Fayoum University Faculty of Engineering, 2025, Vol: 8(1), 143-151
https://fuje.journals.ekb.eg/
ISSN Online: 2537-0634
ISSN Print: 2537-0626

Deep Learning and Fourier Transform for Speaker Recognition(DLFSR)

Taqwa M. Sayed[1,*], Amr M. Gody[2], Sayed T. Muhammad[3]

1 Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum, Egypt
2 Professor of Digital Signals, Faculty of Engineering, Fayoum University, Fayoum ,Egypt
3 Assistant Professor, Computers and Systems Engineering Department, Faculty of Engineering, Fayoum University, Fayoum ,Egypt
*Corresponding author: Taqwa M. Sayed (taqwamahmoud92@gmail.com).

How to cite this paper: Sayed, T.M., Gody, A.M., and Muhammad, S. T. (2025) Deep Learning and Fourier Transform for Speaker Recognition (DLFSR). *Fayoum University Journal of Engineering,* Vol: 8(1), 143-151.
https://dx.doi.org/10.21608/fuje.2024.313518.1090

Abstract

Automatic Speaker recognition (ASR) and verification have gained increased visibility and significance in society as speech technology. Speaker recognition has undergone a revolution due to deep learning techniques, specifically deep neural networks (DNNs). With the use of models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), it is possible to learn discriminative features directly from unprocessed speech signals without the requirement for manual feature extraction. A growing number of people are using end-to-end speaker recognition models because of how well they work and how easily they can link speaker IDs to speech waveforms. It can recognize and authenticate people based on their distinct vocal traits. A lot of Applications of automatic speaker recognition can be found in many areas, such as voice-based digital device authentication, forensic analysis of audio recordings, access control, and phone-based customer support identification. Through our study, we introduce a Deep Learning and Fourier Transform for Speaker Recognition model (DLFSR) that based on Short Term Fourier Transform (STFT) in which the input speech can be transformed into spectrogram then we apply deep learning especially Convolutional Neural Network (CNN) to the spectrogram images to extract feature and classify the spoken person. The training and validation test are applied on speaker recognition dataset 16000pcm. This model performs excellent result with 98.8% correct identification and classification.

Keywords

ASR, STFT, CNN, RNN, DLFSR, pcm dataset.

1. Introduction

Speaker recognition is a form of biometric technology that used to verify a user's identity by identifying particular traits in their voice utterances. Utilizing the speaker's speech utterances, the speaker recognition system establishes the speaker's identity and manages access to services like voice calling, voice mail, security control, etc. A typical speaker recognition system analyzes a speaker voice or speech features to determine how unique they are. The most sensible way to change people's perspectives is through voice or words. In the past 60 years, ASR systems have developed thanks to the development of human-computer research tools. These sophisticated technologies are now utilized in a variety of settings, including voice dialing, online banking, phone purchasing, security control, and forensic applications. The four primary fields of speech recognition research—speaker verification, identification, diarization, and robust speaker recognition—are covered up in this paper. Figure (1) shows the flowcharts of the four branches, while difficulties with noise and domain mismatch are addressed by robust speaker recognition [1].

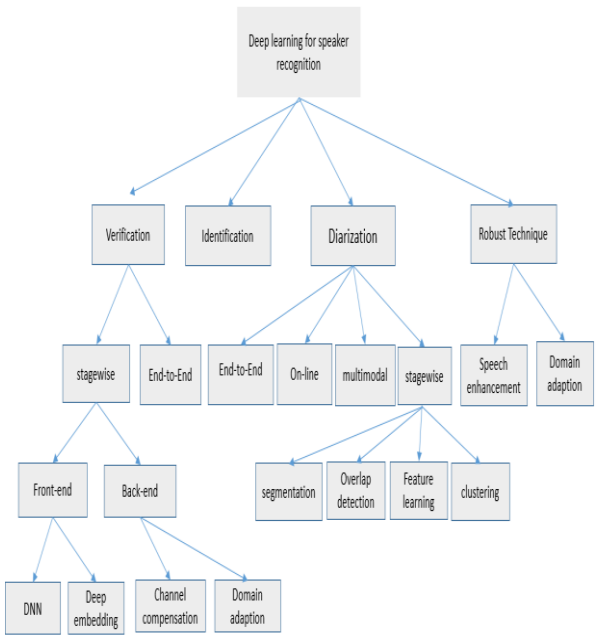

Figure 1. Speaker recognition using deep learning research areas.

