Another experiment is conducted on the intelligibility improved speeches of dysarthric speakers with an application of the phase spectrum compensation technique for speech enhancement. Spectrogram-based feature selection is better in terms of attaining good accuracy as compared to other features. Decision-level fusion of outcome of the experiments for the features with and without speech enhancement technique proves to be good in attaining the very good accuracy of 99.92% for all the isolated digits uttered by dysarthric speakers. Table 5 gives the comparative analysis of the proposed work with existing works mentioned in the literature.

Conclusions

In this paper, the development of a speech recognition system for recognizing the isolated digits uttered by dysarthric speakers is analysed and assessed using two-dimensional spectrogram features and deep CNN. Spectrogram, Melspectrogram, and Gammatonegram features are extracted from the speech utterances corresponding to the speech utterances of isolated digits ['1', '2', '3', '4', '5', '6', '7', '8', '9', '0']. Eighty percent of the derived two-dimensional time–frequency representational features

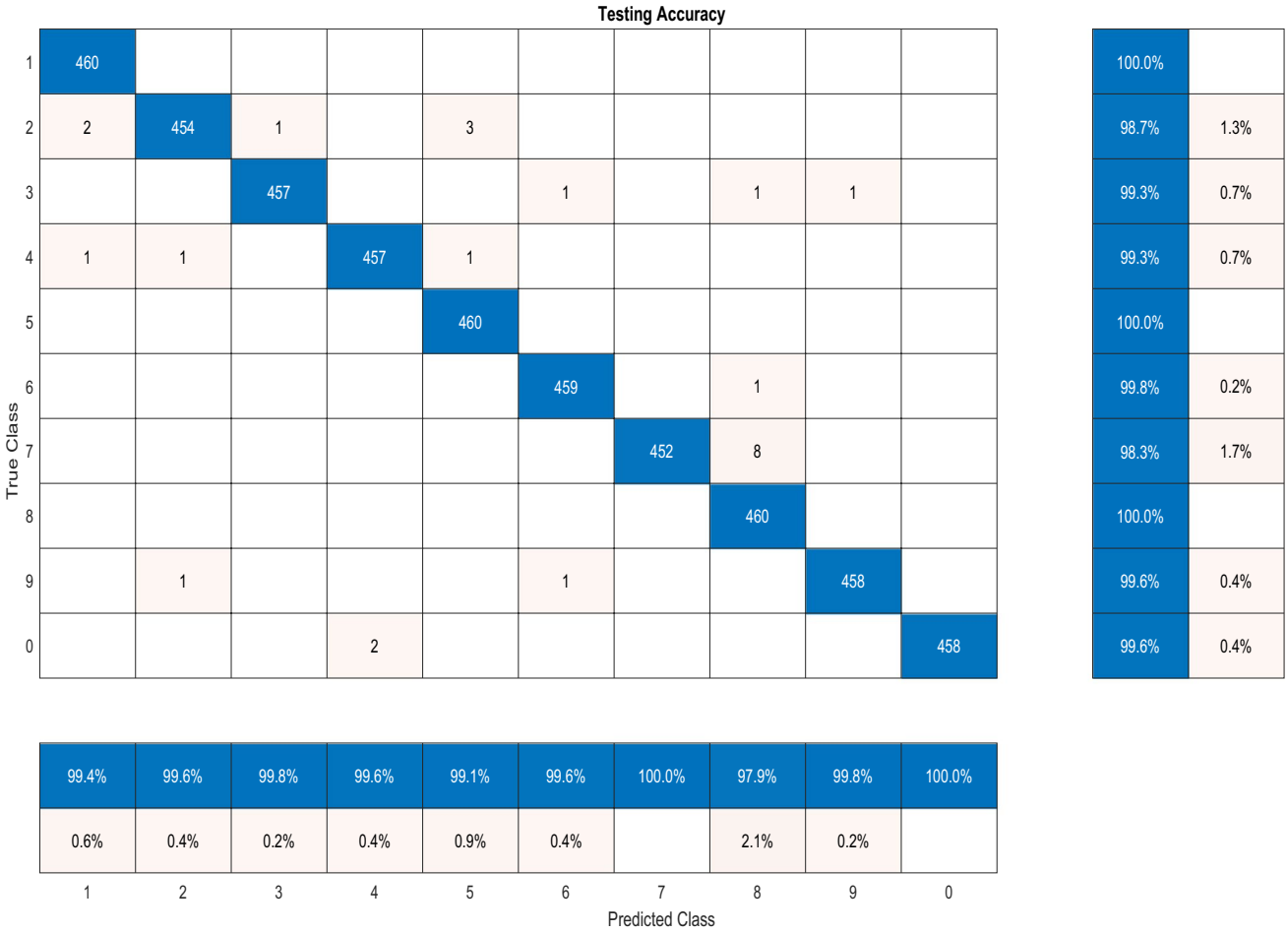

Fig. 15 Performance chart—spectrogram and CNN-based system with PSC for speech enhancement

Fig. 16 Performance chart—Melspectrogram and CNN-based system with PSC for speech enhancement

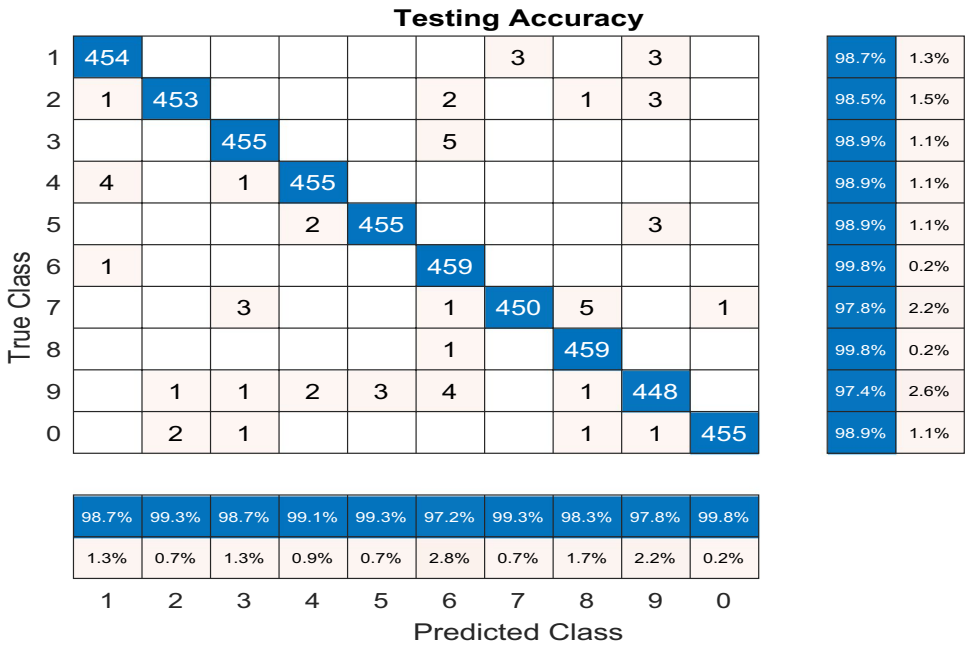

Table 4 Performance assessment—dysarthric digit recognition with PSC for speech enhancement and decision-level fusion classifier

Isolated digits	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero
%RA	100	100	99.78	100	100	100	99.57	100	100	99.78

