Enhancement in Speaker Identification through Feature Fusion using Advanced Dilated Convolution Neural Network

Original Scientific Paper

Hema Kumar Pentapati
Department of Electrical Electronics and Communication Engineering,
GITAM School of Technology
Visakhapatnam-530045, India
hpentapa@gitam.in | https://orcid.org/0000-0002-9373-9132

Sridevi K
Department of Electrical Electronics and Communication Engineering
GITAM School of Technology
Visakhapatnam-530045, India
skataman@gitam.edu | https://orcid.org/0000-0002-6716-6705

Abstract *– There are various challenges in identifying the speakers accurately. The Extraction of discriminative features is a vital task for accurate identification in the speaker identification task. Nowadays, speaker identification is widely investigated using deep learning. The complex and noisy speech data affects the performance of Mel Frequency Cepstral Coefficients (MFCC); hence, MFCC fails to represent the speaker characteristics accurately. In this proposed work, a novel text-independent speaker identification system is developed to enhance the performance by fusion of Log-MelSpectrum and excitation features. The excitation information is obtained due to the vibration of vocal folds, and it is represented using Linear Prediction (LP) residual. The various types of features extracted from the excitation are residual phase, sharpness, Energy of Excitation (EoE), and Strength of Excitation (SoE). The extracted features were processed with the dilated convolution neural network (dilated CNN) to fulfill the identification task. The extensive evaluation showed that the fusion of excitation features gives better results than the existing methods. The accuracy reaches 94.12% for 11 complex classes and 91.34% for 80 speakers, and Equal Error Rate (EER) is reduced to 1.16% for the proposed model. The proposed model is tested with the Librispeech corpus using Matlab 2021b tool, outperforming the existing baseline models. The proposed model achieves an accuracy improvement of 1.34% compared to the baseline system.*

Keywords*: Log-MelSpectrum, MFCC, Speaker Identification, excitation features, Convolution Neural Network(CNN), LP Residual, deep learning, Deep Neural Network*

1. INTRODUCTION

Speaker recognition is one of the most prominent bio-feature based methods. Identifying the speaker depends on the speaker's voice signal [1]. The naturally produced voice is one of the significant human biometric characteristics. The vocal tract, larynx sizes, and various organs responsible for speech generation are unique for each individual [2]. It contains rich information which provides many things about the speaker, including age, gender, and emotions [3,1]. The unique characteristic of speech enables us to focus on speaker recognition. Advanced speaker recognition systems are used in many fields, including online banking, security control, voice access, forensics, and military applications [4,2,5]. Speaker recognition may be mainly classified into two classes: speaker identification (SI) and verification. The SI identifies the unknown speaker's identity by comparing their speech characteristics with those of recognized speakers. The speaker with the highest utterance score is identified as an actual speaker [1]. Thus, it is considered a 1:N match where each utterance is compared against many sets of utterances [4,3]. The speaker verification verifies the claimed identity of the person by accepting /rejecting the speaker [5,3]. It is the 1:1 match where the claimed speaker is compared with their characteristics [3]. Speaker recognition can be divided into text-dependent and Text-independent systems. Text-dependent modules work with the same set of phrases in training and testing [6]. Whereas in text-independent systems, there are no such limitations on text phrases in training and testing.

This paper aims to focus on a text-independent speaker identification system.

Artificial intelligence (AI) approaches can be helpful to Speaker recognition since it is considered a pattern recognition approach [7]. Deep learning addresses complex recognition tasks and effectively helps implement speaker identification systems [8]. Various factors affect the performance of speaker identification. The noisy environment may lead to the misclassification of speakers. The quality of speech signal is severely degraded in forensic applications, leading to difficulty in identifying the correct speaker [9]. Thus, extracting more discriminative features for each individual is the challenge in speaker identification [10]. The researchers found various feature engineering techniques in their study on speaker identification, such as MFCC, LPCC, and several time-domain and spectral-domain features. However, these features cannot represent the unique characteristics under complex and noisy data such as LibriSpeech [11,10]. Moreover, dependence on only a single type of feature set restricts the performance and reliability of the system. In this regard, we propose the model in relation to the researcher's work in this area.

1.1. RATIONALE BEHIND USING EXCITATION FEATURES IN SPEAKER IDENTIFICATION

The MFCC projects the useful speaker's characteristics for speaker identification. However, it is necessary to understand the acoustic cues that describe the detailed speaker characteristics, including different voice qualities and emotions. Then, utilize this information to identify the speaker [12]. Hence, it is required to capture the features describing excitation and the vocal filter in speech production. Thus, excitation features provide supportive information to the frequently used vocal tract features of various speakers. The different methods are well established to describe the vocal tract filter, but the researchers showed less interest in excitation features [12]. The study of [13] demonstrated the methods to capture the excitation features effectively and mentioned the future scope to combine excitation and vocal tract features. This motivates us to combine the derived excitation features from the LP Residual method and vocal tract filter characteristics.

1.2. CONTRIBUTIONS

The significant contributions of the paper are as follows: 1) proposed an efficient feature engineering technique by combining the log-MelSpectrum (vocal tract filter feature) and the various excitation-based features (source features) to improve the identification accuracy. 2) proposed a deep learning-based model referred to as an advanced dilated convolution network that takes the combined feature set of each sample as input to reduce memory consumption and training time. The reason for choosing the dilation convolution network for model construction is it allows us to learn sparse relationships. Primarily, it helps learn speaker-dependent features, as elaborated in the following sections. 3) Evaluate the performance of the system on standard complex LibriSpeech corpus and compare it with baseline methods. 4) Evaluate the proposed model performance by comparing its accuracy with baseline models and analyze the performance with the different number of speakers.

