FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

A Systematic Approach of Advanced Dilated Convolution Network for Speaker Identification

Hema Kumar Pentapati[1*] and Sridevi K[2]

[1*]*Research Scholar, Department of EECE, GITAM School of Technology, Visakhapatnam, India, hpentapa@gitam.in*
[2]*Associate Professor, Department of EECE, GITAM School of Technology, Visakhapatnam, India, skataman@gitam.edu*

***Correspondence:** Hema Kumar Pentapati; hpentapa@gitam.in

ABSTRACT- Over the years, the Speaker recognition area is facing various challenges in identifying the speakers accurately. Remarkable changes came into existence with the advent of deep learning algorithms. Deep learning made a remarkable impact on the speaker recognition approaches. This paper introduces a simple novel architectural approach to an advanced Dilated Convolution network. The novel idea is to induce the well-structured log-Mel spectrum to the proposed dilated convolution neural network and reduce the number of layers to 11. The network utilizes the Global average pooling to accumulate the outputs from all layers to get the feature vector representation for classification. Only 13 coefficients are extracted per frame of each speech sample. This novel dilated convolution neural network exhibits an accuracy of 90.97%, Equal Error Rate (EER) of 3.75% and 207 Seconds training time outperforms the existing systems on the LibriSpeech corpus.

Keywords: Log-Mel Spectrum, MFCC, Dilated Convolution neural networks, Speaker Identification, Deep Learning.

ARTICLE INFORMATION

Author(s): Hema Kumar Pentapati and Sridevi K;
Received: 16/11/2022; **Accepted**: 16/01/2023; **Published**: 05/02/2023;
e-ISSN: 2347-470X;
Paper Id: IJEER 1611-06;
Citation: 10.37391/IJEER.110104
Webpage-link:
www.ijeer.forexjournal.co.in/archive/volume-11/ijeer-110104.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Speaker recognition refers to authenticating the person based on the specific voice characteristics from their speech utterances [1]. Applications of speaker recognition include many domains, such as forensics, banking services, and security services[2]. Speaker recognition is categorized into two classes, Speaker Identification (SI) and speaker verification. Automatic SI refers to identifying the speaker from a set of speakers known to the system. Speaker verification includes finding whether the speaker of the given utterance refers to a set of speakers[3].

Implementation of the Speaker Recognition system consists of two phases (i) Feature Extraction and (ii) Classification. The most popular technique for extracting most speaker-related information from the given speech utterance is Mel Frequency Cepstral Coefficients (MFCC)[4,5,6]. Over the years, speaker recognition has used baseline Vector Quantization(VQ), Gaussian Mixture Models(GMM), and Artificial Neural Networks (ANN) gained success in its implementation[7,8]. Furthermore, Deep learning approaches can handle more complex structures[8]. Significantly, the convolution neural network (CNN) learns features even from the slight variations present in the data[8] and is more robust against noise[9]. Deep learning extends its capabilities in various domains, including speaker recognition. Researchers find the usefulness of deep learning methods in finding the solution to recognition problems[10,11,12]. They found new ways to use deep learning approaches[13].The existing speaker identification systems are not highly accurate. In addition, high model complexity[5] is one issue that bothers most researchers in speaker recognition. The other problem is the training time required to achieve a high recognition rate.

The main contribution of this paper is as follows. Inspired by the success of speaker recognition using deep learning models and novel ideas on CNN, a novel architecture with Advanced Dilated Convolution is proposed to implement end to end speaker identification system. The extracted Mel-spectrograms from speech utterances are fed as input to the dilated CNN model for learning the feature maps and classification after processing. Finally, the conducted experiments on the proposed model for speaker identification showed numerous advantages, including *(a)* More flexible convolution operation and significant reduction in model complexity, *(b)* The dilation convolution concept reduces data loss and prevents learning from dense regions, which makes the system robust to noise *(c)* Effective training time is the added advantage.