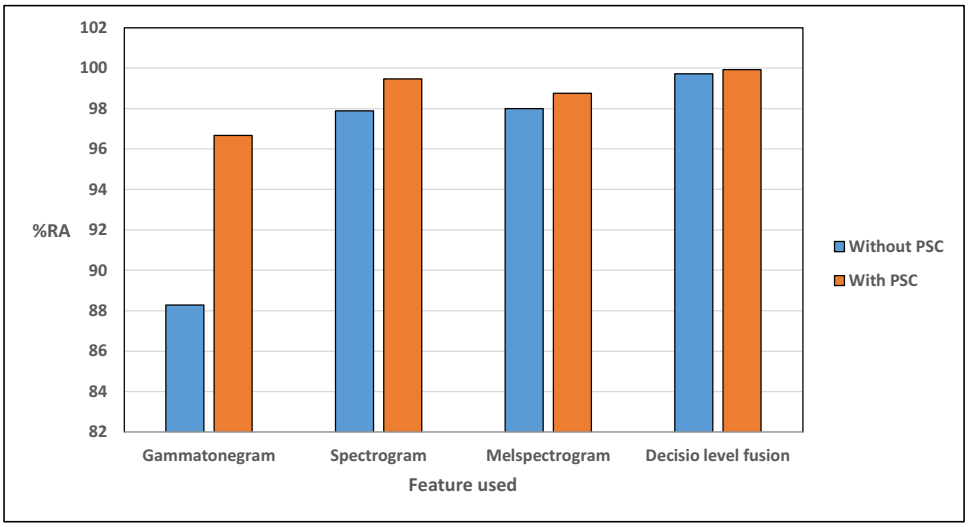

Fig. 17 Comparative analysis with and without PSC speech enhancement technique

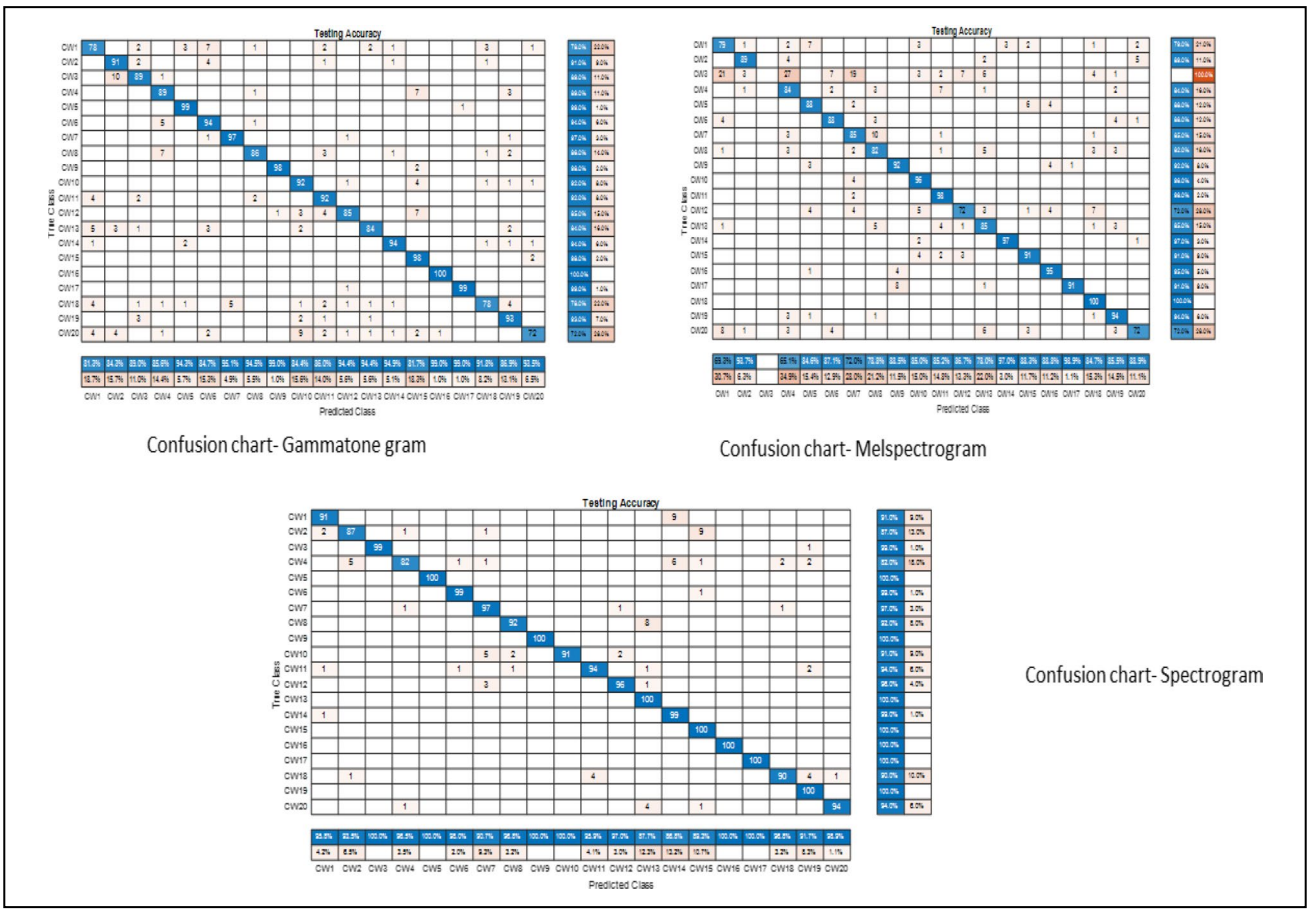

Fig. 18 Results—confusion chart—proposed features and CNN-based systems—connected word recognition

