IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 1, 2021, accepted November 16, 2021, date of publication November 22, 2021,
date of current version December 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129847

Investigation of DNN-HMM and Lattice Free Maximum Mutual Information Approaches for Impaired Speech Recognition

VISHNIKA VENI S, (Member, IEEE), AND CHANDRAKALA S, (Member, IEEE)
Intelligent Systems Group, SASTRA Deemed to be University, Thanjavur, Tamil Nadu 613402, India

Corresponding author: Chandrakala S (chandrakala@cse.sastra.edu)

This work was supported by the Cognitive Science Research Initiative (CSRI), Department of Science and Technology, Government of India, under Project DST/CSRI/2017/131(G).

ABSTRACT Assistive tools that recognize impaired speech due to neurological disorders are emerging and its a fairly complex task. An Intelligent Impaired Speech Recognition system helps persons with speech impairment to improve their interactions with outside world. Impaired speakers have difficulty in pronouncing words which results in partial or incomplete speech contents. Existing Automatic Speech Recognition systems are not effective for Impaired Speech Recognition due to the speaker specific variations which depend on the severity of the neurological disorders. In this work, we have investigated two important approaches namely, Deep Neural Network-Hidden Markov Model and Lattice Free Maximum Mutual Information approach for effective recognition of impaired speech in Tamil language. The training and testing samples are collected from persons with different neurological disorders at varied intelligibility levels such as high, medium, low and very low. The recognition accuracy is evaluated and compared using two datasets namely 20 acoustically similar words and 50 words Impaired Speech Corpus in Tamil.

INDEX TERMS Assistive technology, DNN-HMM, impaired speech recognition, lattice free maximum mutual information, neurological disorders.

I. INTRODUCTION

Developing an assistive system for speech impairment due to neurological disorders is one of the complex pattern recognition tasks. According to the Global Burden of Diseases (GBD) Injuries and Risk Factors report [1], the neurological disorders are considered as the global cause for different types of disabilities around the world. The speech production system is mainly affected by various neurological diseases such as stroke, brain injury, tumors, Parkinson's disease and multiple sclerosis. Dysarthria [2] is a motor speech disorder in which the muscles involved in speech production are damaged or weakened. Cerebral palsy is a kind of disability which affects the speech articulation and the affected people find difficult to speak, write and move without any assistance. The impaired speech is characterized by mispronunciation, low precision, poor articulation, omissions, distortions, and substitutions of phonemes and consonants, slow speaking rate, hypernasality, hoarseness, mono loudness, mono-pitch, slurry speech, distorted vowels, and consonants that degrade the intelligibility of speech [4], [5]. People with speech impairment feel depressed and isolate themselves from the outside world. The impaired speakers usually communicate with the help of keyboard or other input devices. To improve their quality of life, there is a high demand to develop a robust Assistive Speech system that can recognize impaired speech.

Every impaired speaker produce their own phonetic patterns which are incomplete leading to lot of variations in speech utterances. Hence, existing Automatic Speech Recognition (ASR) techniques applied to Impaired Speech Recognition provides poor performance. The ASR systems are ineffective in mapping the impaired speech signals to phonemes correctly. Impaired speech recognition (ISR) converts impaired speech to text [6]–[8]. This text is then synthesized to normal speech in a speech assistive system. Another important challenge is the availability of limited amount of training data. Collecting huge amount of impaired speech

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

samples from neurological disordered person is quite challenging task as it makes the impaired speakers to feel stressed. Handling insufficient dysarthric speech data and issues in pronunciation modeling for impaired speech are addressed in [9]–[13].

Recently, deep model based approaches outperform traditional machine learning approaches for Automatic Speech Recognition. DNN-HMMs are proved to be effective for Automatic Speech Recognition [16], [17] and [18]. DNN-HMM combines the sequential modeling ability of HMM and the representational ability of the deep neural network. Output units of DNN are trained to determine the posterior probabilities of HMM. Though the DNN-HMM is advantageous over the traditional Gaussian Mixture Model- Hidden Markov Model (GMM-HMM), DNN-HMM gives moderate performance only for Impaired Speech Recognition. A bidirectional Deep Recurrent Neural Network (biRNN) based DNN-HMM is used for phoneme recognition [15]. In a recent work [19], the authors used a phonetic posterior feature space for matching and verifying the impaired speech with the control speakers data. Several parameters such as Linear Discriminant Analysis (LDA), context dependent states, Feature space Maximum Likelihood Linear Regression (FMLLR) are used with Teacher-Student network [20] to increase the accuracy.

In [22], DNN pretraining with sequence discriminative training is performed and experimented using a 300-hour switchboard telephone conversation data. The different sets of features such as the FMLLR, 40-LDA, LDA + Semi-Tied Covariance (STC)+ Feature-space Maximum Likelihood Linear Transformation (FMLLT), and single STC over LDA features obtained with various transformations are studied. Another improvement over the DNN-HMM is aligning transcripts using a two-step alignment process [23]. The preprocessed input is aligned in first step. The next step performs insertions, deletions, and substitutions to identify the correct word with the help of National Institute of Standards and Technology (NIST) sclite utility. Reduction in WER is achieved using sequential discriminative training with regularization techniques [24]. In another work [25], the phone posterior along with the regularization techniques such as L2 regularization is used to differentiate among the dysarthric severity levels. It mainly handles the mismatch between the normal and the dysarthric speech.

To address lack of sufficient training data, augmentation [26] is performed by perturbing the data with respect to time and tempo which resembles the dysarthric data. Then DNN-HMM is trained on the synthesised dysarthric speech. In [27], authors proposed a two-step adaptation. The first step is adapting an ASR model to multiple Dysarthric speakers and then further adapted to target dysarthric speaker. The authors used the Connectionist Temporal Classification (CTC) [28] based recognition system and proposed a voice conversion system to synthesize the new set of speech samples from the existing set of samples.

In recent literatures, DNN-HMM is proved to be effective in complex acoustic modelling, discriminative feature extraction, pronunciation error correction and knowledge transfer between normal speech and impaired speech. In this paper, we focus on investigating DNN-HMM approach and a Lattice Free Maximum Mutual Information (LF-MMI) approach for Impaired Speech Recognition. Section II deals with DNN-HMM based ISR. Section III presents the Lattice Free-Maximum Mutual Information approach. Experimental studies and performance analysis are discussed in Section IV.