2. RELATED WORKS

The development of speaker recognition systems is proliferating in artificial intelligence. Over the decades, great progress has been made in speaker identification. The Hidden Markov Models and Gaussian Mixture Models(GMM)[14] became very popular along with parametric methods such as Vector Quantization(VQ) and Dynamic Time Warping. Tirumala *et al.*[15] were conducted a review on the feature extraction methods include MFCC, LPC for speaker recognition and provided criteria for appropriate features. Further, different MFCC variations and approaches are prevalent in recent research conducted by Hourri *et al.* [16].

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

Jahangir *et al.*[3] proposed the Deep Neural Network(DNN) based feature combination of MFCC and MFCC-time-based features. They evaluated the model on LibriSpeech Corpus and achieved an accuracy of 89% for 100 speakers. But, it still fails to classify some of the samples correctly. Chowdhury *et al.*[2] proposed fusion-based feature modeling using the 1D triplet CNN. They conducted various experiments on different databases to evaluate their model. Experiments show the model's effectiveness and misclassify 14% of the total samples. Ting Lin *et al.*[17] experimented with stacking the MFCC features. The long-term acoustic feature (LTA) based on DNN was evaluated on LibriSpeech corpus, it shows efficiency by achieving 90% accuracy for ten speakers.

Zheli Liu *et al.*[7] proposed GMM- CNN-based hybrid model for short utterances. The experimental results show that it achieved 90% test accuracy, and EER was reduced to 2.5%. The spectrograms were processed to detect the deep features of the short speech utterance and gained improvement in accuracy due to the discovery of rich features[7]. Speaker recognition systems should learn and recognize speakers even with minimal data, and researchers found it to be a great challenge[18]. A vast amount of data is required for training to learn the various features[17,18]. Meng H *et al.*[5] proposed and designed the Dilated CNN with residual and attention mechanism(ADRNN) network and the Log-Mel Spectrum was fed to the network for emotion recognition.

3. EXTRACTION OF LOG MELSPECTRUM

The raw speech signal should be in the form of .wav format, and process the signal to extract the features. The speech utterance can be split into overlapping segments of the same duration, called frames. For the sampling rate of speech signal 16000Hz and window of 25ms, the vector dimension of each frame is computed and obtained 400 samples per frame. Here, every frame should pass through a Hamming window. For frame step of 10ms, the step size can be calculated as 160samples. The total number of frames N can be computed using *equation (1).* After the signal is blocked into frames, the resulting signal due to windowing is given in *equation (2).*

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

$$x_i(n) = w(n) \cdot s_i(n) \quad (2)$$

The usage of Mel filter bank in accordance with the human perception of hearing makes the MFCC the most popular approach in feature extraction. The envelope of the power spectrum is used to depict the shape of the vocal tract. It presents the vocal tract information more accurately which gives the acoustic properties of the signal[15]. Take the discrete Fourier transform of each windowed frame to yield the complex spectral numbers. Let the size of Fourier coefficients be K, and the window is denoted as w(n). The Fourier coefficients can be computed using *equation (3).* The periodogram estimate of the power spectrum is shown in *equation (4).* The periodogram approximately models the human cochlea, which vibrates according to the sound received[16]. The mel is computed by using (5).

$$X(k) = \sum_{n=1}^{N} x(n) \, . \, e^{-j2\pi kn/N} \quad (3)$$

$$P_i(k) = \frac{1}{N} |X_i(k)|^2 \quad (4)$$

$$Mel(f) = 2595 * log_{10}(1 + \frac{f}{700}) \quad (5)$$

Here f indicates the regular frequency and Mel shows the perceived frequency. Assume the number of triangular filters used is 26. The log-melspectrum can be obtained by multiplying the magnitude power spectrum with each Mel-filter bank and taking logarithm. It forms the structure to the feature map of size 13 x 198 matrix as shown in *equation (6).* The number of frames can be computed using *equation (1).*

$$S1 = \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1n} \\ M_{21} & M_{22} & \dots & M_{2n} \\ \vdots & \vdots & \dots & \vdots \\ M_{m1} & M_{m2} & \dots & M_{mn} \end{bmatrix} \quad (6)$$

