How to Cite:
Banuroopa, K., & Priyaa, D. S. (2022). A novel voiceprint using ensembled Mel-Chromagram for speaker recognition. *International Journal of Health Sciences*, *6*(S4), 8043–8056. https://doi.org/10.53730/ijhs.v6nS4.10404

A novel voiceprint using ensembled Mel-Chromagram for speaker recognition

K. Banuroopa
Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education

Dr. D. Shanmuga Priyaa
Professor, Department of Computer Science, Karpagam Academy of Higher Education

***Abstract*---**This research paper proposes a novel voiceprint generation methodology for recognizing the speakers registered in a system. The proposed methodology is a keyword-dependent closed set speaker classification task. The features used are Mel-Spectrogram, Chromagram, MFCC and a new ensembled feature called Mel-Chroma. Mel-Chroma is generated with the combination of Mel-spectrogram and Chromagram. The Mel-Chroma spectrogram generated is converted into a binary image by using the average as the threshold. The recurrent neural network model LSTM is used for the classification task and the dataset used is FSDD. The proposed method has a higher accuracy than the state-of-art methods for the specific task. The accuracy obtained for the classification of speakers using a binary Mel-Chroma voiceprint is 98.33%.

***Keywords*---**Mel-spectrogram, MFCC, chromagram, LSTM, keyword-dependent speaker recognition, voiceprint.

Introduction

A series or sequence of sounds produced is stored as an audio file. All sounds have a specific frequency with which it could be identified and accessed. The sound is usually in a sine wave form and the intensity changes over time. There are numerous real-world data science applications where audio data is involved. A few of the prominent applications are recommending audio in radio, segmentation of audio and classification of audio. the process of listening, analysing and classifying audio recordings is called as audio classification. Classification of audio has wide application in the speech to text translation, environment sound recognition, classification of animal or bird sounds, music, instruments or song identification, automatic speech recognition, automatic

speaker identification and even virtual assistants in the smart homes and smart phones.

Speaker recognition can be split into speaker identification and speaker verification process. In speaker identification, a speaker is matched with the known speakers in the system; if there is a match, the unknown speaker has been identified to be in the system of known speakers. The exact match of which speaker has spoken is not done in speaker identification. But in speaker verification process, an unknown speaker is matched with known speakers in the system and the identity of the speaker is determined.[1] This is a multi-class classification problem.

Voice of human beings is used as a biometric in the recent years. The vocal cords act as resonators and it is said that two vocal cords resonating at the same frequency is extremely remote. This feature is taken for creating a voice print for the speakers and successfully matching them. [2] The speaker recognition can also be classified as keyword-dependent or keyword independent. In keyword-dependent models, the speaker has to speak the same keyword all the time and this is most suitable for biometric identification process. The keyword-independent systems model the voice of the speaker using phonemes.

Even though many features of the audio file have been used for the task, MFCC (Mel-Frequency Cepstral Coefficients) is treated as main feature in most speaker recognition models. The reason MFCC is used in many systems is that they are extracted over a triangle filter which model the human hearing model. In recent years the audio classification is modelled as a computer vision problem by producing the spectrograms of the audio files and fed into Convolutional Neural networks or Recurrent Neural networks for classification and identification.

A Chromagram specifies the pitches of the voice input audio over time [3]. The pitch classes are denoted by the seven letters A, B, C, D, E, F and G. It takes a Short-time Fourier transform (STFT) or any spectrogram as its input instead of Fast-Fourier Transform (FFT). In this paper, an ensembled features of Mel-Spectrogram and Chromagram of audio of the keyword spoken by the speaker is generated and used to create a voice print of the speaker for classification using Long Short Term Memory networks (LSTM).

Literature Review

Human's voice is one of the behavioural biometrics that gives the data in relevance to an individual's personalities, such as the ethnicity, age, gender, and emotion. Speaker recognition is recognizing and verifying the identity of speakers depending upon the features extracted from their voice. [4]. Machine learning algorithms like Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbours (KNN) are used for speaker recognition task in [5] and have produced remarkable accuracy using the MFCC vectors. The RF algorithm produced the highest accuracy. The CNN is used along with restricted Boltzmann machine based on transfer learning model for voice print generation and classification for a limited dataset [6] and achieved an accuracy of 97% for the classification.