II. DEEP NEURAL NETWORK-HIDDEN MARKOV MODEL (DNN-HMM) BASED IMPAIRED SPEECH RECOGNITION

In a generative model based HMM approach, the observation sequence is generated by a sequence of state transitions where each state is modeled using a GMM. DNN is capable of learning any arbitrary distribution. In DNN-HMM, the temporal characteristics of impaired speech utterances are modeled using HMM and the observational probabilities are estimated using DNN and hence DNN-HMM is termed as a hybrid model. A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its input and output layers as shown in Figure 1. At each hidden layer, a hidden unit typically maps the weighted sum of its inputs from the layer below to a deterministic value using a nonlinear activation function and passes it to the layer above.

A single DNN is used to model posterior probabilities of all states. But in case of GMM-HMM, a separate GMM is used to model each state. Deep Neural Network (DNN) is used to estimate the posterior probabilities of the context dependent tied triphone HMM states. The DNN outputs the posterior probabilities that are scaled using the class wise prior probabilities. The likelihood probability of triphone feature vector are estimated using the posterior probability given by DNN and the prior probability of states given by HMM. The cross entropy criterion is used during the DNN training. Usually, each impaired speech utterance is divided into 9 to 13 frames and the features extracted from these frames are fed as input to DNN. For recognition, the sum of log-likelihood probabilities of triphone feature vectors of impaired speech utterance is used.

Given a feature vector $\mathbf{x}$, the output of the DNN specified by the model parameters $\{\mathbf{W}, \mathbf{b}\} = \{\mathbf{W}^\ell, \mathbf{b}^\ell\}, 0 < \ell \leq N$ can be calculated by computing the activation vectors from layer 1 to layer $N-1$ [31]. The model parameters $\mathbf{W}, \mathbf{b}$ can be learned with the back propagation algorithm. The model parameters can be improved based on the first-order gradient information as

$$\mathbf{W}_{t+1}^\ell \leftarrow \mathbf{W}_t^\ell - \epsilon \Delta \mathbf{W}_t^\ell \tag{1}$$

and

$$\mathbf{b}_{t+1}^\ell \leftarrow \mathbf{b}_t^\ell - \epsilon \Delta \mathbf{b}_t^\ell \tag{2}$$

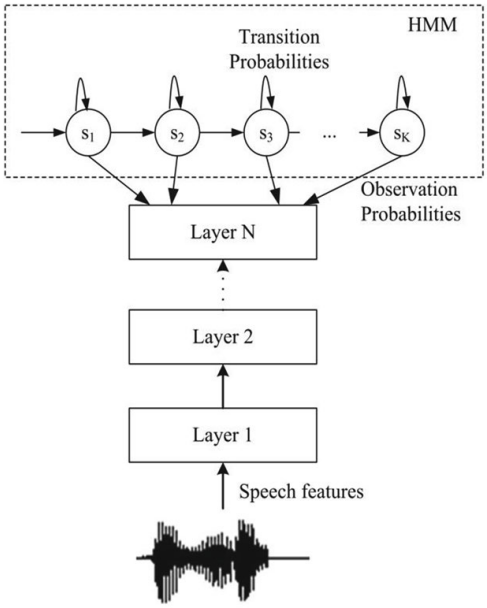


FIGURE 1. The hybrid DNN-HMM architecture.

where $\mathbf{W}_t^\ell$ and $\mathbf{b}_t^\ell$ are the weight matrix and the bias vector at layer ℓ after the t^{th} update.

$$\Delta \mathbf{W}_t^\ell = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{\mathbf{W}_t^\ell} J(\mathbf{W}, \mathbf{b}; \mathbf{x}^m, \mathbf{y}^m) \tag{3}$$

and

$$\Delta \mathbf{b}_t^\ell = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{\mathbf{b}_t^\ell} J(\mathbf{W}, \mathbf{b}; \mathbf{x}^m, \mathbf{y}^m) \tag{4}$$

are the average weight matrix gradient and the average bias vector gradient at iteration t estimated from the training batch of M_b samples, ϵ is the learning rate parameter, x is the feature vector and the corresponding output vector y is the probability distribution.

For an utterance with T frames, the state sequence is given by

$$Q \;=\; q_0\, q_1\, q_2 \cdots\; q_T \tag{5}$$

where q_0 is the initial state. The probability of such a state sequence Q can be written as

$$P(Q \mid \lambda) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \ldots a_{q_{t-1} q_t} \tag{6}$$

where $\pi(q_0)$ and $a_{q_{t-1}q_t}$ are the initial state probability and state transition probability, respectively, determined by the HMM. The embedded Viterbi training algorithm minimizes the average cross- entropy, which is equivalent to the negative log likelihood

$$J_{NLL}(\mathbf{W}, \mathbf{b}; \mathbf{x}, \mathbf{q}) \;=\; -\sum_{t=1}^{T} \log\, p(q_t \mid \mathbf{x}_t; \mathbf{W}, \mathbf{b}) \tag{7}$$

where $\mathbf{Q}$ is the state sequence. If the new model $(\mathbf{W}', \mathbf{b}')$ improves the training criterion over the old model $(\mathbf{W}, \mathbf{b})$ we have

$$-\sum_{t=1}^{T} \log\, p(q_t \mid \mathbf{x}_t; \mathbf{W}', \mathbf{b}') < \; -\sum_{t=1}^{T} \log\, p(q_t \mid \mathbf{x}_t; \mathbf{W}, \mathbf{b}) \tag{8}$$

The score of the aligned utterance

$$\begin{aligned}
&\log p(\mathbf{x} \mid w; \mathbf{W}', \mathbf{b}') \\
&\quad = \log\, \pi(q_0) + \sum_{t=1}^{T} \log\, (a_{q_{t-1}q_t}) \\
&\qquad + \sum_{t=1}^{T} [\log\, p(q_t \mid \mathbf{x}_t; \mathbf{W}', \mathbf{b}') - \log\, p(q_t)] \\
&\quad > \log\, \pi(q_0) + \sum_{t=1}^{T} \log\, (a_{q_{t-1}q_t}) \\
&\qquad + \sum_{t=1}^{T} [\log\, p(q_t \mid \mathbf{x}_t; \mathbf{W}', \mathbf{b}') - \log\, p(q_t)] \\
&\quad = \log\, p(\mathbf{x} \mid w; \mathbf{W}, \mathbf{b})
\end{aligned} \tag{9}$$

The new model improves the likelihood score of the utterance given the correct word sequence.