Different sections of the paper are structured as follows. The recent trends in speaker identification research and the performance of the well-established speaker identification methods are presented in section II. Section III presents the analysis of various features and identification models. Section IV presents the dataset corpus, training, and experiments. Section V reports the results of different experiments and discusses the importance of obtained results. Section VI concludes with overall findings and improvements in the proposed model.

2. LITERATURE REVIEW

The human speech signal is a powerful medium for communication. Due to the unique characteristics of each individual, it is widely used in biometric systems such as speaker recognition and speech recognition. A speaker recognition system should handle various variations, such as environmental and background noises, speaker-based and technology-based variations. It extracts the speaker characteristics effectively from the speech signal [1] and it includes the features commonly used, such as Mel-FrequencyCepstralCoefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). The various feature extraction methods, the variations in MFCC, and fusion [7] have been adopted to address complexity and noise-related issues in speaker identification. Significant progress has been made in the first efforts to build the speaker identification system.

The non-parametric approaches called vector quantization [14] and Dynamic time warping have been popular in the research. Later on, researchers shifted towards parametric methods such as Hidden-Markov Model and Gaussian Mixture Model (GMM) for text-independent based systems. The approaches on MFCC and GMM-based text-independent speaker identification systems [15] were popular. The expectation-maximization algorithm was used for parameter estimation and Maximum likelihood(ML) to train the model, while maximum a posteriori(MAP) was used to train the model in [15]. But it requires long utterances, and the MFCC fails to classify correctly under noisy audio data. In some other approaches, the support vector machine(SVM) trained with a large amount of data and Neural Networks were widely used in SI [3].

Jahangir et al. [10] proposed a feature combination of MFCC and time base features to enhance the overall accuracy of their proposed identification system and

fed to the deep neural network (DNN). They obtained an accuracy of 89% on LibriSpeech for 100 speakers. Although they achieved a good performance, the accuracy is still affected due to similar voice patterns. Also, this is due to their simpler deep learning architecture. Chowdhury et al. [9] recently proposed a speaker recognition system in different degraded audio conditions. The study fused MFCC and Linear Predictive Coefficients (LPC) features to classify speakers using their proposed architecture with dilated convolution. The four other datasets were used to evaluate the model. The total match rate was improved by 12% on degraded TIMIT data compared to the existing method but still failed to verify 14% of samples under this dataset.

Hao Meng et al. [16] proposed dilated CNN-based novel architecture for speech emotion recognition. It was evaluated on two popular emotional databases. They picked up the log-MelSpectrum of the speech signal and obtained a notable improvement in accuracy. In [12], the authors stated that it is required to analyze the generation of acoustical cues in the process of speech production. Extracting the features of both the excitation and vocal tract system is required. Various approaches are used to extract excitation information [17,18]. The authors of [13,17] elaborated on the three excitation features based on strength, energy and derived frequency around Glottal Closure Instant (GCI) locations. The authors [17] concluded that the performance of recognition systems might be improved by combining excitation- based features with the features that describe the vocal tract systems.

Ali et al. [19] proposed an SVM-based speaker identification using an Urdu dataset to identify ten speakers. They fused deep belief network features and MFCC features. The model achieved an accuracy of about 92.6%. But only ten speakers were used for evaluation, and each utterance contained only one word. Nainan Kulkarni et al. [20] evaluated 1D CNN, SVM and GMM based on dynamic MFCC features. The 1D CNN-based model achieved a validation accuracy of about 73.25% on the VidTimit dataset. But it was evaluated for only 43 speakers. Samia Abd El-Moneim et al. [21] proposed an LSTM-RNN-based text-independent speaker recognition system to identify five female speakers using the Chinese mandarin dataset. They extracted the log spectrum from the speech signal and used it as the feature set for the model. The experimental results achieved 98.7% with undistorted data. The spectral subtraction method was used to reduce the noise from the speech signal and improve the recognition performance. Although the model improved accuracy, it was evaluated on only five speakers.

Ting Lin et al. [22] proposed a long-term acoustic features-based DNN to identify ten speakers from the LibriSpeech corpus. The authors extracted MFCC and super MFCC features as LTA features and DNN is employed as a classifier. The experimental results achieved the accuracy of 90%. V Srinivas et al. [23] proposed an efficient adaptive fraction bat-based support vector neural network for speaker recognition. They employed frequency-dependent features like spectral kurtosis. It was evaluated on the ELSDSR dataset and achieved an accuracy of 95%. It can be further extended to test on a large dataset since ELSDSR consists of only 22 speakers. Soleymanpour et al. [24] proposed a text-independent speaker identification model based on an artificial neural network and clustering of MFCC. It was evaluated on the ELSDSR dataset and achieved an accuracy of 93%.

3. PROPOSED METHODOLOGY

This part of the paper deals with the detailed methodology of speaker identification and it is shown in Fig. 2. Firstly, the male and female voices from the Librispeech corpus were collected to conduct the experiments [10]. Second, the speech signals from the database are processed, extracting the various useful features to form the fused feature set. This feature set was given as input to the dilated CNN architecture to design the speaker identification model. The developed model was evaluated with the different number of speakers and complex data and compared with existing baseline systems. The next sections will discuss in detail the preferences and parameters considered for the model.

3.1. SPEECH DATA ORGANIZATION

First, split the speech samples into several frames with a window of 25ms. The log MelSpectrum and residual phase MelSpectrum are derived with the exact window sizes. The log-MelSpectrum of each frame comprises 13 coefficients. For the fixed sampling rate of 16KHz and duration of 2s, we have 198 feature frames from each speech sample in the database. Hence, each speech sample is of size 13 x 198 for the log-Mel spectrum. The extraction of excitation features enriches the feature extraction phase. The residual phase MelSpectrum comprises 13 coefficients and two excitation features per frame.