In [7] the authors have proposed a new CNN architecture for audio recognition through Mel-Spectrograms and achieved an accuracy of 88.9%. The Gaussian Mixture Model - Universal Background Model (GMM-UBM) framework has been used in [8] for speaker recognition of a closed-set speakers and achieved a remarkable accuracy over other methods compared in the proposed model for Romanian speakers. The Chromagram features were used for speech / music classification with SVM algorithm in [9]. The experiments resulted in 95.5% accuracy with Chromagram textural features with 12 bins denoting the pitch classes. Hybrid features of Gamma Tone Cepstral Coefficients and formants [10] were used to generate a voiceprint for speaker recognition. The voiceprint was implemented using Deep Belief Networks and achieved state-of-art accuracy for camouflaged voices.

Voiceprint was created using MFCC and the classification was done with Back Propagation neural networks (BP) and CNN in [11]. The authors achieved 74.5% accuracy for isolated word recognition by the speakers. Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) was used in speaker recognition using the MFCC and Mel-Spectrogram as features in [12]. The MFCC features produced 95.33 % and for log-Spectrogram the accuracy was 98.7% for text independent speaker recognition. The Local Binary Patterns (LBP) operator is used for feature extraction from spectrogram images and the speakers were classified with the spoken isolated words using Deep Neural Networks in [13]. The overall accuracy achieved is 91% for Chinese speakers.

In [14], the authors have used the spectrograms for classifying the speakers using CNN and obtained an accuracy of 97%. They have used clustering techniques for feature selection from the spectrograms. The Convolutional Neural networks (CNN) is vastly used in image classification and is therefore used in classifying spectrogram images. The CNN is used along with GMM for speaker recognition in [15]; the feature used for classification is MFCC. The authors also tested the dataset for i-vector and produced remarkable results for both male and female speakers in the dataset used.

In [16], the authors used MFCC and raw wave form for creating a voiceprint and recognized the speakers with 96% accuracy in both the methods. The classification was done through Convolutional Neural Networks (CNN). Same model is used for MFCC and raw wave form for classification. The model worked well for datasets with or without noise and produced the same accuracy for both features. In [17] a new LSTM model was proposed for text dependent speaker verification task, which used the d-vectors generated from the audio file for the classification task and achieved an accuracy of 96.9% for the proposed model.In their proposed model, the authors of [18] test the classification of both text-dependent and text-independent scenarios for speaker classification based on CNN. They obtain a higher accuracy for text-dependent speaker classification rather than text-independent speaker classification.

The usage of short utterances of a speaker for speaker recognition was explored in [19], [20],[21], [22] with various features like MFCC, Mel-Spectrogram with traditional machine learning algorithms were tested and found that it gives remarkable accuracy in speaker classification even with short utterances. The

authors in [23] explored the usage of DNN for speaker recognition with i-vector and achieved an accuracy of 96.35% which is at par with state-of-art systems. The Deep GRU based model for speaker identification was proposed in [24]. The feature used is Mel-Spectrogram and the model achieved an accuracy of 98%. Most of the models discussed above require huge volume of audio data to create a speaker model and perform feature extraction and classification. The challenge here is to create a methodology through which a speaker can be modelled with short utterances of the speech. The proposed methodology addresses this gap and further explores creation of voiceprints which requires less space for storage through spectrogram images and improves the accuracy of the classification.

Proposed Methodology

The proposed methodology uses the Mel-spectrogram and Chromagram as stacked features to identify a speaker in a pool of speakers already in the database. While the Mel-spectrogram has the details about the audio the Chromagram has the characteristics of the pitch of the speaker which is unique to every human being. The ensemble of these two features creates a unique voiceprint for a speaker for a particular audio file. The type of ASR explored here is the key-word dependent closed set ASR.

The dataset used consists of 6 speakers with each speaker having spoken the digits 0 to 9 fifty times. This dataset is usually used for recognizing the spoken digits. This dataset is used in this research to mimic the short-spoken digits as keywords and as a closed set for evaluation. From each speaker one spoken digit is taken for consideration. This makes that spoken digit as a keyword for that speaker like a password. The aim is to prove that a speaker can be identified even with a very short audio.