During the decoding process, we convert the posterior probability to the likelihood

$$p(\mathbf{x}_t \mid q_t = s) = p(q_t = s \mid \mathbf{x}_t)\;\; p(\mathbf{x}_t) \mid p(s) \tag{10}$$

where $p(s) = \frac{T_s}{T}$ is the prior probability of a state estimated from the training samples, T_s is the number of frames labeled as state s, and T is the total number of frames, $p(\mathbf{x}_t)$ is independent of the word sequence and hence can be ignored.

The decoded word sequence $\hat{w}$ is

$$\hat{w} = \text{argmax}_w \;\; p(\mathbf{x} \mid w)\;\; p(w) \tag{11}$$

where $p(w)$ is the probability given by the language model, and

$$p(\mathbf{x} \mid w) = \sum_q p(\mathbf{x} \mid q, w)\;\; p(q \mid w) \tag{12}$$

$$\approx \max\;\; \pi(q_0) \prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{t=0}^{T} p(q_t \mid \mathbf{x}_t)/p(q_t) \tag{13}$$

is the acoustic model probability, where $p(q_t \mid \mathbf{x}_t)$ is computed from the DNN. The final decoding path is determined by

$$\hat{w} = \text{argmax}_w [\log\, p(\mathbf{x} \mid w) + \lambda\, \log p(w)] \tag{14}$$

where λ is the weight of the language model.

DNNs are powerful in modeling any arbitrary mapping between inputs and outputs. However, it is difficult to train a DNN with many hidden layers. After initializing the DNN weights, supervised fine-tuning is conducted using

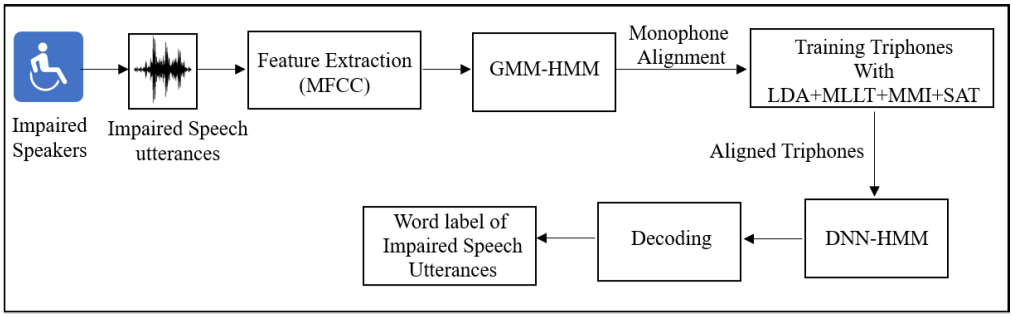


FIGURE 2. Lattice free-maximum mutual information based impaired speech recognition.

back-propagation to adjust the weights which leads to overfitting. To avoid overfitting, weight decays and dropout regularizations are used. Weight decay is applied when the training set size is small compared to the number of parameters in the DNN. Dropout is used to randomly omit a certain percentage of the neurons in each hidden layer for each presentation of the samples during training. During the training each random combination of the remaining hidden neurons need to perform well in the absence of the omitted neurons.

III. LATTICE FREE-MAXIMUM MUTUAL INFORMATION (LF-MMI) APPROACH

Maximum mutual information (MMI) is used to achieve discriminative training of sequences and to maximize the probability of the reference phonetic transcription of a word sequence while minimizing probability of other transcriptions. In MMI training, the HMMs of all the words classes are considered simultaneously. The parameters of the correct word model are updated to maximize its contribution while the parameters of the other word models are updated to minimize its contribution. The training thus provides high discriminative ability leading to improved performance.

Hence, we explore Lattice Free-Maximum Mutual Information(LF-MMI) approach for impaired speech recognition where there is a need for better discrimination among incomplete utterances of different word classes. The diagrammatic representation of Lattice Free-Maximum Mutual Information based Impaired Speech Recognition is shown in Figure 2. In LF-MMI, the output of Deep Neural Network (DNN) corresponds to tied biphone or triphone HMM states, where the state tying is done using a context-dependency tree. Biphone is used to represent a monophone with left or right context dependent monophones. This context-dependency tree is constructed using the GMM-HMM alignments.

The objective function of Maximum Likelihood (ML) estimation [29] is given as

$$f_{ML} = \sum_{u=1}^{U} log\ p_\lambda(x^{(u)}|\mathbb{M}_w^{(u)}) \tag{15}$$

where $x^{(u)}$ is the u^{th} speech utterance with transcription $w^{(u)}$, U is the total number of training utterances and λ is the set of all HMM parameters. The composite HMM graph is denoted by $\mathbb{M}_w^{(u)}$. The objective function of MMI is given as

$$f_{MMI} = \sum_{u=1}^{U} log \frac{p_\lambda(x^{(u)}|\mathbb{M}_w^{(u)})}{p_\lambda(x^{(u)})} \tag{16}$$

The denominator can be estimated as

$$p_\lambda(x^{(u)} = \sum_w p_\lambda(x^{(u)}|\mathbb{M}_w \approx p_\lambda(x^{(u)}|\mathbb{M}_{den}) \tag{17}$$

where $\mathbb{M}_{den}$ is the HMM denominator graph which includes all possible sequences of words and $\mathbb{M}_w$ is the numerator graph. The previously trained cross entropy model or GMM generates the denominator lattices. It compactly encodes a small set of likely alternative word sequences for a training utterance.