Where m = Number of coefficients of log-melspectrum.
n = Number of frames

4. PROPOSED ARCHITECTURE

In this section, our proposed work for text-independent speaker identification is introduced to improve the experiments and to acquire better results. CNN does not require well-pre-processed data input to learn the data effectively; thus, a minimum pre-processing is sufficient [18]. It works to hold the local correlations of the feature map[19]. Let us consider x(n) to be input, then the output y(n) is obtained by convolving the x(n) with the kernel w(n) is given in *equation (7).*

$$y(n) = x(n) * w(n) = \sum_{p=n-k} x(k).w(p) \quad (7)$$

Where k denotes the kernel's size. First, various speakers' speech utterances are collected from LibriSpeech Corpus. The log-Mel Spectrum is extracted to form the feature set of each speech sample. This structured feature matrix is fed as an input to the novel architecture with dilated convolution to accomplish the speaker identification task. The experimental models and methods are discussed in the following sections. The model integrates dilated convolution and removes the max-pooling operation. The dilation factor is increased by 1 in each convolution layer, and the reason behind using dilated convolution is elaborated in next section. Global average pooling consolidates the data at the end [6]. The dilated convolution is shown in *figure 1* and the proposed setup is shown in *figure 2.*

4.1 Dilated Convolution

The learning efficiency of CNN depends significantly on the design of convolution layers. Every convolution layer learns from the data along the network and transforms the data further to the following layers in the network[19]. The convolution with holes establishes the sparse features, and this concept is

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

called dilated convolution, as shown in *figure 1*. This spans the large receptive field with the same number of parameters, and it is given in *equation (9)*.

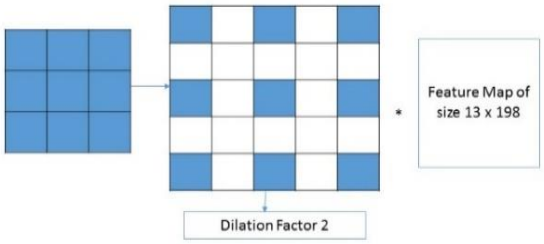

Figure 1: Dilated convolution with dilation factor 2

Figure 1 shows the populated 3*3 kernel in convolution operation. The dilation factor of 2 makes the filter populate the alternative positions, and then it gets convolved with the input matrix derived through the Log-Mel spectrum. Thus, the dilation factor (D) with the new dimension of the kernel (N) and the old dimension (O) is given in *equation (8)*. For the dilation factor of D=2, the output $Y_2(n)$ is convolution of input x(n) with the kernel w(n) and it is given in *equation (9)*.

$$D = \frac{N}{O} \tag{8}$$

$$Y_2(n) = w(n) *_D x(n) = \sum_k w(k).x(n - D.k) \tag{9}$$

The weights of the filters get updated throughout the training till the loss gets minimized. The i^{th} output feature of any k^{th} layer is given in *equation (10)*.

$$Y_i^k(n) = B + \sum Y_i^{k-1}(n).w_{ij} \tag{10}$$

In the process of reducing the dimensionality using the pooling operation, it is revealed that there is some data loss within the network. The Dilated convolution increases the receptive field significantly without much computational cost and reduces the data loss due to max-pooling [2]. These are the two significant advantages of introducing dilated convolution in the network.

4.2 Batch Normalization and Leaky ReLU

The need for batch normalization is essential to normalize the input of the next layer by taking the mean of the feature vector to zero and also adjusting the variance nearly equal to one[5]. This update with batch normalization and β, γ as adjustable parameters is given in *equation (11)*. Leaky ReLU solves the problem of dead neuron by reducing the negative value to a small value of 0.01*x instead of zero and it is shown in *equation (12)*. Applying the activation function for the $Y_{2B}{}^k(n)$ gives the function shown in *equation (13)* and ϕ represents activation function.

$$Y_{2B}^k(n) = \beta + \gamma \frac{(Y_i^k(n) - \mu)}{\sqrt{\sigma^2}} \tag{11}$$

$$a_i = \begin{Bmatrix} x & if\ x \geq 0 \\ 0.01x & if\ x < 0 \end{Bmatrix} \tag{12}$$

$$Y_{2B}^k(n) = \varphi(BN(Y_i^k(n))) \tag{13}$$

The global average pooling averages the features in each channel returns only a single value for each channel instead of a one-dimensional vector and avoids overfitting at this stage[12]. Also, the removal of the max-pooling layer in the consecutive steps introduces an extra computational burden.