Speaker recognition Verification is a method of confirming someone's identity based on the features of their speech. To establish if two voices are identical, the voice of one person is compared to a voice sample that has been saved in the past. This technology is frequently utilized in access control, authentication procedures, and security systems. While the technique of identifying which registered speaker's voice corresponds with a certain audio sample is known as speaker identification. To determine the best match, this entails comparing the voice sample to a database of recognized voices. But the technique for dividing an audio file into pieces that represent various speakers. Without any prior information about the speakers, the objective is to automatically recognize and categorize the speech segments that belong to each unique speaker. This procedure is essential for jobs like conversation transcription, audio recording indexing, and multi-speaker recording information extraction. Finally, Speaker recognition diarization is Speaker recognition robust technique Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models are examples of deep learning techniques that have shown robust in speaker recognition tests. With its ability to acquire intricate patterns and representations straight from unprocessed audio data, these models are useful for managing fluctuations in speech and background noise.

2. Related Work

The interest in automatic speaker recognition is increasing significantly with the advent of artificial intelligence systems. Numerous studies have been carried out on different methods of classifying machine learning algorithms and speech recognition and classification using neural networks.

Suci Dwijayanti et al. [1] investigated a convolutional neural network (CNN), and a CNN-visual geometry group (CNN-VGG) algorithm was applied for recognizing the speakers. The system used fast fourier transform spectrogram in feature extraction method and The accuracy of the suggested approach was high and demonstrated that this architecture is capable of producing an appropriate speaker identification model, which is noticeably better than the accuracy of the Mel-frequency Cepstral coefficients method.

Mahendra Kumar Gourisaria [2] et al provided a comparison between two distinct audio datasets with comparable properties. It also focuses on extracting MFCCs and STFTs features from the audio signals and categorizing them into many categories using a variety of machine learning approaches. the precision, recall, specificity, and F1-score findings both be-

fore and after the noise removal process to examine the impact of the noise on the results. With accuracy rates of 91.41% and 91.27% on the corresponding datasets, this study demonstrated that the ANN model performs better than the other six audio models.

Budiga et al. [3] presented with the use of a dataset of 16000 PCM voice samples and applying the Mel Frequency Cepstral Coefficient (MFCC), a convolutional neural network (CNN) classification algorithm, to extract speech features. The system has many hurdles when it comes to speaker recognition under various recording situations. It was 92.8% accurate in recognition now. The MFCC-CNN model that was provided with filtering techniques produced better results.

Karthikeyan et al. [4] introduced a jump-connected 1-D CNN with a mixed loss function for speaker recognition. To extract speaker-specific properties, the proposed model combines a 1-D convolutional layer with jump connections; this minimizes time-based and frequency-based variability for quicker processing. The suggested compact convolutional neural networks (CCNN) are guided to recognize the proper spokesman with increased efficacy by a combined softmax loss, smooth L1-norm, and steady L2-norm loss function. With applying voiceprint identification model, the average speaker recognition rate is a remarkable 98.76%.

Mokgonyane et al. [5] presented the creation of an automated speaker recognition system that includes Sepedi home language speakers' classification and recognition. The WEKA data mining tool is used to train four classifier models: Random Forest (RF), K-Nearest Neighbors, MLPs, and Support Vector Machines (SVM). The optimal classifier model and its ideal hyper-parameters are found using Auto-WEKA. accuracy surpassing the state-of-the-art with an accuracy of 97%.