The full denominator graph with Deep Neural Network (DNN) based model is used in Lattice Free MMI (LF-MMI) approach. Its similar to lattice based MMI except that LF-MMI uses a numerator graph which makes use of alignment information and a common denominator graph instead of utterance based lattices. The LF-MMI numerator graph is a special acyclic graph that makes use of the GMM-HMM alignments as the time constraints on the phones. It is a finite state acceptor(FSA) where each phone can occur at some number of frames earlier or later than its actual occurrence in the corresponding alignment.

The two forward-backward passes are used to calculate the derivatives of the LF-MMI objective function (i.e) one on the denominator graph and the other on the numerator graph. To make the efficient forward backward pass of the denominator graph, all the utterances are split into a fixed 1.5 second chunks based on the alignment information and training is carried out on these mini batches. The pruned phone level language model trained on the previous GMM-HMM model alignments. In this work, we have used LF-MMI training in the DNN-HMM model with full denominator graph. The LF-MMI based discriminative training is expected to provide better performance than the DNN-HMM approach.

TABLE 1. Vocabulary of tamil 20 words (classes).

S.No	Word	Phonetic Transcription
1	Illa	IH L AH
2	Inga	IY NG G AH
3	Kaayam	K AA AY AH M
4	Kashtam	K AE SH T AH M
5	Marundhu	M AH R UH N D HH UW
6	Mayakkam	M EY AE K AH HM
7	Moodu	M UW D UW
8	Munnadi	M UH N AH D IY
9	Neram	N EH R AH M
10	Nethu	N EH TH UW
11	Paal	P AH L
12	Paapa	P AA P AH
13	Saapadu	S AA P AH D UW
14	Saapdu	S AA P D UW
15	Vaandhi	V AA N D HH IY
16	Vaanga	V AA N G AH
17	Valikudhu	V AE L IH K UH D HH UW
18	Vanakkam	V AH N AE K AH M
19	Vendaam	V EH N D AH M
20	Venum	V EH N UH M

IV. EXPERIMENTAL STUDIES

A. DATASETS

The **Impaired speech corpus in tamil** is formed from 18 impaired speakers of both male and female with various neurological disorders like cerebral palsy, multiple sclerosis, mental retardation, brain and spinal cord injury, muscular dystrophy and stroke. The speakers of all intelligibility levels "High", "Medium", "Low" and "Very Low" are involved in Impaired speech sample collection. The speech samples are recorded using lavalier collar microphone in a laboratory environment. Each Impaired speaker has uttered 50 unique isolated words and repeated those 50 words for 5 times in different sessions. Thus, each speaker has uttered 50*5 = 250 unique examples. The corpus also contains the speech data collected from 6 healthy speakers. We have used two dysarthric speech datasets namely 50 words impaired speech corpus and the other 20 words dataset formed from 50 words impaired speech corpus by picking utterances of word classes that are acoustically similar. First dataset contains 6000 utterances and second dataset contains 2400 utterances. The selected 20 words are listed in the Table 1. For both datasets, 75% utterances were used for training and 25% used for testing. All the training and the testing utterances are selected at equal proportion from all the four intelligibility levels. MFCC features were used as basic features, which are extracted using 25 ms frame size, 50% frameshift with Hamming window. HMM, DNN-HMM and LF-MMI experiments are done in Kaldi [30] toolkit.

B. HMM FOR IMPAIRED SPEECH RECOGNITION

To model impaired speech utterances, due to co-articulation effects, context-dependent triphone units are used as the basic units. In conventional HMM training, the number of states and mixtures are fixed based on the lexicons, silence, phoneme related files and text. The text file contains the utterance-ids and the corresponding word. With the help of these files, the HMM topology is fixed and different alignments are performed by considering the phoneme as a basic unit. Triphones significantly increase the number of parameters to be estimated. The performance of the HMM evaluated using monophones and four different triphone models tri1a, tri2a, tri3a and tri4a for 20 acoustically similar words and 50 words impaired speech corpus in tamil datasets are shown in Table 2. HMM gives poor performance due to challenges in impaired speech such as missing vowels and consonants and overlaps in acoustically similar word classes.

TABLE 2. Performance (%) of the HMM approach for impaired speech corpus in tamil.

Features	20-acoustically similar words Impaired Speech Corpus in tamil Accuracy (%)	50 words Impaired Speech Corpus in tamil Accuracy (%)
Mono	57.50	35.40
Tri1	64.50	45.33
Tri2a	**65.17**	**45.93**
Tri3a	64.00	43.93
Tri4a	49.67	30.40

TABLE 3. Performance (%) of the DNN-HMM approach for impaired speech corpus in tamil.

Features	20-acoustically similar words Impaired Speech Corpus in tamil Accuracy (%)	50 words Impaired Speech Corpus in tamil Accuracy (%)
Mono	58.80	36.73
Tri1	**65.83**	**51.87**
Tri2a	63.83	50.27
Tri3a	64.33	51.40
Tri4a	65.00	51.30

C. DNN-HMM APPROACH

In DNN, various parameters like number of hidden layers, number of neurons in each hidden layer and batch size are fixed based on the HMM aligned data, Weighted Finite state Transducer (WFST) and lexicon file. The maximum number of states that can be active at one time is controlled by the max-active parameter is fixed during the decoding process. The different triphone alignments are studied to achieve better word recognition accuracy even with overlapped and missing phonemes. The DNN-HMM shows slight improvement in performance than that of HMM by 1.01% for 20 acoustically similar words dataset and 11.2% for 50 words impaired speech corpus in tamil dataset. Slight improvement for Impaired speech recognition is due to the limited amount of training data when compared to large datasets available for Automatic Speech Recognition (ASR) task. The performance of DNN-HMM of two datasets are shown in Table 3.

D. CONVOLUTIONAL NEURAL NETWORK APPROACH

Convolutional Neural Network (CNN) is used to learn high level features from Spectrograms. Spectrograms are

generated by applying FFT over the preprocessed impaired speech signal with the help of hamming window. Then the Mel filter bank is applied for converting the spectrum to the Mel spectrum. The dimension of the generated spectrogram is 1368×864 pixels. The generated spectrograms are discriminative even for acoustically similar word classes. These spectrograms are fed as input to CNN to output the word label of impaired speech samples. The architecture of CNN is as follows: The network is composed of four sets of convolutional layers and max pooling layers with pool of size 2×2. Initially, the filter of size 16 is used for convolution operation and gradually increased to 128. Batch normalization and Dropout regularization are applied to avoid overfitting. The rate of dropout is set to 50% in all the layers. The categorical cross entropy and adam optimizer is used to optimize 42,879,892 trainable parameters. The Tensorflow and keras package are used to implement the CNN architecture. The performance of CNN with two datasets is shown in Table 6.