3.2. FEATURE ENGINEERING

The feature set plays a prominent role in the system's overall performance. In deep learning, the speaker identification task depends on the relevant features extracted [10]. Hence, several valuable features are required to extract from the speech samples and these features are appropriate to accomplish the speaker identification mechanism. Therefore, extracting the discriminative features and preparing the feature set through feature engineering is essential. These features are used by the deep learning model and make learning and developing the speaker identification model easier. Thus, the proposed work adopts the extraction and fusion of novel features known as log-MelSpectrum and excitation features to build an accurate system for speaker identification. The functionality and usage of these features are discussed in the following sections.

The vocal tract information can be best depicted by log-MelSpectrum and is used to capture the individual speech characteristics. First, each speech utterance is split into overlapping segments called frames. Each frame of the same length passed through the hamming window. Thus, we obtain the overlapping frames with a step size of 10ms. The number of frames can be calculated as shown in (1).

$$\text{Number of frames N} = \frac{\text{Total size of speech sample - window size}}{\text{Step size}} \quad (1)$$

After frame blocking and windowing of speech signal, the log-Mel spectrum can be extracted by computing the discrete Fourier transform. Then apply the mel-scale filter bank and take the logarithm of the obtained values.

3.3. PROPOSED EXCITATION FEATURES

The speech signal provides acoustical information about the speaker's state. It includes the emotion and health state of the speaker. The speech production system consists of two functional parts [12]. One is excitation produced at the larynx and another is filtering, which is excited by the excitation e(n). Thus, the speech signal s(n) can be produced by exciting the filter v(n) using an excitation.

$$s(n) = e(n) * v(n) \quad (2)$$

The excitation component contains relevant individual speakers' data as the vocal fold vibrations are distinct for a specific individual speaker [25]. The glottal vibrations, the shape of the glottal wave and the SoE of each glottal cycle are the characteristics of the speaker [25]. The variation of physical acoustical movement of excitation leads to changing essential features such as pitch and quality of voice [12]. The excitation features extracted from speech signals are Instantaneous Frequency (Fo), GCI and Glottal Opening Instant (GOI). The glottal flow shows a sudden negative peak in the excitation during the closing phase. This instant where the negative peak occurs is called GCI [12]. This peak is an important excitation to the vocal tract. LP analysis derives the excitation component from the speech signal [13]. Due to the abrupt variations in the movements of the vocal fold, the impulse-like event is produced which gives the prominent characteristic of the excitation [13]. In the LP residual, the instant where the maximum error value occurs refers to the GCI location. In linear prediction, each present sample is predicted from the past *l* samples [25]. This mathematical representation of prediction is given by (3).

$$\hat{s(n)} = -\sum_{k=1}^{l} a_k \cdot s(n-k) \quad (3)$$

The obtained difference between true and predicted sample is said to be the error signal as given in (4).

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{l} a_k . s(n-k) \quad (4)$$

Where a_k indicates prediction coefficients and l indicates the order of prediction. The obtained LP residual signal or glottal excitation suppresses vocal tract characteristics. The glottal excitation phase, or the residual phase, is unique for each individual, and thus the feature is helpful in recognition applications [18]. This feature can be extracted by taking the cosine phase of an analytic signal using the Hilbert transform $e_h(n)$. The analytic signal $e_a(n)$, cosine phase and Hilbert envelope $h_e(n)$ representations are given in (5), (6), (7), respectively.

$$e_a(n) = e(n) + je_h(n) \quad (5)$$

$$\text{cosinephas} = \frac{e(n)}{h_e(n)} \quad (6)$$

$$h_e(n) = \sqrt{(e^2(n) + e_h^2(n))} \quad (7)$$

The detailed flow chart of computation residual phase melspectrum is given in Fig. 1. Various Glottal Closure Instant(GCI) parameters which reflects the residual signal [13], [17] are discussed below.

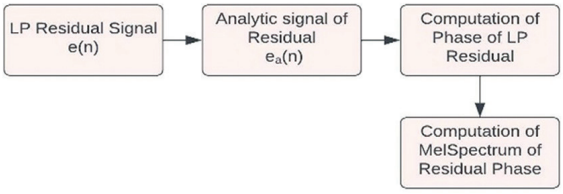

Fig. 1. The Block diagram for computation of Residual Phase based MelSpectrum

- **Strength of Excitation**: The strength is given by the substantial residual error at the prominent excitation and it varies by the rate of glottal closure [25]. The Strength of Excitation is referred to the magnitude of the impulse-like excitation [13]. It is computed as the amplitude of excitation at GCI. From the LP Residual signal, the SoE can be given by (8).

$$SoE = |h_e(n)| \quad \text{for all N} \quad (8)$$

- **Energy of Excitation:** LP residual is closely related to the excitation of the speech sample. EoE given in (9) reflects the vocal effort. This can be calculated from the Hilbert envelope around each GCI of LP Residual.

$$EoE = \frac{1}{(2k+1)} \sum_{i=0}^{2k} h_e^2(i) \quad (9)$$

- **Sharpness of Excitation:** The sharpness of excitation can be computed as the ratio of standard deviation (ŋ) to the mean (μ) [18]. The ŋ and μ are computed for the Hilbert envelope of excitation at the closure instance. The mathematical expression is given in (10).

$$sharpness = \frac{\eta}{\mu} \quad (10)$$

3.4. FUSION OF LOG-MELSPECTRUM AND EXCITATION FEATURES

The four sets of features are obtained as given in (11). The first vector represents the log-Mel spectrum of the

speech sample with 13 coefficients and the second vector represents the LP residual phase MelSpectrum with 13 coefficients for each frame. The third and fourth feature vector represents the sharpness and strength of the excitation. Thus, it forms the 28 features for each frame of the speech segment. Finally, all the feature vectors are fused to create a 28 x 198 size matrix, shown in (11) and fed to the dilated convolution neural network to accomplish the speaker identification task.