Pavan et al. **[6]** introduced a broad phoneme class specific deep neural network (DNN) based speech augmentation technique. To determine the probabilities of each class in each test frame, a classifier network is constructed. They tested out two broad phoneme classes (vowel and non-vowel) and four broad phoneme classes (vowel, stop, fricative, and nasal). The TIMIT **[7]** corpus of speech data, four SNR conditions, and nine noise types (five visible and four invisible) are used in the experiments. The accuracy percentage came out to 84.1%.
Gbaily et al. [8] presented a preprocessing classification strategy for automatic speech recognition. Four hybrid models are offered, the first hybrid model (FS-HMM-GM-MBT). The second hybrid model (FSHMM-GM-MFCC). The third hybrid model is Mel-scaled Best Tree Encoding (VS-HMM-GM-MBT). (VS-HMM-GM-MFCC) are the components of the fourth hybrid model. This study made use of a subset of the TIMIT database. VS-HMM-GM-MBT achieves the highest overall recognition rate (81.01%).
Wei Han et al**.** [9] investigated a brand-new Context Net CNN-RNN-transducer architecture. Context Net has a fully convolutional encoder that adds squeeze-and-excitation modules to convolution layers to incorporate global context information. Furthermore, they suggest a straightforward scaling technique for Context Net widths that strikes a fair balance between computation and accuracy. They show that Context Net obtains a word error rate (WER) of 2.1%/4.6% on the popular Librispeech benchmark without the need for an external language model (LM).

Chao-Han Huck Yang et al. [10]proposed a system that is based on an end-to-end acoustic model (AM) based on recurrent neural networks (RNNs) and a quantum convolutional neural network (QCNN) made up of a quantum circuit encoder for feature extraction. The suggested QCNN encoder outperforms earlier designs that used centralized RNN models with convolutional features, achieving a competitive accuracy of 95.12% in a decentralized model during testing on the Google Speech Commands Dataset.

Ayad Alsobhani et al. [11] suggested a model that use deep convolutional neuro-learning to apply speech recognition features to the building of a word-tracking model. There are six control words: left, right, forward, backward, and stop. remarks made by individuals of various ages. Our voice dataset is contributed by an equal number of men and women, and it is used to train and evaluate proposed deep neural networks. The suggested deep neural network produced a word classification accuracy of 97.06%.

Daria Vazhenina et al. [12] explored the impact of merging spectrogram features from the Short-Time Fourier Transform (STFT) and Hilbert–Huang transform (HHT) with regard to an end-to-end DNN model-based ASR system. Additionally, they contrasted Empirical Mode Decomposition (EMD) and Variational Mode Decomposition (VMD), two mode decomposition techniques used in the HHT spectrum computation. They suggested adding many attention-based combination layers to the model, placing them in between the recurrent and convolutional stacks. Their findings demonstrate that the ASR system noise robustness can be enhanced by combining features without increasing the number of model parameters. The increase in parameters in the suggested attention-based combined feature models varies, with the AV-CNN model having a parameter increase of 13.1% and the SA-CNN model of 0.45%.

3. Motivation

There is still a need for more effective time-frequency representations that retain both temporal and spectral features, even though speaker recognition has been investigated using a variety of feature extraction techniques, including Mel-Frequency Cepstral Coefficients (MFCCs) and deep learning. For real-time high-resolution spectrum analysis, particularly in noisy or dynamic situations, STFT presents a promising, if understudied, approach. Still, there has been little investigation into how STFT might enhance resilience in tasks like speaker verification in challenging environments and short utterance identification.

It is this need for a feature extraction method that strikes a balance between precise spectrum representation and computational ease that motivates us to use STFT. In order to capture speaker-specific characteristics even in brief utterances, STFT offers a more detailed analysis of speech signals. Real-world applications where recordings may be brief, noisy, or contain overlapping voices—like speaker identification in call centers or forensic analysis—make this more important. We hope that our work will help address these issues, which are still not sufficiently addressed by current methods, by utilizing the advantages of STFT.