E. LATTICE FREE MMI APPROACH

The steps followed in LF-MMI experiment is explained as follows. In this work, MFCC features are extracted and computationally efficient Cepstral Mean and Variance Normalization (CMVN) is applied to the extracted MFCC features. These are used as basic features to train a monophone acoustic model. The triphone model tri1a was trained by fixing the number of Gaussians to 9000. The tri2a model was trained using delta and delta-delta features. Once the tri2a model is trained, Linear Discriminant Analysis (LDA) and Maximum Linear Likelihood Transformation (MLLT) is applied on the features and maintained the same number of Gaussians to form Triphone tri2b model. Then we applied MMI on the top of tri2b model to form tri2b MMI model and a boosting of 0.05 was applied on the tri2b model to check the effect of boosting. The Maximum phone error (MPE) is applied on the top of LDA + MLLT. Additionally, the Speaker Adaptive Training (SAT) was used along with LDA + MLLT to form tri3b model. Finally, the tri3b MMI training is done with the LDA + MLLT + SAT + MMI feature transforms.

LF-MMI gives slightly better performance than that of DNN-HMM for impaired speech recognition as shown in table 6. LF-MMI approach shows improvement by 3.38%, 2.33% and 22.93% than that of HMM, DNN-HMM and CNN in 20 acoustically similar words impaired speech corpus in tamil respectively. In case of 50 words impaired speech corpus in tamil dataset, the LF-MMI approach shows better improvement by 25.5%, 11.18% and 35.15% than that of HMM, DNN-HMM and CNN respectively.

F. FIXED DIMENSIONAL REPRESENTATION USING CEPSTRAL FEATURES

The raw impaired speech signal is fed as input to extract Mel Frequency Cepstral Coefficients(MFCC). MFCC is a dominant feature extraction technique which extracts the speaker specific parameters from the impaired speech. The steps involved in MFCC feature extraction are as follows: preprocessing, framing and windowing, Fast Fourier Transform (FFT), processing using Mel Filter bank and Discrete Cosine Transform (DCT). The long windows are used to obtain better frequency resolution and short windows are used for better time resolution. Support Vector Machine is a discriminative classifier proved effective for complex recognition tasks even with small amount of training data. SVM accepts a fixed dimensional feature vector and so the number of windows are fixed by varying the window size for each impaired speech utterance. 39 MFCC features extracted from every overlapping window are concatenated to form fixed dimensional feature vector. The number of windows are varied from 100, 200, 400 and 600 to form different dimensions of the MFCC feature vectors. The MFCC feature vector dimensions for 100, 200, 400 and 600 windows are 100*39, 200*39, 400*39 and 600*39 respectively. The performance of this approach is shown in the Table 4.

TABLE 4. Performance (%) of fixed dimensional representation using MFCC features for impaired speech corpus in tamil.

Number of Windows	20-acoustically similar words Impaired Speech Corpus in tamil Accuracy (%)	50 words Impaired Speech Corpus in tamil Accuracy (%)
100	**42.66**	**29.80**
200	39.00	27.76
400	36.60	25.23
600	34.00	24.10

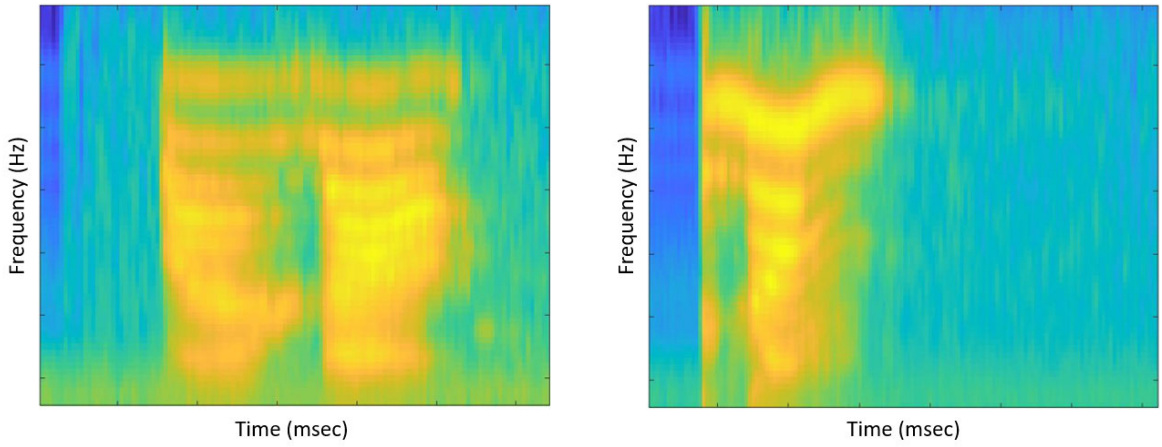
TABLE 5. Performance (%) of Gammantonegram representation for impaired speech corpus in tamil.

Features	20-acoustically similar words Impaired Speech Corpus in tamil Accuracy (%)	50 words Impaired Speech Corpus in tamil Accuracy (%)
BLOB	41.83	24.67
SIFT	**42.16**	**25.00**