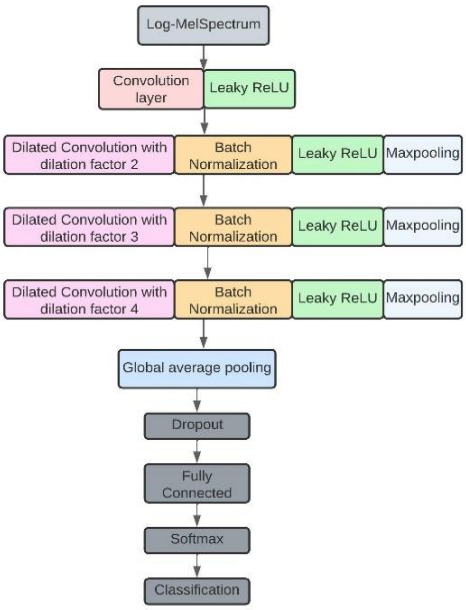

Figure 2: The architecture of proposed advance Dilated Convolution Network

Table 1: Tunable parameters of the System

Name	Activations	Parameters	Dilation factor
Convolution layer	198*13*32	3*3*1*32	1
Dilated Convolution	198*13*64	3*3*32*64	2
Maxpooling	97*5*64		
Dilated Convolution	97*5*96	3*3*64*96	3
Maxpooling	49*3*96	-	-
Dilated Convolution	49*3*128	3*3*96*128	4
Maxpooling	25*2*128	-	-
GlobaAveragePooling	1*1*128	-	-
Dropout (20%)	1*1*128	-	-
Fully Connected	1*1*11	11*128	-
Softmax	1*1*11	-	-

The Dilated Convolution holds much-extended information and appears superior to the max-pooling operation. The input is initially convolved with the convolution filters of size 32. The details of each layer in the proposed work are illustrated in *table 1*.

5. DATASETS AND EXPERIMENTS

5.1 Database

To confirm the ability of the proposed framework, we use the Libri speech database in our experiments. It is part of the audiobooks of LibriVox, which contains nearly 1000hours of data sampled at 16 KHz[20]. This database can be available freely, along with training data. Over 2628 speech utterances of

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

male and female speakers are extracted from the audiobooks to classify the speakers. The utterances of the speech corpus are improvised by taking equal-duration samples.

5.2 Experimental Setup

All the experiments on the proposed model are conducted on i5-6500 CPU and implemented using latest MATLAB environment. Our experiments are conducted to train the proposed model for 25epochs with a minibatch of 32. It is proposed to deploy the Adam optimization technique with a learning rate of 0.0001initially. Also, weights of the convolution kernel are updated with random values initially and later updated through the backpropagation along the network based on the loss function at the end. The dropout of 0.1 is deployed to overcome the overfitting problem in the training process. The overall progress of the training process of the network is shown in *figure 3*.