Features Used in the proposed model are:

1. Mel-Spectrogram
2. MFCC
3. Chromagram
4. Mel-Chroma(proposed ensembled feature)

Mel-Spectrogram

A Mel-Spectrogram is generated by applying Discrete Fourier Transform (DFT) on the frames of the audio signal, which produces the magnitude spectrum. The next step is applying the triangular filters of the Mel Scale and take log() of it, which then produces the Mel-Spectrogram. The Figure-1 shows the steps

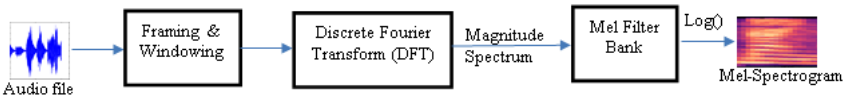


Figure-1 Generation of Mel-Spectrogram

The equation which is used to convert given frequency(freq) to Mel-Scale frequency (melfreq) is given below:

$$melfreq = 2595 * \log_{10}((freq/700) + 1)) \quad (1)$$

The Libarosa package of Python uses "librosa.feature.melspectrogram()" function to generate the Mel-Spectrogram.

MFCC

Mel Frequency Cepstral Coefficients (MFCC) are short representations of the frequency spectrum over the Mel-frequency filter banks. The process of obtaining the MFCC is shown in Figure-2. The Fast Fourier Transform is applied over the frames of the audio signal. The MFCC are obtained by applying Discrete Cosine Transform (DCT) over the triangular Mel-scale frequency filter applied signal. The DCT decorrelates the highly correlated Mel Frequency Spectrum to obtain MFCC. MFCC is said to be the most used audio feature in content-based audio retrieval and identification.

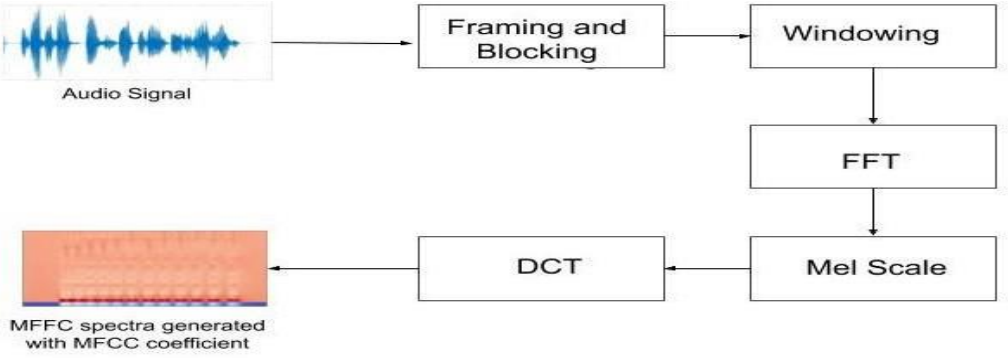


Figure-2 MFCC Generation

The MFCC coefficients can be made dynamic by getting delta and delta-delta derivatives of it. The first order will result in 13 coefficients and further each derivative would produce 13 more. The Figure-3 below shows MFCC generated for an audio file in the dataset with X-axis showing the time in seconds and Y-axis showing the Mel frequency.

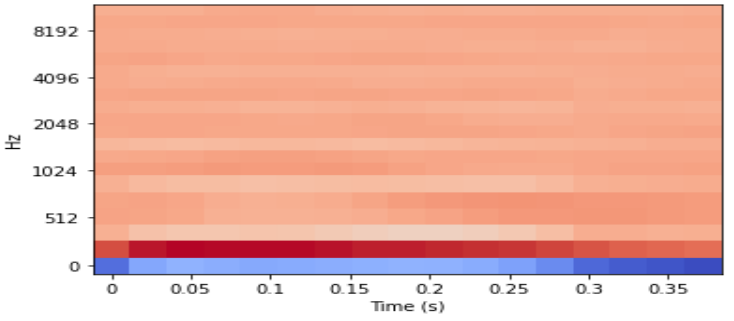


Figure- 3. MFCC

Chromagram

The pitch of the audio signal is embodied in the Chromagram. The pitch signifies fundamental frequency of vibration of the vocal cords of a speaker. Chromagram represents the whole spectral audio information mapped into one octave. Each octave is divided into 12 bins representing each one semitone. The Chromagram is generated from the audio file by generating its spectrum and then applying the Chroma filters on it. These filters project the frequency over time onto the 12 bins which signify the 12 octaves or pitch class. The Chromagram generated in Libarosa package is of dynamic nature as it uses Short Time Fourier Transform (STFT) to compute it rather than Fast Fourier Transform (FFT). The Figure-4 represents the Chromagram of an audio from the dataset.