G. VISUAL REPRESENTATION USING GAMMATONEGRAM

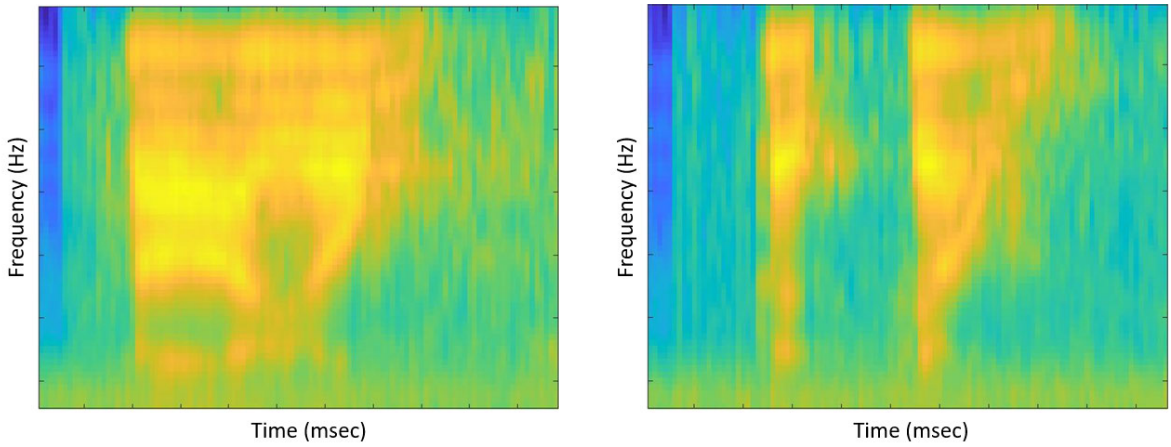
We explored another fixed dimensional representation using Gammatonegrams which perform better than the spectrograms [35]. Gammatonegram is a visual time vs frequency representation of energy of speech signal obtained using Short Time Fourier Transform (STFT) and Gammatone filterbank. Gammatonegram, the visual representation on gammatone filterbank. The gammatonegram generation is quite simple and requires only matrix multiplication and Discrete Fourier Transform (DFT). It is more robust than traditional spectrogram, since the gammatone bandpass filter's magnitude gain is proportional to the bins of the DFT.

The key difference between the spectrogram and the gammatonegram depends on the bandwidth. In spectrogram, the input speech signal is processed by bandpass filter with same bandwidth. But in case of gammatonegram representation, the bandwidth of the bandpass filter changes with the central



(a) Vendaam (b) Venum

FIGURE 3. Sample gammatonegram's of acoustically similar words "Vendaam" and "Venum" belonging to "Very Low" intelligibility.



(a) Kaayam (b) Kashtam

FIGURE 4. Sample gammatonegram's of acoustically similar words "Kaayam" and "Kashtam" belonging to "Very Low" intelligibility.

frequency. It implies that the difference in frequency is not observed strongly in high frequency region than at low frequency. The input speech signal is divided into n number of frames and the gammatonegram representation $y(t, f_c)$ is formed by concatenating the output response of the frame $x(t)$ with the gammatone filterbank $g(t, fc)$ [35].

$$g(t, f_c) = at^{n-1} e^{-2\pi bt} cos(2\pi f_c t + \phi) \quad (18)$$

$$y(t, f_c) = x(t) * g(t, f_c) \quad (19)$$

where each column of gammatonegram is the filterbank response at time t, central frequency $fc(inHz)$, a is the amplitude which is kept constant that controls the gain and n denotes the order of the filter. The bandwidth of the filter is determined by the impulse response duration and decay factor b.

The gammatonegram's generated for acoustically similar words ''Vendaam'', ''Venum'' and ''Kaayam'', ''Kashtam'' uttered by the ''Very Low'' intelligibility speaker is depicted in Figure 3 and 4. It shows how the interest points are spread over the gammatonegram image and proves the representational power of gammatonegram. But in case of spectrogram, the interest points are localized in low frequency. The

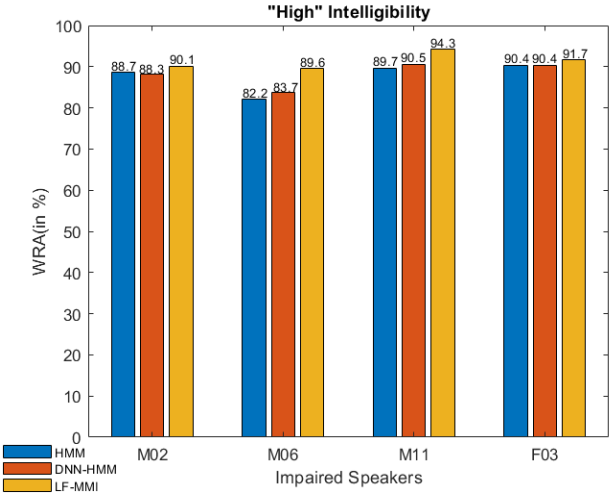


FIGURE 5. Speaker wise WRA (in %) for "HIGH" intelligibility.

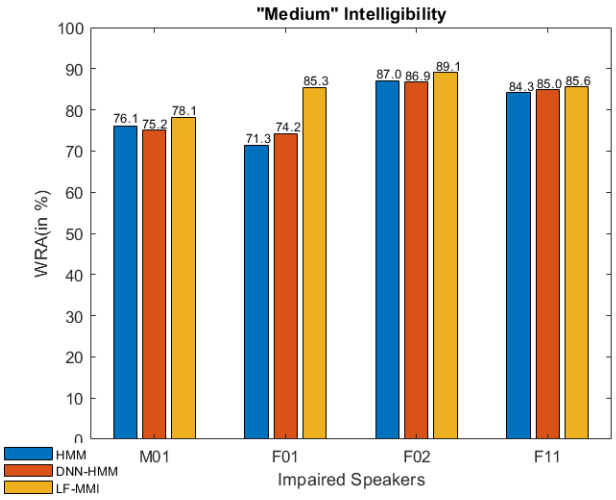


FIGURE 6. Speaker wise WRA (in %) for "MEDIUM" intelligibility.

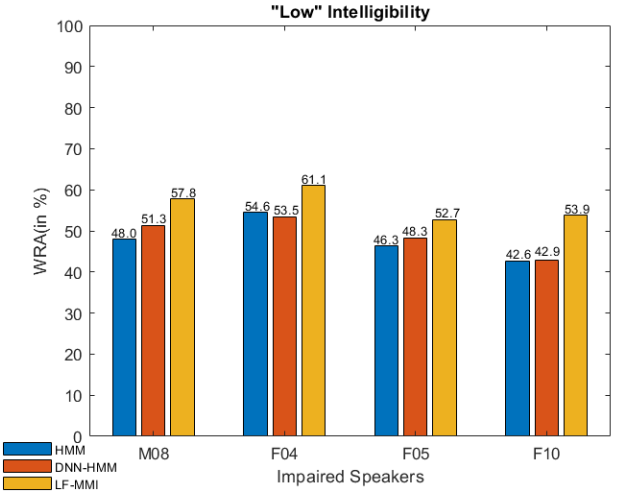


FIGURE 7. Speaker wise WRA (in %) for "LOW" intelligibility.

generated gammatonegrams show better discrimination even for two acoustically similar words. The dimension of the generated gammatonegram is 875 × 656 and to reduce the computational complexity it is resized to 100 × 100 pixels.

