Multimedia Tools and Applications (2024) 83:31587–31601
https://doi.org/10.1007/s11042-023-16811-x

Real time implementation of voice based robust person authentication using T-F features and CNN

A. Revathi[1] · **N. Sasikaladevi**[2] · **N. Raju**[1]

Received: 7 April 2022 / Revised: 5 June 2023 / Accepted: 31 August 2023 /
Published online: 18 September 2023

Abstract
A forensic investigation uses personal traits to identify the persons involved in criminal offences. In this work on person authentication, the recorded voice samples can also be used to narrow down the search to identify persons. Time-frequency (T-F) features obtained from the concatenated training set of utterances are given to the convolutional neural networks (CNN), with layers configured for creating templates. Testing utterances are tied, and T-F features are derived. These features are applied to the CNN templates, and based on the match claimed, recognition accuracy is computed to validate the feature selection and CNN technique. Decision-level fusion of features with CNN for modelling and classification provides an overall authentication rate of 98%. This system is also implemented in real-time using Raspberry Pi hardware. This automated system would be helpful in identifying convicts in forensic sectors and perform secured online transactions against fraudulent attacks in financial sectors.

Keywords Forensic investigation · Spectrogram · CNN · Machine learning · Person authentication · Raspberry Pi hardware

1 Introduction

In the digital world, online banking using credit cards is often necessary, and authentication is initiated using a sequence of alphanumeric characters. It is often possible to duplicate the passwords as a sequence of characters, and impostors pose as genuine persons and try to loot the money. It cannot be replicated if a voice is used as a password. Suppose it is necessary to give actual users access to privileged information. In that case, voice can also be used as an additional biometric to check whether the clients are honest. Voice

As the authors of the manuscript, we do not have a direct financial relation with the commercial Identity mentioned in our paper that might lead to a conflict of interest for any of the authors.

✉ A. Revathi
revathi@ece.sastra.edu

[1] Department of ECE/SEEE, SASTRA Deemed University, Thanjavur, India

[2] Department of CSE/SEEE, SASTRA Deemed University, Thanjavur, India

biometric authentication could find applications in call or contact centres where authentication is done using the voice passwords of customers/clients. Speech is biometric for multi-level framework-based person authentication [4]. Voice password, Text-dependent and independent modules are integrated as a multi-level framework for speaker verification. MFCC features and dynamic time-warping models are used to authenticate speakers, and the performance is analyzed in terms of the time complexity of the testing procedures adopted for different modules in the work. An equal error rate (EER) is a performance metric for comparative analysis between the various modules. A person authentication system [15] is assessed against replay attacks using voice as a biometric and is applied for security services. An EER is used as a performance metric to compare the different modelling techniques used. Signature and speaker verification systems [14] are combined to authenticate speakers. False rejection and acceptance rates are performance metrics used in the work.

Voice-based biometric system [17] is developed using whispered speech to verify speakers. Weighted instantaneous frequencies are used as features, and their performance is compared with a basic MFCC-GMM-based system using EER as a performance metric. MFCC features and i-vector-based modelling [5] are used to authenticate speakers, and the recognition rate is the performance metric. Face, fingerprint, and speech are Bigun et al. [3] multimodal biometrics for person authentication applications, and EER is a performance metric. Audio and visual features are fused [6] to augment the speaker verification system's performance. Face and speech are combined [11] to analyze the performance of the speaker verification system using EER as a metric. Face and speech are combined [16] to assess the performance of the noisy speaker verification system using the different classifiers. Speech-based biometric system [12] is developed to authenticate speakers, and its performance is assessed against voice conversion attacks. The GMM-SVM-based approach performs better than the GMM-UBM approach in evaluating the speaker verification system. An MFCC-SVM-based speaker verification system [13] is implemented using FPGA hardware. Bio-inspired architecture powered by CM1K chip [18] is implemented for person authentication using speech and face as biometrics. The voice-based security system is implemented [8] using Raspberry Pi hardware.

1.1 Motivation of the work

It is necessary to develop a hardware-based speaker authentication system. A speaker authentication system using voice as a biometric has been implemented, and the related works are recently in practice. Voice passwords may be different or the same to authenticate speakers. For example, suppose the voice password is the same for all the speakers during testing. In that case, the systems are considered more robust against mimicry. Test speeches are transmitted through Raspberry Pi hardware for the identification of persons remotely.