$$SS1 = \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1n} \\ M_{21} & M_{22} & \dots & M_{2n} \\ \vdots & \vdots & \dots & \vdots \\ M_{m1} & M_{m2} & \dots & M_{mn} \\ R_{11} & R_{12} & \dots & R_{1n} \\ R_{21} & R_{22} & \dots & R_{2n} \\ \vdots & \vdots & \dots & \vdots \\ R_{r1} & R_{r2} & \dots & R_{rn} \\ P_1 & P_2 & \dots & P_n \\ S_1 & S_2 & \dots & S_n \end{bmatrix} \quad (11)$$

Where:

SS1= first speech sample

M= log-melspectrum of speech sample 'SS1'

m= number of coefficients of M

R = melspectrum of residual phase for speech

r = number of coefficients of R

P = Sharpness of Excitation

S = Strength of Excitation

n = Number of frames

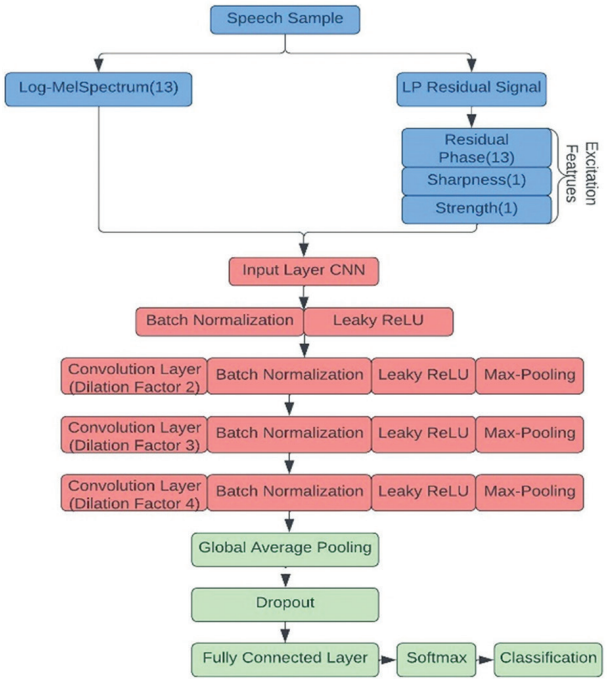

Fig. 2. Proposed Dilated Convolution Network with Excitation Features

3.5. DILATED CONVOLUTION

The learning capability of CNN depends on the design and depth of the convolution layers. Each layer learns and then gives the transformed data to the layers in the network and has sparse connectivity [26]. Since the speech signal changes continuously due to articulatory movements, it is considered non-stationary. So, working on short time segments of the speech signal is desired to get stable speech characteristics [9]. We used dilated convolution instead of traditional convolution due to the following reasons. The pooling layer imposes the data loss in the network by reducing the dimensionality. Without imposing a computational burden, dilated convolution increases the receptive field extensively. It learns the sparse relationship between the features. For example, a one-dimensional convolution kernel of size 4 x 1learns the sparse filter of 7 x1 with a dilation factor of 2. The dilation convolution kernel populated alternative positions when the dilation factor is two and this reduces the data loss. Also, learning the sparse relationship between the features is helpful in degraded speech as the frequency bands are degraded in dense regions [9]. Hence, dilated convolution prevents it from learning local dense regions due to sparse filters.

4. DATASET AND EXPERIMENTS

4.1. SPEECH DATASET

The LibriSpeech dataset in the experiments is used for evaluating certain aspects of speaker identification. The class or category with speech utterances of various speakers is named an unknown class. The experiments on the LibriSpeech corpus aim to analyze the proposed algorithm under the different number of speakers and different classes, including the unknown category. The publicly available LibriSpeech corpus provides the clean speech data of nearly 100 speakers. Each speaker's training and testing sets are disjoint to make it text-independent. It consists of 250 utterances by each speaker in the English language. We split the available data into the training and test with 80% and 20% proportions, respectively.

4.2. TRAINING

The proposed network schemes are implemented using Matlab 2021b and the deep learning toolbox is used to build the network. The model is trained for 25 epochs for all the experiments with a minibatch of 128. The number of iterations is changed for the different number of speakers. The validation data is used to tune hyper-parameters. The proposed dilated convolution network shown in Fig. 2 is trained using Adam optimization with an initial learning rate of 0.00001. The convolution kernel size and learning rate are selected based on a trial and error approach. The weights are updated throughout the learning process.

The tendency of the gradient of the sigmoid activation function is to shrink at every step throughout the depth of the network [1]. It could lead to a vanishing gradient problem. The ReLU creates the dead neuron

when the gradient becomes zero. Hence, in our experiments, the Leaky ReLU activation function was used after each convolution layer.

4.3. EXPERIMENTAL SETUP

The experimental overview of the proposed feature fusion and dilated convolution network is discussed in this section. Many experiments are conducted to measure the performance of the developed system. The following sets of experiments on the proposed model are performed to evaluate and compare its performance with baseline systems.

Firstly, the proposed excitation features are extracted from speech utterances to identify 11 classes, including one unknown class. The excitation features such as residual phase, sharpness of excitation, EoE and SoE are combined with log-MelSpectrum and given as input to the dilation convolution neural network for the classification of speakers. The four sets of features (four experiments) are given as input to Dilated CNN separately to evaluate the performance.

In the second set of experiments, the performance of the different number of speakers (say 100, 70, 56, 30,10) for the proposed model is evaluated. This setting allows us to evaluate the classification performance in classifying the large number of speakers.

Lastly, our model is compared with the baseline system and evaluates the performance of five female speakers. The improvement of the proposed feature fusion-based system is compared with several existing baseline- models by computing accuracy, EER and ROC.