The robust features like Scale Invariant Fourier Transform (SIFT) and Binary Large Object (BLOB) are extracted from the gammatonegrams. The dimension of BLOB and SIFT feature vector is 30000. The performance of this representation using Auditory Image features for 20 acoustically similar words and 50 words impaired speech corpus in tamil dataset is shown in the Table 5.

H. MULTIVIEW REPRESENTATION USING CEPSTRAL FEATURES AND GAMMATONE IMAGE FEATURES

The cepstral features are combined with the auditory image features to form the multi view representation. The feature dimensions of the Multiview representation is 30000 (blob feature dimensions) + 3,900 (MFCC features with

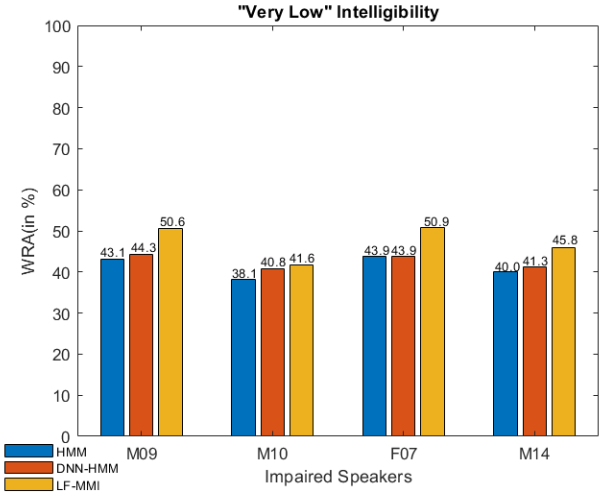


FIGURE 8. Speaker wise WRA (in %) for "VERY LOW" intelligibility.

100 windows). These combined features are fed as input to the discriminative classifier SVM and an improved performance is obtained when compared to fixed dimensional MFCC representation and Gammatonegram representation. The word recognition accuracy of the multi-view representation is shown in Table 6.

I. PERFORMANCE ANALYSIS

1) OVERALL COMPARISION

We have compared the LF-MMI approach with the conventional HMM, DNN-HMM, Fixed dimensional MFCC representation, gammatonegram representation, multiview representations and CNN respectively. The word recognition accuracy (WRA) is calculated for all the experiments to evaluate the performance.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ words}{Total\ number\ of\ words\ per\ class} * 100 \tag{20}$$

The overall performances of HMM, DNN-HMM, Fixed dimensional MFCC representation, Gammatonegram representation, Multiview representations, CNN and LF-MMI approach are given in Table 6. Though the performance of DNN-HMM is better than HMM, Fixed dimensional MFCC representation, Gammatonegram representation, Multiview representations and CNN, the LF-MMI approach attains a better recognition accuracy even in the presence of high overlapping word classes like "Kaayam" and "Kashtam", "Paal" and "Paapa", "Saapadu" and "Saapdu", "Venum" and "Vendaam".

2) PERFORMANCE ANALYSIS WITH VARIED INTELLIGIBILITY LEVELS

The performance of LF-MMI approach for different speakers belonging to different intelligibility levels of Impaired speech corpus in Tamil is evaluated. The words uttered by

TABLE 6. Comparision of performance (%) of the HMM, DNN-HMM, Fixed dimensional MFCC, Gammatonegram, Multiview, CNN and LF-MMI approach for impaired speech corpus in tamil.

Methods	20-acoustically similar words Impaired Speech Corpus in tamil Accuracy (%)	50 words Impaired Speech Corpus in tamil Accuracy (%)
HMM [32]	65.17	45.93
DNN-HMM [33]	65.83	51.87
Fixed Dimensional (MFCC)	42.66	29.80
Gammatonegram Representation (SIFT)	42.16	25.00
Multi View (MFCC + SIFT) [37]	42.86	30.31
CNN (Spectrogram) [36]	54.80	42.67
LF-MMI [34] (LDA+MLLT+MMI+SAT)	**67.37**	**57.67**

speakers belonging to ''Very Low'' and ''low'' intelligibility levels are correctly recognized by LF-MMI approach than other conventional approaches. Even with limited amount of training dataset, high overlapping word classes and missing phonemes, the LF-MMI approach provides better performance than HMM based approach and improved the performance by 2.56%, 3.68%, 20.54% and 7.35% for ''High'', ''Medium'', ''Low'' and ''Very Low'' intelligibility levels respectively. The Figures 5, 6, 7 and 8 shows speaker wise word recognition accuracy of varied intelligibility levels ''High'', ''Medium'', ''Low'' and ''Very Low'' respectively. For the purpose of demonstration, four out of six impaired speakers of each intelligibility levels are shown.

V. CONCLUSION

We have investigated the performance of DNN-HMM approach and LF-MMI approach for Impaired Speech Recognition in Tamil language. LF-MMI approach provides an improved discrimination among impaired speech utterances of acoustically similar word classes with missing vowels and consonants. The performance of the LF-MMI approach was evaluated using 20 acoustically similar words and 50 words dataset of Impaired speech corpus in Tamil. The LF-MMI approach shows better performance than the conventional HMM, DNN-HMM, CNN and MVR representation. Though DNN-HMM and LF MMI approaches are promising for healthy speech recognition systems, studies show that there is still a need for robust methodologies to improve the performance of impaired speech recognition task.