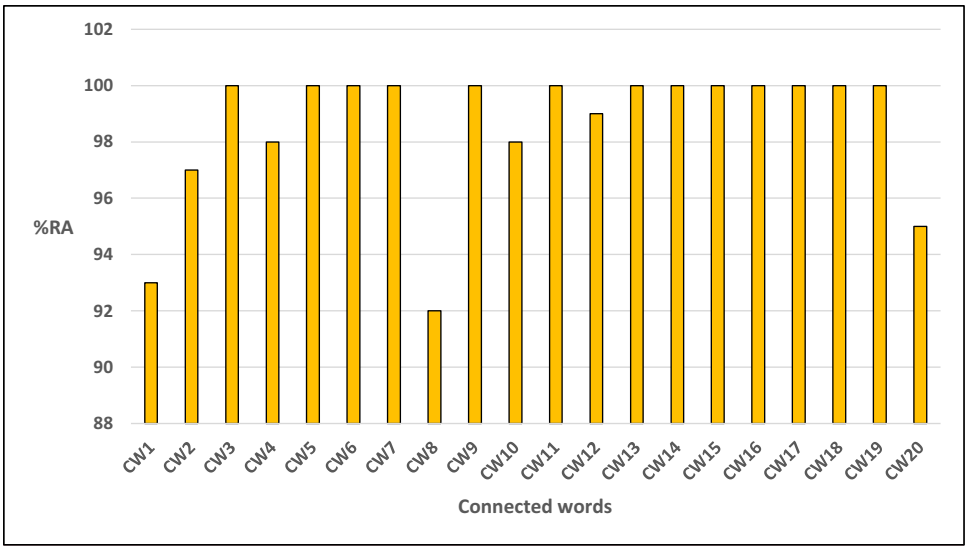

Fig. 19 Results—decision-level fusion—connected word recognition

Table 5 Comparative analysis—proposed work with related works in literature

Reference	Database used	Features and models used	Highest performance in %
[2]	Torgo database	Spectrogram, CNN, Resnet	RA-Resnet—98.9%
[4]	Dutch database, Flemish database	Bottleneck features and TFCNN	WER—10%
[5]	Created dysarthric dataset	MFCC, CLSTM-RNN	PER—30.6%
[6]	Torgo database	MFCC, CRNN	RA—40.6%
[7]	ATR Japanese database	MFCC, pBLSTM	CER-Top-1—23.2%
[8]	Taiwan Mandarin database	MFCC, gated CNN	RA—87.75%
Proposed method	Torgo database	Spectrogram, Melspectrogram, Gammatonegram, CNN, decision-level fusion classifier	Without PSC—99.72% With PSC—99.92%

are applied to the CNN layered architecture, and combined group CNN models are created. The remaining 20% of the features are applied to the CNN group models, and classification is done based on the association of feature frames with one of the groups in training. Gammatonegram-, spectrogram-, and Melspectrogram-based CNN have an overall accuracy of 88.3%, 97.89%, and 98%, respectively. Decision-level fusion of correct classification indices of three CNN-based systems has yielded a good overall accuracy of 99.72%, with 100% individual accuracy for some isolated digits. The system is evaluated by applying the PSC speech enhancement technique to the raw dysarthric speeches, and a 9% increase in overall accuracy is ensured for Gammatonegram and CNN-based systems. The speech enhancement technique ensures a 1% increase in accuracy for spectrogram- and Melspectrogram-based recognition systems with marginal improvement for decision-level fusion of all CNN-based systems. If there is a system to recognize the distorted speeches of dysarthric persons, this automated system would be useful for caretakers to provide required help/assistance to the persons affected with dysarthria.