5. EXPERIMENTAL RESULTS

5.1. EVALUATION METRICS

We compare the Accuracy, Equal Error Rate (EER), Receiver Operating Characteristics (ROC) curves and confusion matrices for both baseline and proposed systems of a varying number of speakers and various classes of speakers. These are chosen as performance metrics of the proposed method. We also evaluated and reported the accuracies of male and female speaker identification for five speakers of the proposed system. Accuracy is computed as the number of samples that are correctly identified divided by the total samples in the test subset. Equation (12) gives the mathematical expression for accuracy.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{12}$$

Equal Error Rate (EER) is computed by finding the same value of FAR and FRR. The system with low EER indicates good performance.

ROC is used to analyze the classifier's performance for each speaker class and gives a precise performance overview by plotting curves of each class. The Area Under the Curve(AUC) measures how efficiently the architecture separates classes [4]. A value of AUC nearly equal to one indicates the good conduct of the classifier, while this value of less than 0.5 indicates poor performance [8,10] . To understand and plot the ROC, the mathematical expressions of true positive (TP) and negative (TN) rates are useful as shown in (13), (14).

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

5.2. MODEL RESULTS

In this section, the results of various experiments are discussed as follows. We trained the dilated convolution neural network using Librispeech corpus with an available speaker training set to obtain the results and analyze the performance. Then, evaluate the trained model for identification of speakers' test data. First, the results of the proposed model with excitation and log-MelSpectrum with various classes, including unknown class are obtained. The results of the proposed method with the different number of speakers are also obtained. Lastly, the results of the comparison of the proposed network with baseline systems and with respect to gender are obtained. All these results are discussed in the following sections.

In the first set of experiments, the proposed model's overall accuracies and Equal Error rate with various excitation features are evaluated and compared with the previously developed models in the research. The reported accuracies ranged from 76.61% to 94.12% with the librispeech dataset shown in Table 1. As shown in Table 1, the proposed deep learning model with 28 features outperformed the existing models by obtaining an accuracy of 94.12% and an EER of 1.16% for speaker identification. In the other three methods, the advanced dilated convolution network, when trained and tested on the fusion of log-MelSpectrum and excitation features such as sharpness of excitation and EoE, obtained the highest accuracy of about 91.57% with EER 1.16%. The plot of accuracies of various methods is shown in Fig. 3.

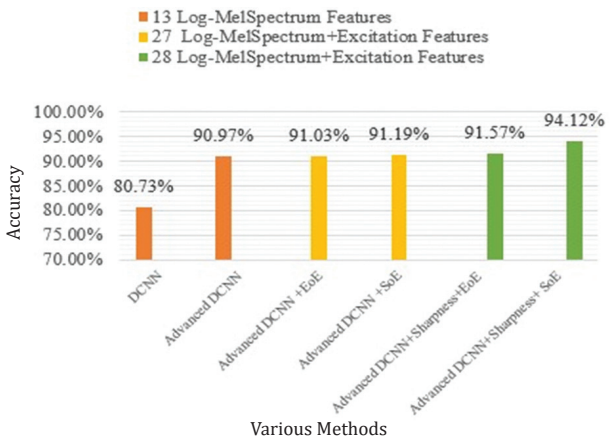

Fig. 3. Plot of Accuracies for various methods

Table 1. The performance of the identification models with 11 classes of speakers (including unknown class) using LibriSpeech data

Type of Model	Features	Dilation factor	Accuracy (%)	EER
CNN	MFCC	1	76.61%	5.90 %
Dilated CNN	Log-MelSpectrum	2	80.73%	4.82 %
AlexNet	Log-MelSpectrum	1	82.40%	-
Advanced Dilated CNN	Log-MelSpectrum	4	90.97%	2.38%
Advanced Dilated CNN	Log-MelSpectrum + LP Residual Phase MelSpectrum+EoE	4	91.03%	2.28%
Advanced Dilated CNN	Log-MelSpectrum + LP Residual Phase MelSpectrum + Sharpness	4	91.19%	2.27%
Advanced Dilated CNN(Proposed)	Log-MelSpectrum + LP Residual Phase MelSpectrum + Sharpness of Excitation + EoE	4	91.57%	1.16%
Advanced Dilated CNN (Proposed)	**Log-MelSpectrum + LP Residual Phase MelSpectrum + Sharpness of Excitation + SoE**	**4**	**94.12%**	**1.16%**

Across the eight experiments in Table 1, the proposed method correctly identifies an extra 2.59% of the test samples over advanced dilated CNN with sharpness and EoE. Also, an additional 3.15% of the test samples were over the Log-MelSpectrum model. Fig. 4 shows the confusion matrix that analyzes the number of test samples over false acceptance and rejection for the proposed method. The proposed system also improved the EER from 5.90% to 1.16%.

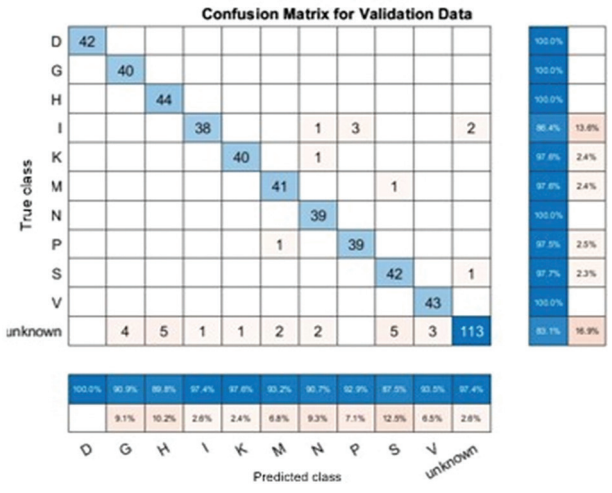

Fig. 4. Confusion matrix for validation data of the proposed method

The efficiency of the proposed method is revealed by comparing its performance with baseline systems. The experimental results of the baseline and proposed methods are shown in Table 2. The ratio of train and test sets is the same (80:20) for all the methods mentioned in Table 2. In general, the performance of the classification model decreases when the number of speakers increases. As presented in the Table 2, the proposed excitation features based dilated convolution network outperformed the baseline methods. The plot of training and validation accuracy is given in Fig. 5.