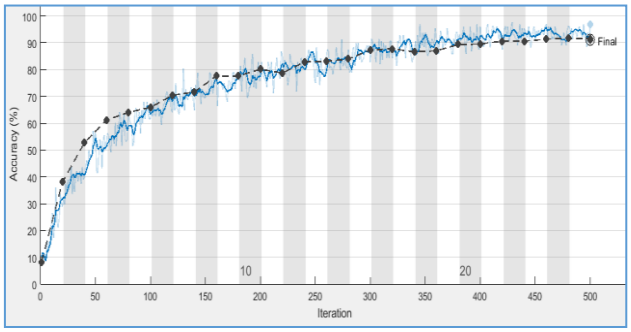

Figure 3: Accuracy vs Training iterations of proposed method

6. EXPERIMENTAL RESULTS AND ANALYSIS

6.1 Evaluation Metrics

The performance measurement of the speaker identification systems depends on two metrics, namely, false acceptance and rejection rates. FAR refers to falsely accepting the non-target speaker as the actual speaker, whereas FRR refers to falsely rejecting the true speaker for a non-target speaker. These are obtained through confusion matrices. The accuracy can be obtained by the ratio of correctly predicted samples to the number of validation data as given in (14). EER is the intersecting point of FAR and FRR curves. In addition to accuracy, some other important metrics like Precision, Recall, and F1 Score are used to analyze each class's robustness and overall misclassifications[21][22].

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (14)$$

6.2 Results and Discussions

The proposed dilated CNN is designed to classify 11 different classes of speakers. The data is arranged as 80% samples of each class are used to train the network, and 20% samples are used for validation. The speaker identification experiment with the LibriSpeech corpus of our proposed architecture acquires more significant superiority in classifying the speakers correctly than the architecture with a dilation factor of 2. For example, the speaker 'I' is easily confused with other classes. Thus only 19 utterances are identified correctly with the architecture of dilated factor 2, as shown in *figure 5*. Whereas 34 utterances are correctly identified for the identification of the speaker 'I' with the proposed architecture. It is shown in *figure 4*.

Despite good accuracy, the class 'I" still records a low 77.3% true positive rate. It improves by 33.4% when compared with the architecture with dilation factor 2. Similarly, we acquired 18.61% distinguishable improvement in the class 'V' of 97.69% when compared to other experimental tests.

Predicted Class

True Class	D	G	H	I	K	M	N	P	S	V	unknown
D	39			2							1
G		40									
H			44								
I	1			34	1			1		7	
K					40		1				
M				1		41					
N					1		38				
P								40			
S			1						42		
V										42	1
unknown	5	2	2	4	1	1	6		7	4	104

Figure 4: Confusion matrix of the proposed system with increased dilation factor of 4

Predicted Class

True Class	D	G	H	I	K	M	N	P	S	V	unknown
D	39			1			1	1			
G		39			1						
H			38					1			5
I	6		1	19	2			6		8	2
K		2		1	35		1			1	1
M						40		1			1
N							39				
P						2		37			1
S			1						38		4
V	1			3	3		1			34	1
unknown	9	1	8	6	1	2	3	1	11	3	82

Figure 5: Confusion matrix of the proposed system with dilation factor of 2

Further, the total processing time of the proposed model is compared with different models to prove the efficiency of the model. Because of our advanced architecture, the model does not introduce time complexity. We developed different models and evaluated them with the same data at various times during the research. The *table 2* shows the accuracy, EER and training time of those different models. The previous work of dilated CNN with dilation factor 2 had 80.73% accuracy and thus had an improvement of over 10% in the comparison. From the Table II, it is evident that there exists a slight difference of 100s in training time. Still, it raised the 90.97% accuracy rate in the proposed system compared with the 80.73% accuracy of dilated CNN with dilation factor 2. The improvement in EER is evident from *table 2* as it reduces from 5.90% to 3.75%. It attains the apparent improvement on 207s training time and 90.97% accuracy compared with AlexNet of 743s training time and 82.40% accuracy.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

Table 2: Accuracy, EER and Training Time of the Proposed Dilated CNN Models

Type of Model	Number of Coefficients	Dilation factor	Accuracy	EER	Training time per Epoch
CNN+MFCC	13	1	76.61%	5.90%	37.24s
DilatedCNN+Melspectrum	26	2	76.05%	5.67%	204s
DilatedCNN+Melspectrum	13	2	80.73%	4.82%	107s
AlexNet+Melspectrum	26	1	82.40%	4.55%	743s
Proposed method	**13**	**4**	**90.97%**	**3.75%**	**207s**

Across the four experiments in *table 3*, the proposed model's precision, recall, and F1 score improve upon the baseline system CNN-MFCC by 0.1179, 0.1314, and 0.1251, respectively. It also enhances the required training time to 207s per epoch, as given in *table 2*.