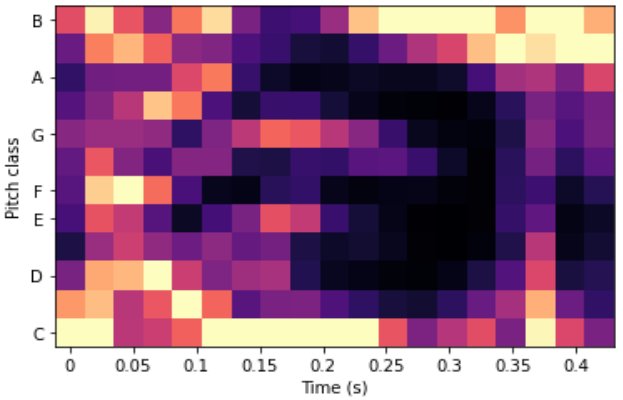
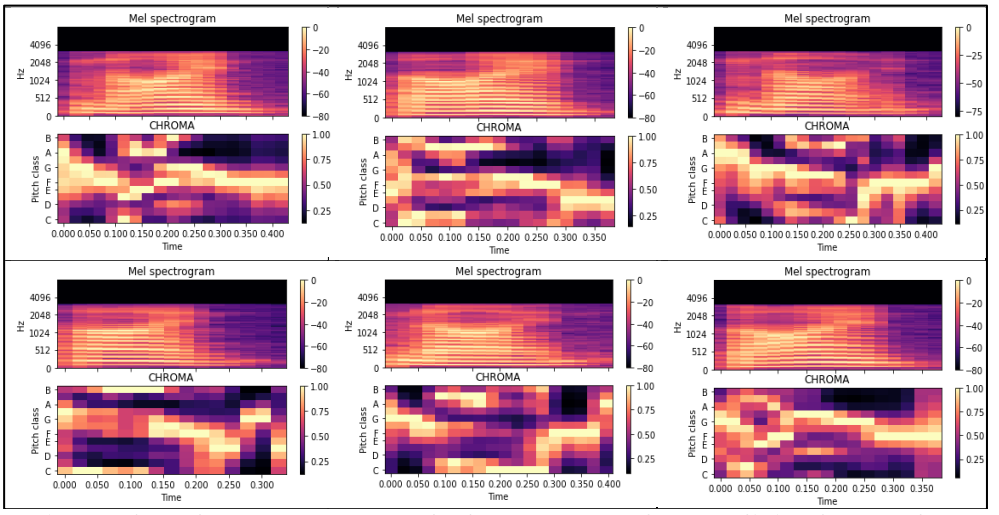


Figure-4. Chromagram

The focus of this work is on Chromagram which embodies the unique characteristics of speech apparatus of a speaker; the glottal source, vocal tract, nasal tract and lips. The glottal source is the space between the vocal folds which is responsible for the pitch and loudness of the voice and the vocal tract is the tube where the sound produced resonates through the throat and mouth. The Figure-5 shows the Mel-spectrogram and the Chromagram of the digit nine of the single speaker. From the Figure- it can be observed that there are slight changes in the Mel-Spectrogram and Chromagram even when the same speaker has spoken the same keyword. This will enable the recognition of the speaker even if he speaks the keyword little differently each time.



Fiqure-5. Mel-Spectrogram and Chromagram of same digit of a speaker

The Figure-6 shows the spoken digit zero by all the six speakers. It is evident from the Mel-Spectrogram and Chromagram displayed for the sample audio files that even though the same keyword or password is spoken by different speakers, the pitch, frequency of the audio file varies. Thus, the spectrogram images of the audio file can be unique for human for a specific keyword and can be used for speaker recognition task.