REFERENCES

[1] W. M. Carroll, ''The global burden of neurological disorders,'' *Lancet Neurol.*, vol. 18, no. 5, pp. 418–419, May 2019.
[2] F. L. Darley, A. E. Aronson, and J. R. Brown, ''Differential diagnostic patterns of dysarthria,'' *J. Speech Hearing Res.*, vol. 12, no. 2, pp. 246–269, 1969.
[3] J. R. Duffy, *Motor Speech Disorders E-Book: Substrates, Differential Diagnosis, and Management*. Amsterdam, The Netherlands: Elsevier, 2019.
[4] R. D. Kent, H. K. Vorperian, J. F. Kent, and J. R. Duffy, ''Voice dysfunction in dysarthria: Application of the multi-dimensional voice program,'' *J. Commun. Disorders*, vol. 36, no. 4, pp. 281–306, Jul. 2003.
[5] S. Chandrakala and N. Rajeswari, ''Representation learning based speech assistive system for persons with dysarthria,'' *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1510–1517, Sep. 2016.
[6] F. Xiong, J. Barker, Z. Yue, and H. Christensen, ''Source domain data selection for improved transfer learning targeting dysarthric speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7424–7428.
[7] E. Hermann and M. Magimai.-Doss, ''Dysarthric speech recognition with lattice-free MMI,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6109–6113.
[8] Y. Liu, T. Lee, T. Law, and K. Y.-S. Lee, ''Acoustical assessment of voice disorder with continuous speech using ASR posterior features,'' *IEEE/ACM Trans. Audio, Speech, Langauge Process.*, vol. 27, no. 6, pp. 1047–1059, Jun. 2019.
[9] K. T. Mengistu and F. Rudzicz, ''Adapting acoustic and lexical models to dysarthric speech,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4924–4927.
[10] H. Christensen, M. B. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, ''Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech,'' in *Proc. Interspeech*, Aug. 2013, pp. 3642–3645.
[11] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, ''Severity-based adaptation with limited data for ASR to aid dysarthric speakers,'' *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86285.
[12] S. O. C. Morales and S. J. Cox, ''Modelling errors in automatic speech recognition for dysarthric speakers,'' *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–14, Dec. 2009.
[13] W. K. Seong, J. H. Park, and H. K. Kim, ''Multiple pronunciation lexical modeling based on phoneme confusion matrix for dysarthric speech recognition,'' *Adv. Sci. Technol. Lett.*, vol. 14, pp. 57–60, Jun. 2012.
[14] P. Sharma, V. Abrol, and A. K. Sao, ''Deep-sparse-representation-based features for speech recognition,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2162–2175, Nov. 2017.
[15] J. Yu, K. Markov, and T. Matsui, ''Articulatory and spectrum information fusion based on deep recurrent neural networks,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, pp. 742–752, Apr. 2019.
[16] A. Waris and R. K. Aggarwal, ''Optimization of deep neural network for automatic speech recognition,'' in *Proc. Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2018, pp. 524–527.
[17] T. Tanaka, R. Masumura, T. Moriya, and Y. Aono, ''Neural speech-to-text language models for rescoring hypotheses of DNN-HMM hybrid automatic speech recognition systems,'' in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Conf. (APSIPA ASC)*, Nov. 2018, pp. 196–200.
[18] G. E. Dahl, D. Yu, L. Deng, and A. Acero, ''Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,'' *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[19] J. Fritsch and M. Magimai-Doss, ''Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features,'' *IEEE Signal Process. Lett.*, vol. 28, pp. 224–228, 2021.
[20] N. M. Joy and S. Umesh, ''Improving acoustic models in TORGO dysarthric speech database,'' *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 637–645, Mar. 2018.
[21] Y. Takashima, T. Takiguchi, and Y. Ariki, ''End-to-end dysarthric speech recognition using multiple databases,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6395–6399.
[22] Y. Miao and F. Metze, ''Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training,'' in *Proc. Interspeech*, vol. 13, Aug. 2013, pp. 2237–2241.
[23] N. Dugan, C. Glackin, G. Chollet, and N. Cannings, ''Intelligent voice ASR system for Iberspeech 2018 speech to text transcription challenge,'' in *Proc. IberSPEECH*, Nov. 2018, pp. 272–276.
[24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, ''Purely sequence-trained neural networks for ASR based on lattice-free MMI,'' in *Proc. Interspeech*, Sep. 2016, pp. 2751–2755.
[25] T. Lee, Y. Liu, Y. T. Yeung, T. K. T. Law, and K. Y. S. Lee, ''Predicting severity of voice disorder from DNN-HMM acoustic posteriors,'' in *Proc. Interspeech*, Sep. 2016, pp. 97–101.

[26] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 471–475.
[27] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6104–6108.
[28] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6428–6432.
[29] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proc. Interspeech*, Sep. 2018, pp. 12–16.
[30] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6465–6469.
[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Cognit. Model.*, vol. 5, no. 3, p. 1, 1985.
[32] L. R. Rabiner and L. R. Juang, *Fundamentals of Speech Recognition*, vol. 14. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[33] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.
[34] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, p. 2326, Apr. 2020.
[35] A. Greco, N. Petkov, A. Saggese, and M. Vento, "AReN: A deep learning approach for sound event recognition using a brain inspired representation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3610–3624, 2020.
[36] M. S. Yakoub, S.-A. Selouani, B.-F. Zaidi, and A. Bouchair, "Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network," *EURASIP J. Audio, Speech, Music Process.*, vol. 2020, no. 1, pp. 1–7, Dec. 2020.
[37] M. Srinivasan, C. Shanmuganathan, S. M. K. Gupta, and M. Y. Sikkandar, "Multi-viewrepresentation based speech assisted system for people with-neurological disorders," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 1–12, Jan. 2021.

VISHNIKA VENI S (Member, IEEE) is currently pursuing the Ph.D. degree with SASTRA University, India. Her research interests include machine learning, speech technology, and computer vision. She is a member of IEEE Signal Processing Society and IEEE Engineering in Medicine and Biology Society.

CHANDRAKALA S (Member, IEEE) received the Ph.D. degree from the Indian Institute of Technology, Madras, India. She is currently working as a Professor with the School of Computing, SASTRA University, India. Her research interests include machine learning, data analytics, intelligent systems, speech technology, and computer vision. She is a member of the IEEE Signal Processing Society and IEEE Systems, Man, and Cybernetics Society. She is a Reviewer of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and IEEE ACCESS.