Acknowledgements The authors thank the Department of Science & Technology, New Delhi, for the FIST funding (SR/FST/ET-I/2018/221(C)). In addition, the authors wish to express their sincere thanks to the SASTRA Deemed University, Thanjavur, India, for extending infrastructural support to carry out this work.

Data availability All relevant data are within the paper and its supporting information files.

Declarations

Ethics approval This article does not contain any studies with human participants or animals performed by any authors.

Conflict of interest The authors declare no competing interests.

References

Albaqshi H, Sagheer A. Dysarthric speech recognition using convolutional recurrent neural networks. Int J Intell Eng Syst. 2020;13(6):384–92. https://doi.org/10.22266/ijies2020.1231.34.

Arias-Vergara T, Klumpp P, Vasquez-Correa JC, et al. Multi-channel spectrograms for speech processing applications using deep

learning methods. Pattern Anal Applic. 2021;24:423–31. https://doi.org/10.1007/s10044-020-00921-5.

Binh PH, Hoang PV, Ba DX. A high-performance speech-recognition method based on a nonlinear neural network, 2021 international conference on system science and engineering (ICSSE). 2021:96–100. https://doi.org/10.1109/ICSSE52999.2021.9537942.

Chen C-Y, Zheng W-Z, Wang S-S, Tsao Y, Li P-C, Lai Y-H. Enhancing intelligibility of dysarthric speech using gated convolutional based voice conversion system. Interspeech. 2020;2022:4686–90. https://doi.org/10.21437/Interspeech.2020-1367.

Emre Y, Vikramjit M, Sivaramand G, Franco H. Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. Comput Speech Lang. 2019;58:319–34. https://doi.org/10.1016/j.csl.2019.05.002.

Gupta S, Patil AT, Purohit M, Parmar M, Patel M, Patil HA, Capobianco Guido R. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. 2021;139:105–117. https://doi.org/10.1016/j.neunet.2021.02.008.

Joshy AA, Rajan R. Automated dysarthria severity classification using deep learning frameworks. In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE; 2021. pp. 116–20.

Kim M, Cao B, An K, Wang J. Dysarthric speech recognition using convolutional LSTM neural network. Proc Interspeech. 2018;2018:2948–52. https://doi.org/10.21437/Interspeech.2018-2250.

Kim H, Hasegawa-Johnson M, Perlman A, Gunderson J, Huang T, Watkin K, Frame S. Dysarthric speech database for universal access research. INTERSPEECH 2008, 9th annual International Speech Communication Association conference, Brisbane, Australia. (2008). https://www.isca-speech.org/archive/archive_papers/interspeech_2008/i08_1741.pdf.

Sangwan P, Deshwal D, Kumar D, Bhardwaj S. Isolated word language identification system with hybrid features from a deep belief network. Int J Commun Syst. 2020;e4418.

Sidi Yakoub M, Selouani S, Zaidi BF, et al. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. J Audio Speech Music Proc. 2020;2020(1). https://doi.org/10.1186/s13636-019-0169-5.

Soliman A, Mohamed S, Abdelrahman IA. Isolated word speech recognition using convolutional neural network, 2020 international conference on computer, control, electrical, and electronics engineering (ICCCEEE). 2021:1–6. https://doi.org/10.1109/ICCCEEE49695.2021.9429684.

Stark AP, Wójcicki KK, Lyons JG, Paliwal KK. Noise driven short-time phase spectrum compensation procedure for speech enhancement. In: Ninth annual conference of the international speech communication association. 2008.

Takashima Y, Takiguchi T, Ariki Y. End-to-end dysarthric speech recognition using multiple databases, ICASSP 2019 - 2019 IEEE international conference on acoustics, speech, and signal processing (ICASSP). 2019:6395–6399. https://doi.org/10.1109/ICASSP.2019.8683803.

Vavrek L, Hires M, Kumar D, Drotár P. Deep convolutional neural network for detection of pathological speech. In 2021 IEEE 19th world symposium on applied machine intelligence and informatics (SAMI) (pp. 000245–000250). IEEE. 2021.

Zhang J, Xiao S, Zhang H, Jiang L. Isolated word recognition with audio derivation and CNN, 2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI). 2017:336–341. https://doi.org/10.1109/ICTAI.2017.00060.