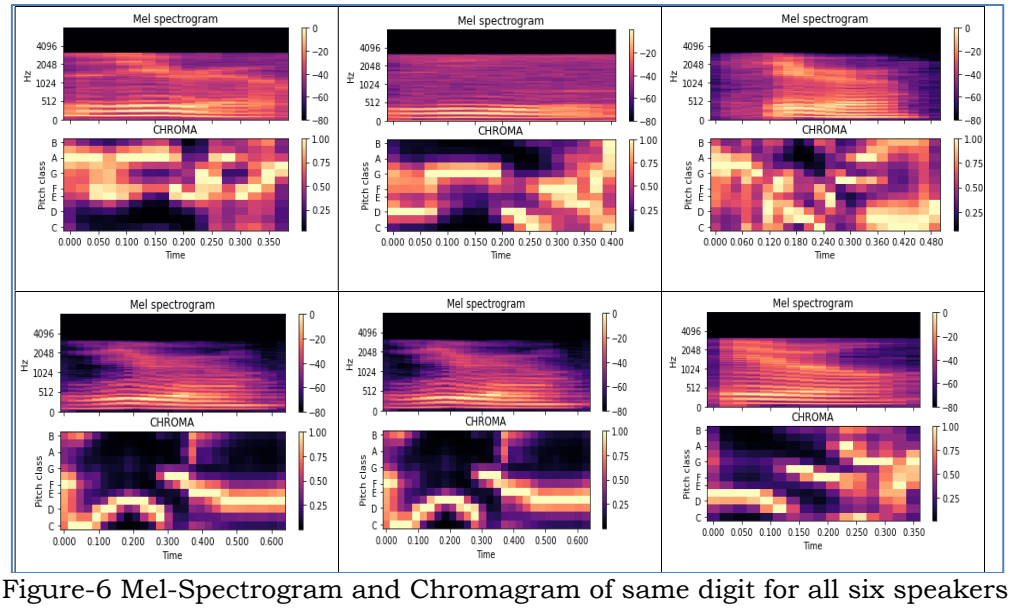


Figure-6 Mel-Spectrogram and Chromagram of same digit for all six speakers

A novel Chromagram called MEL-CHROMA is produced in this experimental setup by giving a Mel-Spectrogram as input to create a Chromagram. This novel ensemble of features of an audio file is then used as a voice print to classify the speakers in the database. The Figure-7 shows the Mel Spectrogram, Chromagram and Mel-Chroma of the same audio file.

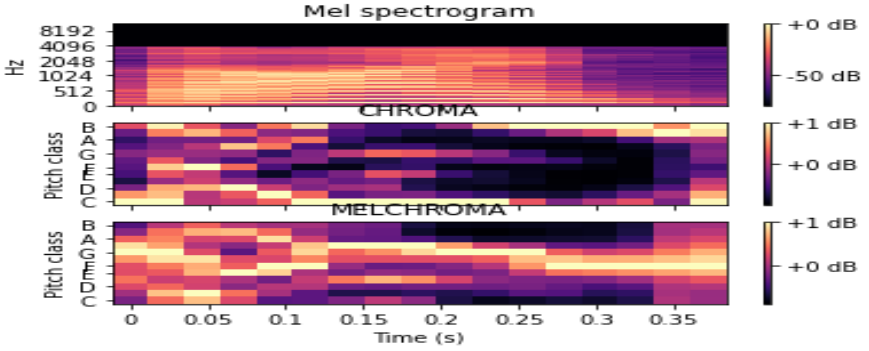


Figure- 7. Mel Spectrogram, Chromagram and Mel-Chroma of the same audio file

The audio files of digits spoken by speakers are taken one digit per speaker and converted into the corresponding voiceprint (ie) the ensembled or stacked feature of Mel-Chroma. The two most used feature for speaker recognition are MFCC and the Chromagram. Both are combined together in this experiment to produce a novel and unique voiceprint which is used in the classification of audio files through LSTM.

Global mean voiceprint

The spectrogram images are converted to binary images by calculating the mean or average of the spectrogram matrix and then changing the values to either 1 or 0 depending upon the threshold value. If the value in the spectrogram matrix is greater than or equal to the threshold value it is changed to 1 and if it is less than the threshold value then it is 0. In the Libarosa audio processing package of the Python, the values are true or false. Instead of a binary matrix of 0 and 1, a Boolean matrix of True and False are created. Then the Boolean matrix is displayed as a spectrogram in this case a Chromagram.

$$MC(x)=\begin{cases}1\ if\ x \geq AVG\\0\ if\ x < AVG\end{cases} \qquad (2)$$

Where MC is the MEL-CHROMA array and x is the value in the array. AVG denotes the average calculated, which acts as the threshold value for converting the MEL-CHROMA to a binary image. The Boolean images or the binary MEL-CHROMA images are calculated for 0.8* AVG, AVG, 1.2*AVG as threshold values and the sample of an audio files resultant images are displayed in the Figure- 8.