4. Fast Fourier Transform

The Fast Fourier Transform (FFT) is a common transform used for speech signal analysis. For a voice signal, the frequency domain standard representation is provided via FFT. Though time frequency variations can be held using the Short Fourier Transform. Since the FFT implies that signals are fixed in nature, it has the problem of not being suited for signals whose frequencies shift over time. Utilizing the frequency spectrum of the voice signal in place of the waveform enables for working in the frequency domain thanks to the FFT. It may be easier to identify the speakers when using the frequency domain since it offers more details about the voice signal. Certain techniques then employ the voice signal obtained immediately through sampling for speech recognition [13].

To convert a signal between the time and frequency domains, one can apply the Fourier Transform mathematical transformation. It has reversible functionality, meaning it can switch between domains. The Fourier series coefficients are a discrete set of complex amplitudes that can be computed by applying the Fourier transform to a periodic function over time. A description of the time-frequency analysis technique is given:

$$\text{STFT}(t,\omega) = \int_{-\infty}^{\infty} \boldsymbol{f(\tau)w(\tau - t)e^{-i\omega\tau}d\tau}, \qquad (1)$$

Where f(τ) is considered as the input signal in the time domain, $\boldsymbol{w(\tau - t)}$ is identified as hamming window or sliding window function located at time t. The Fourier kernel for frequency ω is represented by the complex exponential $\boldsymbol{e^{-i\omega\tau}}$.

The purpose of Equation (1) is to intercept the original signal in the time domain by using the window function; the signal within the window is interpreted as a smooth signal. The spectrum at that precise moment is then obtained via FFT transformation of the local information. STFT is the continual application of Equation. (2), or shifting the window over, and t. the Discrete Fourier Transform case:

$$\text{STFT}(m,\omega) = \sum_{n=-\infty}^{\infty} \boldsymbol{f[n]w[n - m]e^{-i\omega n}} \qquad (2)$$

The variable m is discretized and ω variable is continuous [14]. This is the discretized time Fourier transform (DTFT), which is continuously in the frequency domain but discretized in the time domain.

Figure (2) demonstrate how Fourier Transform Can convert waveforms to spectrogram using Fourier Transform of Oscillating function.

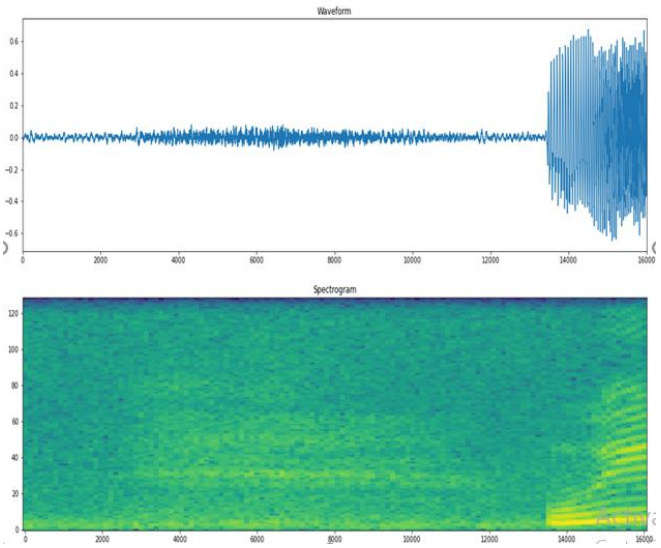

Figure 2. Waveforms to spectrogram conversion using Fourier Transform.

On the X-axis:
Label: Time (in seconds) illustrates how the speech signal has changed over time. Depending on how lengthy the signal under analysis is, the units are typically in milliseconds or seconds.

on the Y-axis:
Label: Frequency (Hz) indicates the frequency components that the STFT was able to extract from the signal. The units, which display the amount of energy present at particular frequencies across time, are Hertz (Hz).
The amount or power of the frequency components at each time step is displayed on the color scale, also known as intensity. Decibels (dB) are typically used to express magnitude, which makes tiny energy differences easier to see.
In recent times, the image recognition task has been much more effective because to the Convolution Neural Network (CNN). For speech recognition using spectrum images, many researchers employ CNN. Based on these techniques, a set of features that can characterize this signal is often generated.