Table 3: Comparison of Model Performance Parameters

Type of Model	Number of Coefficients	Dilation factor	Precision	Recall	F1 Score
CNN+MFCC	13	1	0.8227	0.7722	0.7966
Dilated CNN+Melspectrum	13	2	0.8415	0.8017	0.8211
AlexNet+Melspectrum	26	1	0.8596	0.8215	0.8401
Proposed method	**13**	**4**	**0.9406**	**0.9036**	**0.9217**

6.3 Comparison of various baseline systems

The performance of the proposed model is compared with some baseline models to analyze the effectiveness of the proposed method. The results proved that the proposed model with advanced dilated convolution outperformed some baseline systems[17], [23], as shown in *table 4.* However, there is only a marginal difference in performance between the proposed and baseline methods[3]. But still, the proposed model consumes less training time because of low model complexity. Besides, the number of speakers used in the baseline method[3] is quite large compared to the proposed model. Though the proposed model obtained better performance, the number of speakers might still affect the performance. Further, it is desired to extend the model for the large corpus. Also, the 9.02% validation error is better than the 19.26% validation error of other designed models. The overall accuracies of various designed models and existing methods are shown in *figure 6*.

Table 4: Performance Comparison of various existing methods

Methods	Datasets	Accuracy (%)
MFCCs +DNN[17]	LibriSpeech	90.0
GMM UBM[23]	LibriSpeech	86.0
MFCCT+DNN[3]	LibriSpeech	89.0
MFCC derivatives+1D CNN[4]	VidTimit	73.25
Proposed Dilated CNN	**LibriSpeech**	**90.97**

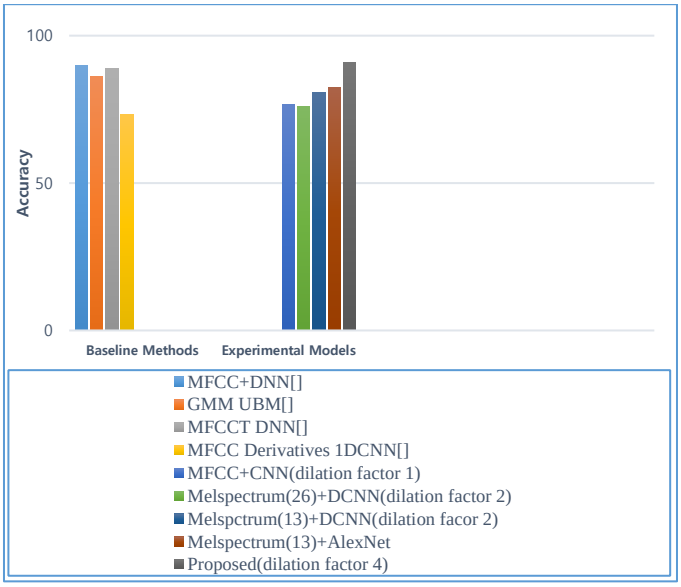

Figure 6: The overall performance comparison of proposed model with baseline models

7. CONCLUSION AND FUTURE SCOPE

In this paper, a novel architecture for speaker identification is proposed. The importance of the Log-Mel spectrum in extracting and learning local contextual information is investigated. The dilation factor is increased by one unit for every convolution layer and removed the max-pooling layer.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 25-30 | e-ISSN: 2347-470X

The performance of the designed system is tested on the standard Libri speech corpus. The comparison of the results showed that the proposed dilated CNN has a more significant advantage over baseline models in 90.97% overall accuracy, 3.75% EER, and 207s training time. Though the proposed architecture has obtained satisfactory results, a few areas still require enhancement. In the future, it is desired to reduce the classification errors due to similar speech patterns using much deeper architecture. Maintaining the system's stability under noisy speech data is necessary. In addition, it is required to improve the model further to classify a large number of speakers.