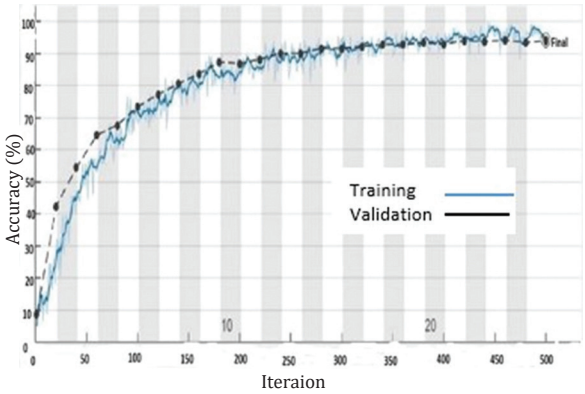

Fig. 5. Accuracy Vs Iterations plot of proposed method

Table 2. The performance comparison of proposed model with existing baseline models

Literature	Methodology	Features	Database	Speakers	Accuracy
Jahangir et.al. [10]	DNN	MFCC + MFCCT	Libri Speech	50	90 %
Ting Lin et.al. [22]	DNN	MFCC and Super MFCC	Libri Speech	10	90 %
S. Chakraborthy et. al. [27]	GMM-UBM	MFCC	Libri Speech	40	86 %
Pentapati et.al.[6]	Dilated CNN (dilation factor 2)	Log-Mel Spectrum	Libri Speech	11	80.6 %
Proposed	**Dilated CNN (dilation factor 4)**	**Excitation features and Log-Mel Spectrum**	**Libri Speech**	**80** **11 (with Unknown)**	**91.34 %** **94.12 %**

The baseline method[10] obtained 90% and 89% accuracy with 50 and 100 speakers, respectively. Compared with the baseline method [10], the proposed model shows improvement in performance by obtaining an accuracy of 91.34% with 80 speakers. To confirm this, the number of speakers gradually increased from 10 to 80. The experiments are conducted to evaluate the performance of each set of speakers as shown in Table 3. In all these four experiments, the proposed method outperformed the baseline method.

Table 3. Performance comparison of the proposed method with different number of speakers

Number of Speakers	Accuracy (%)	EER	Training error (%)	Validation Error (%)
10	97.61%	1.02%	0%	2.3923%
26	96.26%	1.16%	0.1348%	3.7409%
56	93.54%	2.08%	0.6827%	6.4551%
80	91.34%	3.11%	0.9093%	8.6585%

The comparative analysis of the proposed and baseline method [10] can be observed in Fig. 6. The accuracy of the proposed system is slightly lower than that of the baseline method [10] by 1.9% for ten speakers. Despite that lower value, the accuracy of our approach is much better on 26, 56 and 80 speakers compared to the method proposed in [10].

Across the several methods in Table 2, the proposed method improves the accuracy by 7.61% compared to [22] baseline method for 10 speakers. The accuracy of the proposed method further improves by nearly 8% and 14% compared to the baseline methods [27] and [6] respectively. The proposed work does not introduce computational complexity since extracting the log-melspectrum and excitation features have been done offline. Thus, the proposed model achieved improved accuracy with training time of 115 seconds per epoch.