Convolutional neural networks, or CNNs, generate filtered feature cards stacked preceding each other by performing the mathematical process known as convolution in place of general matrix multiplication. A CNN is consisting of six-layer Convolutional layer, pooling layer, ReLU (rectified linear unit), fully connected layer, Output layer, Softmax layer.

This research focuses on finding a solution to the classification challenge on speaker recognition dataset 16000 pcm applying Fourier transform and CNN. Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Thatcher, and Nelson Mandela are the five well-known leaders whose speeches are included in this dataset. The folder names are likewise represented in this list. There is a 16000 sample rate PCM encoded audio file in each folder, each lasting one second. For easier workability, I divided each speaker's speech into a one-second segment from its original, lengthy audio. Each speaker's entire speech can be formed by combining the chunked audio files from 0.wav to 1500.wav.For easier workability, I divided each speaker's speech into a one-second segment from its original, lengthy audio. Each speaker's entire speech can be formed by combining the chunked audio files from 0.wav to 1500.wav.

5. Convolutional Neural Network

Convolutional neural network is an especially effective kind of deep learning model for visual data analysis, Medical image analysis, image classification, picture identification, and other visual data-related tasks all make extensive use of them.

The visual sense served as inspiration for CNN architecture. Artificial neurons are equivalent to biological neurons; CNN kernels are receptors that can react to diverse aspects; activation functions mimic the process by which neural electric signals can only pass to the next neuron when they surpass a specific threshold. In order to train the CNN system as a whole to learn what is needed, people devised loss functions and optimizers. Specifically, four parts are usually required to construct the CNN model, convolution layer, pooling layer, activation layer and fully connected layer [11]as dedicated in Figure (3).

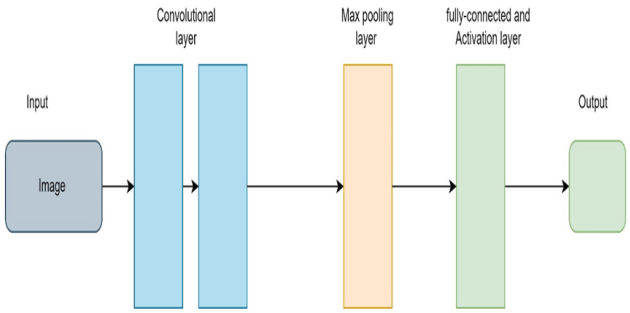

Figure 3. Convolution neural network layer diagram.

5.1. Convolution Layer

This layer is a fundamental structural component in CNN. Convolutional operations are applied to the input data. Convolution is the process of creating a feature map by swiping a tiny filter, sometimes referred to as a kernel, over the input data and carrying out element-wise multiplication and summing. Convolution is used to capture spatial hierarchy by identifying local patterns and features. The output of this layer is then computed as the result of an element-by-element convolution of the filters and receptive field of the input. A weighted summation is incorporated into the subsequent layer. According to Figure (4), the concentrated region (left matrix), represented by the colors blue and red as its center, is multiplied by the filter matrix (middle). The multiplication result will be kept in the layer below at the location that corresponds to the center of focus.

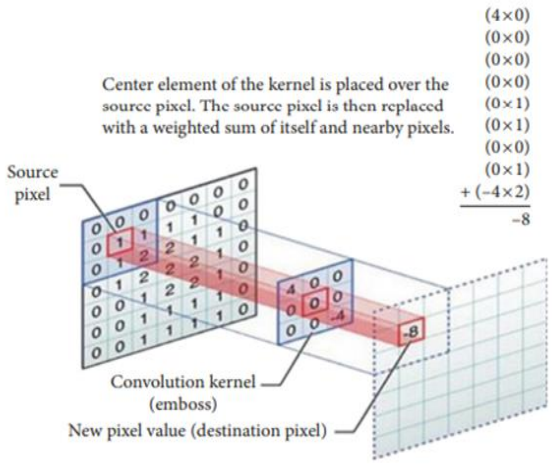

Figure 4. Sliding the filter over an input by the convolution layer [15].