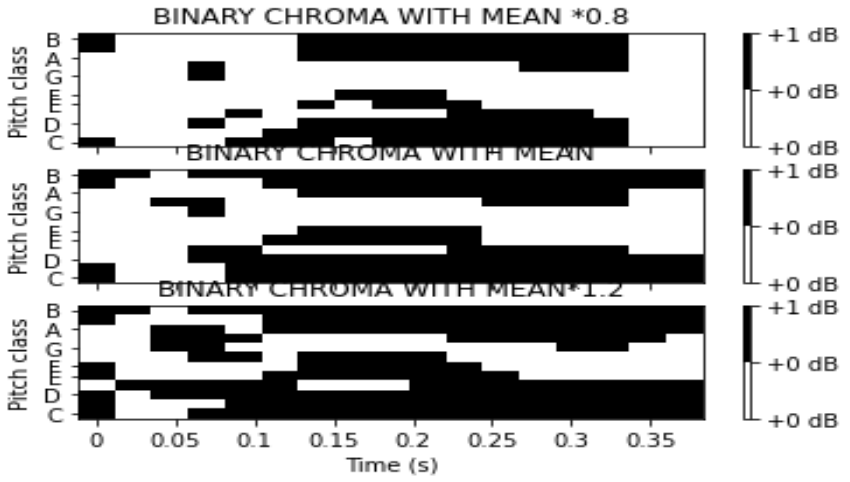


Figure-8. Binary MEL-CHROMA

The size of the MEL-CHROMA matrix is 12 x18. The code snippet for calculation of the average and converting the matrix back to a spectrogram is shown below:

```
pathAudio = "C:/Users/Dell/OneDrive/Desktop/FSDD/"
files = librosa.util.find_files(pathAudio, ext=['wav'])
for y in files:
data, sr = librosa.load(y)
ps = librosa.feature.melspectrogram(y=data , sr=sr)
melchroma = librosa.feature.chroma_stft(s=ps,sr=sr)
mcarr=np.array(melchroma)
avg=mcarr.mean()
bar=np.where(mcarr < avg, True, False)
librosa.display.specshow(bar,y_axis='chroma',x_axis='s')
```

Description of dataset used

The dataset used is Free Spoken Digits Dataset (FSDD). It is a publicly available open dataset from GITHUB. The dataset is usually used for classifying the spoken digits. This dataset is chosen for this task as it meets the requirement of keyword dependent closed set speaker recognition system. The audio files are in the .wav format. All the spoken digit speech audio files of the 6 speakers are sampled at 8KHz. The recorded audio files have minimum silence and noise. In the total number of speakers in this dataset is six. The language spoken is English. The speakers are all male and have different accents like American, French, Greek and German.

The dataset required for the proposed methodology is prepared as a subset from the 3000 audio files. First subset contains the 50 audio files of a single digit by an individual speaker. Unique digit is chosen for each speaker, thus making the total audio files as 300. The second subset is created by choosing a single for all the speakers. This subset also contains 300 audio files. The subsets mimic the speakers having unique keywords and all the speakers having the same keyword

respectively. The hypothesis is that even when the keyword is same for two speakers, they can be recognized through the characteristics of their voice rather than the content of the audio file.

Experimental results

The LSTM model used is a stacked LSTM layer model which was used in [26] and gave a good classification results for classifying environmental sound. The LSTM model called Deep LSTM has input layer connected to two stacked LSTM layers one after the another. The LSTM layers are followed by two Time Distributed Layers for stepping through the steps of time. This is followed by a flattening layer and a dense layer with activation functions Rectified Linear Unit (ReLU) and Softmax. Softmax is needed as it a multiclass classification problem. Categorical cross entropy is used as a loss function, for multiclass classifications and the metric used is accuracy.

The Table-1 below shows the accuracy of the speaker recognition through stacked LSTM networks. The features used for the experiment are Mel-Spectrogram, MFCC Cepstrum, Chromagram, Mel-Chromagram, binary Mel-Chromagram generated over the threshold values 0.8* AVG, AVG and 1.2*AVG, where AVG is the average or mean value of the Mel-Chromagram generated. The features are evaluated for two datasets Same-keyword for all speakers and unique-keyword for different speakers. From the table it is evident that Binary Mel-Chromagram with the threshold of AVG has the highest accuracy of 98.3 % and the Mel-Chromagram also has a remarkable accuracy of classification of 97.6% for unique keyword dataset. For the same-keyword dataset, the highest accuracy of 97.8% is produced again by the Binary Mel-Chromagram, but with the threshold of AVG *1.2.