REFERENCES

[1] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," IEEE Access, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.

[2] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 1616–1629, 2020, doi: 10.1109/TIFS.2019.2941773.

[3] R. Jahangir et al., "Text-Independent Speaker Identification through Feature Fusion and Deep Neural Network," IEEE Access, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[4] S. Nainan and V. Kulkarni, "Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN," Int. J. Speech Technol., vol. 24, no. 4, pp. 809–822, 2021, doi: 10.1007/s10772-020-09771-2.

[5] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network," IEEE Access, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[6] Mahesh K. Singh, S. Manusha, K.V. Balaramakrishna and Sridevi Gamini (2022), Speaker Identification Analysis Based on Long-Term Acoustic Characteristics with Minimal Performance. IJEER 10(4), 848-852. DOI: 10.37391/IJEER.100415.

[7] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3244–3252, 2018, doi: 10.1109/TII.2018.2799928.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[9] X. Wang, F. Xue, W. Wang, and A. Liu, "A network model of speaker identification with new feature extraction methods and asymmetric BLSTM," Neurocomputing, vol. 403, pp. 167–181, 2020, doi: 10.1016/j.neucom.2020.04.041.

[10] Mahesh K. Singh, P. Mohana Satya, Vella Satyanarayana and Sridevi Gamini (2022), Speaker Recognition Assessment in a Continuous System for Speaker Identification. IJEER 10(4), 862-867. DOI: 10.37391/IJEER.100418.

[11] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," Sensors (Switzerland), vol. 20, no. 21, pp. 1–18, 2020, doi: 10.3390/s20216008.

[12] T. W. Sun, "End-to-End Speech Emotion Recognition with Gender Information," IEEE Access, vol. 8, pp. 152423–152438, 2020, doi: 10.1109/ACCESS.2020.3017462.

[13] S. Hourri and J. Kharroubi, "A deep learning approach for speaker recognition," Int. J. Speech Technol., vol. 23, no. 1, pp. 123–131, 2020, doi: 10.1007/s10772-019-09665-y.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process. A Rev. J., vol. 10, no. 1, pp. 19–41, 2000, doi: 10.1006/dspr.1999.0361.

[15] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," Expert Syst. Appl., vol. 90, pp. 250–271, 2017, doi: 10.1016/j.eswa.2017.08.015.

[16] S. Hourri, N. S. Nikolov, and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," Int. J. Speech Technol., vol. 24, no. 2, pp. 389–400, 2021, doi: 10.1007/s10772-021-09795-2.

[17] T. Lin and Y. Zhang, "Speaker recognition based on long-term acoustic features with analysis sparse representation," IEEE Access, vol. 7, pp. 87439–87447, 2019, doi: 10.1109/ACCESS.2019.2925839.

[18] A. Q. Ohi, M. F. Mridha, M. A. Hamid, and M. M. Monowar, "Deep Speaker Recognition: Process, Progress, and Challenges," IEEE Access, vol. 9, pp. 89619–89643, 2021, doi: 10.1109/ACCESS.2021.3090109.

[19] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, 2018, doi: 10.1109/LSP.2018.2860246.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2015-August, pp. 5206–5210, Aug. 2015, doi: 10.1109/ICASSP.2015.7178964.

[21] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, Deep learning approaches for speech emotion recognition: state of the art and research challenges, vol. 80, no. 16. Multimedia Tools and Applications, 2021. doi: 10.1007/s11042-020-09874-7.

[22] T. J. Sefara and T. B. Mokgonyane, "Emotional Speaker Recognition based on Machine and Deep Learning," 2020 2nd Int. Multidiscip. Inf. Technol. Eng. Conf. IMITEC 2020, 2020, doi: 10.1109/IMITEC50163.2020.9334138.

[23] S. Chakraborty and R. Parekh, "An improved approach to open set text-independent speaker identification (OSTI-SI)," in 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2017, pp. 51–56. doi: 10.1109/ICRCICN.2017.8234480.