4.2. Pooling Layer

Pooling layers are typically inserted after each convolution layer in CNN construction in order to decrease the representation spatial size. By lowering the parameter counts, this layer lowers the computational complexity. Furthermore, pooling layer aids in addressing the overfitting issue. By choosing the maximum, average, or sum values contained inside these pixels, the number of parameters can be limited by choosing a pooling size. The max pooling and average pooling operation are indicated in Figure (5) [13].

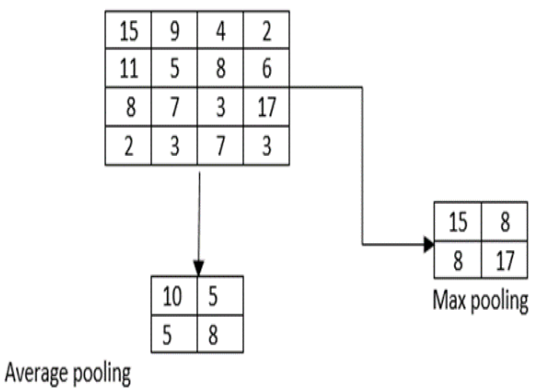

Figure 5. The operations of maximum and average pooling.

4.3. Fully-connected Layer

In an artificial neural network, a fully-connected layer, sometimes called the dense layer. It is one in which every node or neuron is connected to every other layer node and every layer neuron. To put it another way, every neuron output in one layer influences every other layer neuron input. The neural network can identify complicated patterns and relationships in the data thanks to its fully-connected architecture.

Nevertheless, it also adds more parameters to the model, increasing its computing cost and increasing the likelihood of overfitting, particularly with high-dimensional data. Fully-connected layers are a common neural network design for building element found throughout various deep learning frameworks and tools. They frequently work in tandem with rectified linear units (ReLU), tanh, and sigmoid activation functions to add non-linearity to the model.

4.4. Activation Layer

Typically, the activation function applied to the final completely connected layer differs from that of the previous layers. For every task, the appropriate activation function must be chosen. The Softmax function is an activation function used in the multiclass classification job. It normalizes the output real values from the last fully connected layer to the target class probabilities, where all values add up to one and each value ranges from 0 to 1 [14]. Figure (6) illustrates different type of activation function.

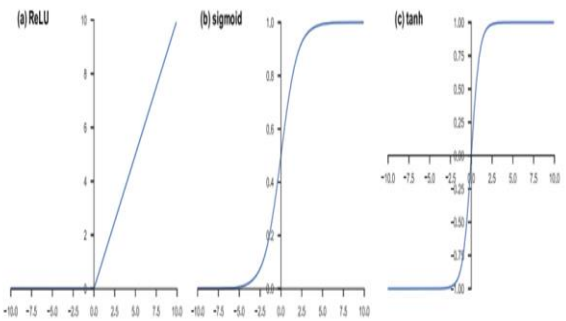

Figure 6. Activation function types.

5. Proposed DLFSR Model

In the field of digital signal processing, speech recognition is typically accomplished by employing a vector of characteristics that simply and clearly characterize speech. Consequently, as per Figure (7), speech recognition based on spectrogram images may be split into multiple components. The process of synthesizing speech signals dependent on the setting of fundamental parameters, such as the speech signal discrete frequency, bit depth, and channel count, is implemented in the first block. Initial speech signal processing, such as filtering, normalizing, and averaging, is used in the second block to eliminate speech noise and interruptions.