Feature	Accuracy	
	UNIQUE KEYWORD	SAME KEYWORD
Melspectrogram	94.5 %	93.4 %
MFCC	97.3 %	97.2 %
Chromagram	93.7 %	92.9 %
MEL-CHROMA	97.6 %	96.2 %
Binary MEL-CHROMA with 0.8*AVG	97.2 %	96.7 %
Binary MEL-CHROMA with AVG	98.3 %	97.4 %
Binary MEL-CHROMA with 1.2*AVG	96.3 %	97.8 %

Table-1. Accuracy of speaker classification for various Features

The Table-2 shows the comparison of the state-of-art models and the proposed model. In this comparison it is observed that the proposed model has the better accuracy in classification than others. Further, LSTM-based models are showing a better performance than other models.

Model	Feature Used	Accuracy (%)
GMM-UBM [8]	MFCC	94.5
Multimodal LSTM [23]	i-vector	96.25
CNN [16]	MFCC	96

LSTM [17]	d-vector	96.9
Deep GRU [24]	Mel Spectrogram	91.56
Proposed Binary MEL-CHROMA with LSTM	Ensemble of Mel-Spectrogram & Chromagram	98.4

Table-2. Comparison of accuracy of speaker classification with proposed model

The chart below in Figure 9 Shows the comparison graph of accuracy values in Table -1

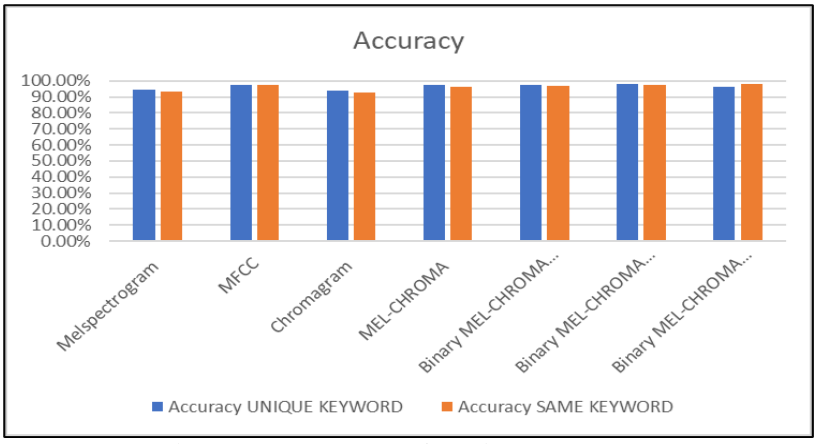


Figure-9 Accuracy of Various Features

The confusion matrix for the classification of the speakers with Binary MEL-CHROMA with the threshold of the average of the spectrogram is shown in Figure-10

	Speaker1	Speaker2	Speaker3	Speaker4	Speaker5	Speaker6
Speaker1	49	0	0	0	0	0
Speaker2	0	49	0	1	0	0
Speaker3	0	0	50	0	0	0
Speaker4	1	1	0	49	1	0
Speaker5	0	0	0	0	49	1
Speaker6	0	0	0	0	0	49

Figure-10. Confusion Matrix

Conclusion

The methodology proposed is a keyword-dependent closed set speaker classification task. The features used are Mel-Spectrogram, Chromagram, MFCC and a new ensembled feature called Mel-Chroma. Mel-Chroma is generated with the combination of Mel-spectrogram and Chromagram. The spectrogram generated is converted into a binary image by using the average as the threshold. The recurrent neural network model LSTM is used for the classification task and the dataset used is FSDD. The proposed method has a higher accuracy than the state-of-art methods for the specific task. The accuracy obtained for the classification of speakers using a binary Mel-Chroma voiceprint is 98.3% with threshold as Average for the different keywords and 97.8% with threshold as AVERAGE*1.2 than other methods used for voiceprint generation. In future, other DNN models can be explored for increasing the accuracy. The number of speakers in the dataset can be increased and noise can be added to mimic the real time implementations.