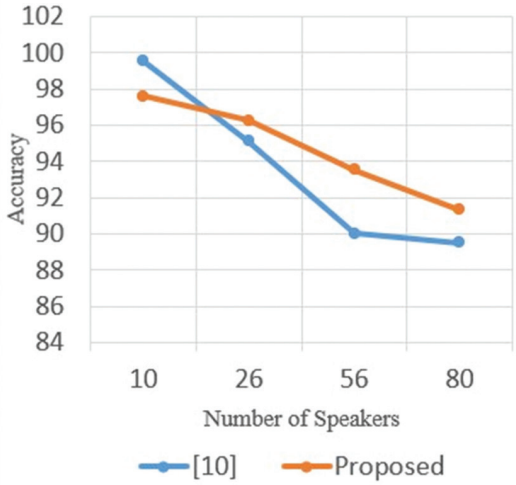

Fig. 6. Comparative analysis of proposed and baseline method

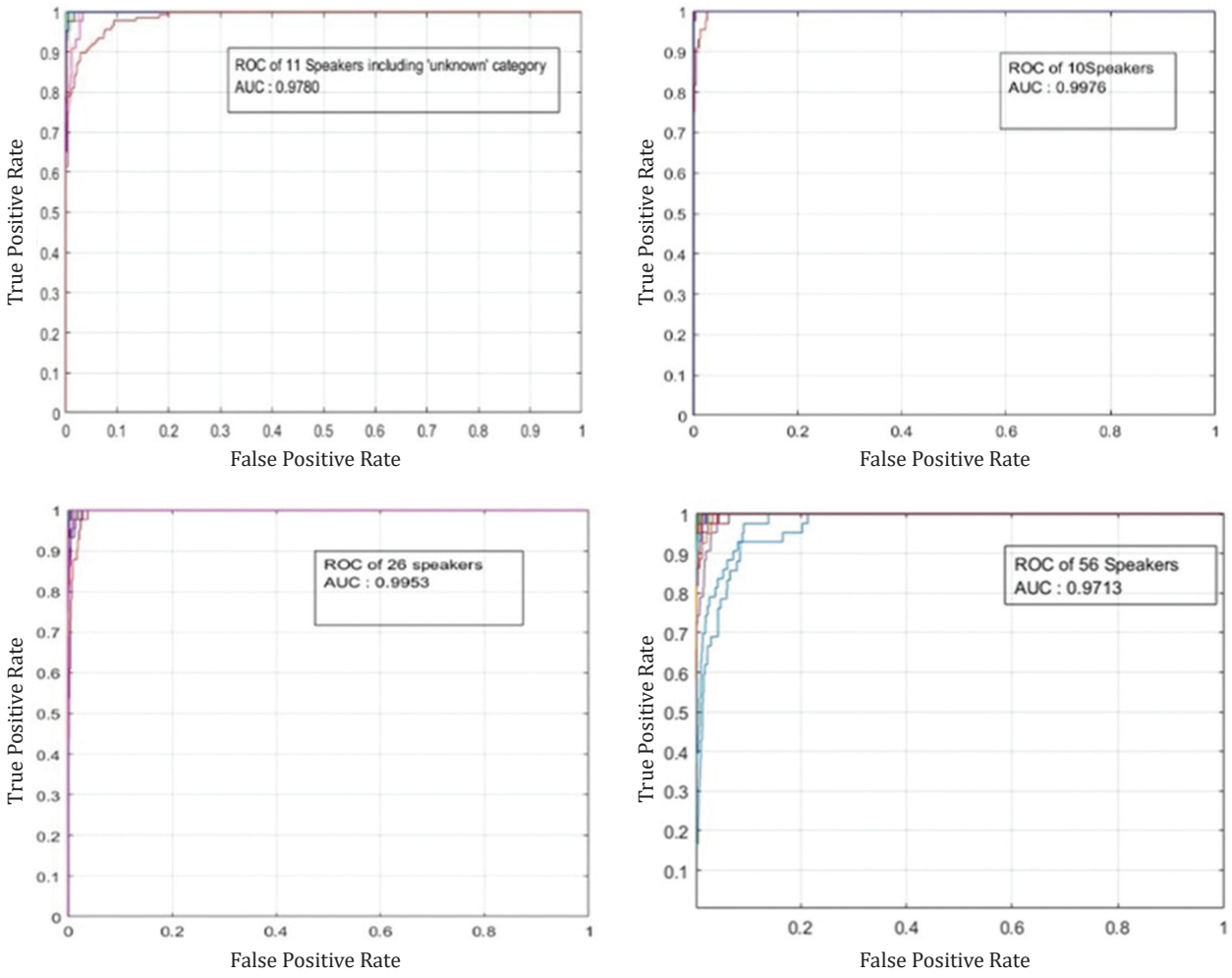

Fig. 7. ROC and AUC of the proposed method for different number of speakers

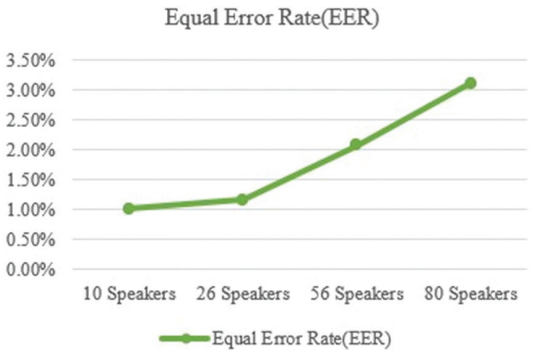

Fig. 8. Variation of EER for different number of speakers

In Table 2, Jahangir et al. [10] were achieved an accuracy of 90% for 50 speakers since an efficient MFCCT-based feature fusion was employed. Although they used novel feature fusion, the model didn't reach the desired accuracy due to simpler deep learning architecture. Thereby, misclassification happens due to similar voice patterns. Ting Lin et al. [22] were also extracted novel LTA-based features along with MFCC. They used simple DNN and tested only ten speakers on the Librispeech corpus and achieved 90% accuracy. In contrast, the proposed method used an advanced dilated convolution network with feature fusion of excitation features. Hence, enhancing the accuracy to 91.34% for 80 speakers. In Table 1, the previous model with advanced dilated CNN achieved only 90.97% accuracy since only log-melspectrum was used in the feature extraction phase.

Across the four experiments shown in Table 3, the proposed model with 10 speakers correctly identifies an additional 1.35% of the test samples over 26 speakers, 4.07% over 56 speakers, 6.27% over 80 speakers. The Fig. 7 shows ROC and AUC plots of proposed method with 10,11,26 and 56 speakers. This shows AUC of 10 speakers achieved the highest value of 0.9976 when compared with other sets. As shown in Fig.8, It also improved the EER from 3.11% to 1.02%.

The log-spectrum based baseline method [21] obtained the highest accuracy of 98.7% with five female speakers on the Chinese mandarin database. The Five female speakers with only 100 utterances per speaker, a frame size of 256 samples and 128 features were considered in the model proposed by [21]. In contrast, the proposed model considered the 250 utterances of each speaker with a frame size of 400 samples.

The Librispeech and Chinese Mandarin datasets considered the undistorted samples with the same sampling rates. Also, the training and testing sets are disjoint and collected under the same environmental conditions in both datasets. As the number of utterances to classify increases, the model performance decreases. To confirm the improvement in the proposed design, the experiment is conducted to test the proposed method with five male and five female speakers. As shown in Table 4, the proposed model obtained an accuracy of 99.52% for female speakers. It correctly identifies an additional 0.82% of the test samples over the baseline [21].