The third block produces spectrogram images from the speech signal processing's first processing and construct a data collection as described in forth block. In the last block, the procedure of feature extraction from the speech tone-created image spectrograms is carried out. CNN is used for feature extraction since it makes it possible to automatically retrieve crucial features for image learning. The best experimental findings are used to create a classification model utilizing the neural network architecture.

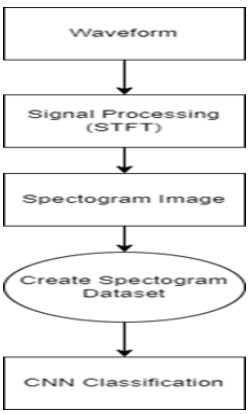

Figure 7. Procedure for proposed DLFSR model.

6. Experiment and Results

The proposed DLFSR classification model is formed through a series of procedures that included evaluating the implementation of the classification model, examining the dataset, and producing training and validation sets. As seen in Figure (8), the model is trained on the training set and tested on the validation set; the output accuracy falls among 97 and 98.8%. convergence between the trained and tested accuracy is illustrated in Figure (9).

the dataset was specifically designed for a configuration of 5 speakers. This focused approach was primarily due to resource constraints that limit our ability to process larger datasets, such as LibriSpeech or MGB-2. By concentrating on a smaller set of speakers, we were able to optimize the model's performance based on the unique characteristics within this group.

In response to your inquiry, we have updated the manuscript to explicitly mention this rationale for our dataset design. While we recognize that expanding the number of speakers could affect accuracy due to increased complexity and variability, our current design allows for more effective training and evaluation given the available resources.

The confusion can be clearly seen in Figure (10). The graph shows the actual label versus the expected label for the five speakers that are labeled on the x and y axes of the matrix. Determining each word's misclass and correctness is made easy using this graph.

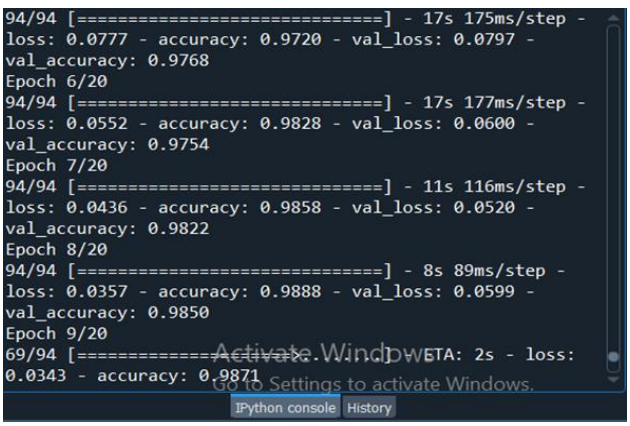

Figure 8. training the model on the spyder platform.

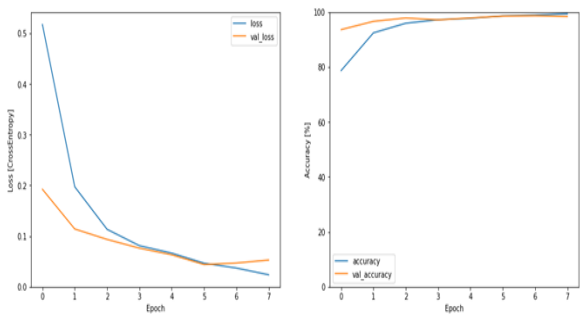

Figure 9. Differences in Training and Testing Losses with Dataset's Number of Epochs.