References

1. Banuroopa, K., & Shanmuga Priyaa, D. (2022). MFCC based hybrid fingerprinting method for audio classification through LSTM. International Journal of Nonlinear Analysis and Applications, 12(Special Issue), 2125-2136.
2. Birajdar, G.K., Patil, M.D. Speech/music classification using visual and spectral chromagram features. J Ambient Intell Human Comput 11, 329–347 (2020).
3. Birajdar, Gajanan K. and Mukesh D. Patil. “Speech/music classification using visual and spectral chromagram features.” Journal of Ambient Intelligence and Humanized Computing (2020): 1-19.
4. Dong Zhipeng et al 2019 Voiceprint recognition based on BP Neural Network and CNN. J. Phys.: Conf. Ser. 1237 032032
5. El-Moneim, S. A., Nassar, M. A., Dessouky, M. I., Ismail, N. A., El-Fishawy, A. S., El-Samie, A., & Fathi, E. (2020). Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimedia Tools and Applications, 79(33), 24013-24028.
6. El-Moneim, S.A., Sedik, A., Nassar, M.A. et al. Text-dependent and text-independent speaker recognition of reverberant speech based on CNN. Int J Speech Technol 24, 993–1006 (2021).
7. Georgescu, A.; Cucu, H. GMM-UBM Modeling for Speaker Recognition on a Romanian Large Speech Corpora. In Proceedings of the 2018 International Conference on Communications (COMM), Bucharest, Romania, 14–16 June 2018; pp. 547–551
8. Hourri, S., Nikolov, N.S. & Kharroubi, J. A deep learning approach to integrate convolutional neural networks in speaker recognition. Int J Speech Technol 23, 615–623 (2020).
9. Jiang, N., Liu, T. Research on Voiceprint Recognition of Camouflage Voice Based on Deep Belief Network. Int. J. Autom. Comput. 18, 947–962 (2021).
10. Li, X.K., Zheng, Y.L., et al. (2018) Research on voiceprint recognition method based on deep learning. Journal of engineering, Heilongjiang University, 009(001): 64-70.

11. Lu Y.N., Shan B.Y., Guan C. (2017) Current situation and application of voiceprint recognition technology. Information system engineering, 02:13-13.
12. Lukic, Y., Vogt, C. Durr, O., Stadelmann, T. (2016). Speaker identifcation and clustering using convolutional neural networks. In IEEE international workshop on machine learning for signal processing, Sept. 13–16, 2016.
13. Novoa, R. B. (2021). State of the art and future applications of digital health in Chile. *International Journal of Health & Medical Sciences*, *4*(3), 355-361. https://doi.org/10.31295/ijhms.v4n3.1772
14. P. Li, M. Chen, F. Hu and Y. Xu, "A spectrogram-based voiceprint recognition using deep neural network," *The 27th Chinese Control and Decision Conference (2015 CCDC)*, 2015, pp. 2923-2927, doi: 10.1109/CCDC.2015.7162425.
15. Rafizah Mohd Hanifa, Khalid Isa, Shamsul Mohamad. (2021). A review on speaker recognition: Technology and challenges, Computers & Electrical Engineering, Volume 90.
16. Shanthi, T. S., Lingam, C. (2014). Isolated word speech recognition system using htk. International Journal of Computer Science Engineering and Information Technology Research, 4(2), 81-86.
17. Sun, Cunwei, Yuxin Yang, Chang Wen, Kai Xie, and Fangqing Wen. 2018. "Voiceprint Identification for Limited Dataset Using the Deep Migration Hybrid Model Based on Transfer Learning" *Sensors* 18, no. 7: 2399.
18. 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," *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019, pp. 141-146
19. Tian, Y., Cai, M., et al. (2016) Speaker recognition system based on deep neural network and bottleneck feature. Journal of Tsinghua University (NATURAL SCIENCE EDITION), (11): 1143-1148.
20. Togneri R, Pullella D (2011) An overview of speaker identification: accuracy and robustness issues. IEEE Circuits and Systems Magazine 11:23–61
21. Widana, I.K., Sumetri, N.W., Sutapa, I.K., Suryasa, W. (2021). Anthropometric measures for better cardiovascular and musculoskeletal health. *Computer Applications in Engineering Education*, *29*(3), 550–561. https://doi.org/10.1002/cae.22202
22. Xin, W., Zhang, H.R. (2018) Robust i-vector speaker recognition method based on DNN processing. Computer Engineering & Applications.
23. Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, “Deep feature for text-dependent speaker verification,” Speech Communication, vol. 73, pp. 1–13, 2015.
24. Ye, F.; Yang, J. A Deep Neural Network Model for Speaker Identification. Appl. Sci. 2021, 11, 3603
25. Ye, Feng, and Jun Yang. 2021. "A Deep Neural Network Model for Speaker Identification" Applied Sciences 11, no. 8: 3603. https://doi.org/10.3390/app11083603.
26. Z. Liu, Z. Wu, T. Li, J. Li and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3244-3252, July 2018, doi: 10.1109/TII.2018.2799928.

27. Zeng, C.Y., Ma, C.F., et al. (2020) Robust speaker recognition method based on convolutional neural network. Journal of Huazhong University of science and Technology (NATURAL SCIENCE EDITION), 48(06): 39-44.