Table 4. Performance comparison of the proposed method and baseline method [21] for five Speakers

Method	Dataset	Gender	Features	Frame size in samples	Accuracy in %
Samia Abd El-Moneiet. al. [21]	Chinese mandarin Corpus	Female	Log-Spectrum	256	98.7%
Proposed	Libri Speech Corpus	Female	Log-Melspectrum + Residual Phase + Sharpness +SoE	400	99.52%
Proposed	Libri Speech Corpus	Male	Log-Melspectrum + Residual Phase + Sharpness + SoE	400	99.49%

6. CONCLUSION

In this paper, the efficient excitation-based features are considered and combined with Log-MelSpectrum of the speech sample. The excitation-based features such as the Residual phase, Sharpness of excitation, EoE and SoE are extracted using Linear prediction analysis. Extensive experiments are conducted on the proposed advanced dilated convolution network with excitation features using the LibriSpeech corpus. It showed that the accuracy and EER are 94.12% and 1.16%, respectively, with 11 complex classes including the unknown category. For 80 speakers, the proposed method achieved an accuracy of 91.34% and the accuracy reached 99.52% for five female speakers. The experimental results proved that the proposed speaker identification system is effective with respect to the number of speakers and computational complexity compared to existing baseline systems. In the future, we aim to develop a speaker identification system with a more profound architecture to handle a much more extensive database and reduce the misclassified samples. The hyper-parameter tuning process can be enhanced to improve the system's performance further.

7. REFERENCES:

[1] F. Ye, J. Yang, "A deep neural network model for speaker identification", Applied Sciences, Vol. 11, No. 8, 2021, pp. 1-18.

[2] X. Wang, F. Xue, W. Wang, A. Liu, "A network model of speaker identification with new feature extraction methods and asymmetric BLSTM", Neurocomputing, Vol. 403, 2020, pp. 167-181.

[3] S. Farsiani, H. Izadkhah, S. Lotfi, "An optimum end-to-end text-independent speaker identification system using convolutional neural network", Computers and Electrical Engineering, Vol. 100, 2022, p. 107882.

[4] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities", IEEE Access, Vol. 9, 2021, pp. 79236-79263.

[5] S. Hourri, J. Kharroubi, "A deep learning approach for speaker recognition", International Journal of Speech Technology, Vol. 23, No. 1, 2020, pp. 123-131.

[6] H. K. Pentapati, Sridevi K, "Dilated Convolution and MelSpectrum for Speaker Identification using Simple Deep Network", Proceedings of the 8th International Conference on Advanced Computing and Communication Systems, Coimbatore, India, 25-26 March 2022, pp. 1169-1173.

[7] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, R. Wang, "Speaker identification features extraction methods: A systematic review", Expert Systems with Applications, Vol. 90, 2017, pp. 250-271.

[8] R. Jahangir, Y. W. Teh, F. Hanif, G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges", Multimedia Tools and Applications, Vol. 80, No. 16, 2021, pp. 23745-23812.

[9] A. Chowdhury, A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals", IEEE Transactions on Information Forensics and Security, Vol. 15, 2020, pp. 1616-1629.

[10] R. Jahangir et al., "Text-Independent Speaker Identification through Feature Fusion and Deep Neural Network", IEEE Access, Vol. 8, 2020, pp. 32187-32202.

[11] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 19-24 April 2015, pp. 5206-5210.

[12] S. R. Kadiri, P. Alku, B. Yegnanarayana, "Extraction and Utilization of Excitation Information of Speech: A Review", Proceedings of the IEEE, Vol. 109, No. 12, 2021, pp. 1920-1941.

[13] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, B. Yegnanarayana, "Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference", Circuits, Systems and Signal Processing, Vol. 39, No. 9, 2020, pp. 4459-4481.

[14] M. K. Gill, R. Kaur, J. Kaur, "Vector Quantization based Speaker Identification", International Journal of Computer Applications, Vol. 4, No. 2, 2010, pp. 1-4.

[15] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, Vol. 10, No. 1, 2000, pp. 19-41.

[16] H. Meng, T. Yan, F. Yuan, H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network", IEEE Access, Vol. 7, 2019, pp. 125868-125881

[17] S. R. Kadiri, P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection", IEEE Access, Vol. 8, 2020, pp. 60382-60391.

[18] T. Thomas, Spoorthy, N. V. Sobhana, S. G. Koolagudi, "Speaker Recognition in Emotional Environment using Excitation Features", Proceedings of the Third International Conference on Advances in Electronics, Computers and Communications, Bengaluru, India, 11-12 December 2020.

[19] H. Ali, S. N. Tran, E. Benetos, S. Artur, A. Garcez, "Speaker recognition with hybrid features from a deep belief network", Neural Computing and Applications, Vol. 29, No. 6, 2018, pp. 13-19.

[20] S. Nainan, V. Kulkarni, "Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN", International Journal of Speech Technology, Vol. 24, No. 4, 2021, pp. 809-822.

[21] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement", Multimedia Tools and Applications, Vol. 79, No. 33-34, 2020, pp. 24013-24028.

[22] T. Lin, Y. Zhang, "Speaker recognition based on long-term acoustic features with analysis sparse representation", IEEE Access, Vol. 7, 2019, pp. 87439-87447.

[23] Vasamsetti Srinivas, Ch Santhi Rani, "Optimization-Based Support Vector Neural network for Speaker Recognition", The Computer Journal, Vol. 63, No. 1, 2020, pp. 151-167.

[24] M. Soleymanpour, H. Marvi, "Text-independent speaker identification based on selection of the most similar feature vectors", International Journal of Speech Technology, Vol. 20, No. 1, 2017, pp. 99-108.

[25] S. R. Mahadeva Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", Speech Communication, Vol. 48, No. 10, 2006, pp. 1243-1261.

[26] Z. Liu, Z. Wu, T. Li, J. Li, C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition", IEEE Transactions on Industrial Informatics, Vol. 14, No. 7, 2018, pp. 3244-3252.

[27] S. Chakraborty, R. Parekh, "An improved approach to open set text-independent speaker identification (OSTI-SI)", Proceedings of the Third International Conference on Research in Computational Intelligence and Communication Networks, 3-5 November 2017, pp. 51-56.