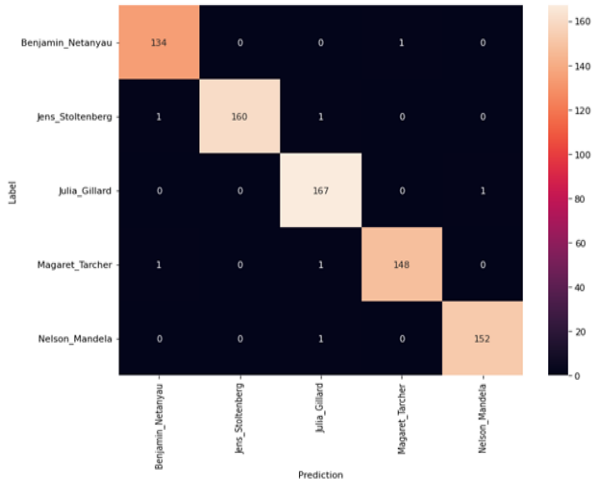

Figure 10. Confusion matrix diagram for DLFSR model

7. Conclusion

The proposed model offers a practical interpretation for getting over the convolutional neural network and deep learning extremely computational requirements. The CNN model and STFT were introduced in the planned research project for speaker recognition. On the 16000pcm speaker recognition dataset, the training and validation tests are running. The accuracy rate of this model's identification and classification is 98.8%. For real-time applications, the proposed model is appropriated. Additionally, depending on the analysis's significance and noise level, more dataset building is advised. Furthermore, experimenting with different deep learning and machine learning methods.

References

[1] S. Dwijayanti, A. Y. Putri, and B. Y. Suprapto, "Speaker Identification Using a Convolutional Neural Network," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 140–145, 2022, doi: 10.29207/resti.v6i1.3795.

[2] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," *Discov. Internet Things*, vol. 4, no. 1, 2024, doi: 10.1007/s43926-023-00049-y.

[3] P. Budiga, B. Bhavana, G. Gunisetty, N. D. Moka, and G. V. S. Reddy, "CNN trained speaker recognition system in electric vehicles," in *2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON)*, IEEE, 2022, pp. 1–6.

[4] V. Karthikeyan and S. S. Priyadharsini, "Text-independent voiceprint recognition via compact embedding of dilated deep convolutional neural networks," *Comput. Electr. Eng.*, vol. 118, p. 109408, 2024.

[5] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, IEEE, 2019, pp. 141–146.

[6] P. Karjol and P. K. Ghosh, "Broad phoneme class specific deep neural network based speech enhancement," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, IEEE, 2018, pp. 372–376.

[7] S. Language and S. Group, "SPEECH D A T A B A S E D E V E L O P M E N T AT MIT : TIMIT A N D B E Y O N D Victor ZUE , Stephanie SENEFF , and James GLASS Spoken Language Systems Group , Laborato ; qv for Computer Science . Massachusetts lnstimte of Technology . Cambridge , (arising," vol. 9, pp. 351–356, 1990.

[8] M. O. Gbaily, "Automatic Database Segmentation using Hybrid Spectrum-Visual Approach," *Egypt. J. Lang. Eng.*, vol. 8, no. 2, pp. 28–43, 2021.

[9] W. Han *et al.*, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv Prepr. arXiv2005.03191*, 2020.

[10] C.-H. H. Yang *et al.*, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6523–6527.

[11] A. Alsobhani, H. M. A. ALabboodi, and H. Mahdi, "Speech recognition using convolution deep neural networks," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 12166.

[12] D. Vazhenina and K. Markov, "End-to-end noisy speech recognition using fourier and hilbert spectrum features," *Electron.*, vol. 9, no. 7, pp. 1–18, 2020, doi: 10.3390/electronics9071157.

[13] V. K. Sarkania and V. K. Bhalla, "International Journal of Advanced Research in," *Android Internals*, vol. 3, no. 6, pp. 143–147, 2013, doi: 10.13140/RG.2.1.2722.0969.

[14] M. Musaev, I. Khujayorov, and M. Ochilov, "Image Approach to Speech Recognition on CNN," *ACM Int. Conf. Proceeding Ser.*, no. June, 2019, doi: 10.1145/3386164.3389100.

[15] S. Albawi, O. Bayat, S. Al-Azawi, and O. N. Ucan, "Social touch gesture recognition using convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2018, 2018, doi: 10.1155/2018/6973103.