1.2 Contribution to the work

This hardware-based work is preceded by simulation work, emphasizing feature extraction as a first stage. The extracted two-dimensional image-like features are given to the 2D-CNN for creating templates, and 20% of the test images are given to the models. Based on analyzing the classification indices, the test image is classified as associated with one of the speaker models. Recognition accuracy is taken as a metric to assess the system's performance. These codes are made compatible to be deployed in Raspberry Pi hardware.

If the test speech is given to the Raspberry Pi hardware, features are extracted and given to the CNN templates for the classification of a speaker remotely. This work highlights using the same utterances for training and testing to authenticate the speakers. Even if the test utterance is the same for speakers, it is possible to ascertain the speakers using speech as a biometric.

1.3 Organization of the work

This paper is organized in Section 2, which emphasizes the details of the database used, description of feature extraction procedures, techniques for developing CNN templates and results and discussion based on the application of testing procedures for authenticating persons. Section 3 demonstrates the real-time implementation of the Raspberry Pi-based person authentication system, and Section 4 summarizes the performance of the work on the person authentication system.

1.4 The proposed research framework

Figure 1 indicates the flow diagram of the proposed research work.

2 The implementation of a voice-based person authentication system

This voice-based person authentication is implemented by using voice prints of persons for training and testing. Voice prints of the words (Enter, Erase, Go, No, Help, Start, Stop, Repeat, Rugby, One, Two, Three, Four, Five, Six, Seven, Eight, Nine and Zero) spoken by the persons are commonly used for developing templates for each person. In addition, voice prints of the word "Yes" are used for testing/authenticating the persons.

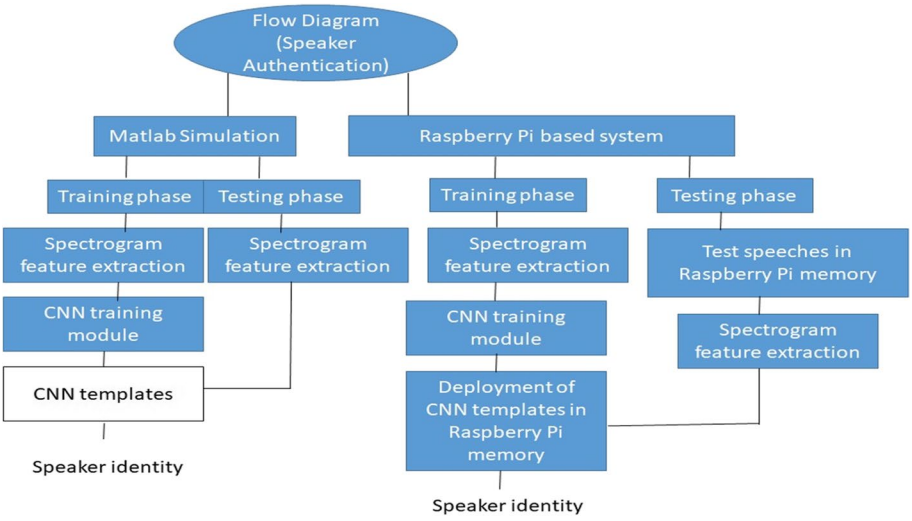

Fig. 1 Flow diagram – The Proposed Research

2.1 Details about the database used

The database used for the voice-based person authentication system is "TIMIT". It contains spoken utterances by eight female and eight male speakers at various trials for the isolated words and digits. This database could also perform speaker-dependent/independent isolated word recognition. Spoken utterances are present for isolated words: Enter, Erase, Go, No, Help, Start, Stop, Repeat, Rugby, Yes, One, Two, Three, Four, Five, Six, Seven, Eight, Nine and Zero. Out of the twenty isolated words, spoken utterances in nineteen words except "Yes" are used for developing training models/templates and spoken phrases in the word "Yes" are used uniformly for authenticating sixteen persons. The motivation behind this is that even if the spoken utterances of the words are the same for training and testing, it is possible to distinguish the persons using voice as a biometric. If other persons impersonating the genuine speakers speak the same word, this system earmarks the impostor against the claimed Identity. These utterances are saved in a database with 12500 as the sampling rate and 16 bits allocated to each voice sample.

2.2 Development of person authentication system

This voice-based person authentication system comprises training and test phases. During training, utterances corresponding to the speaker are grouped and taken as one data file. Two-dimensional T-F features are extracted from the training data and given to CNN layered architecture for creating templates for a person to perform person authentication. Test sets of utterances are grouped for a person, and T-F features are extracted. Features are applied to the CNN templates. According to the match between the test matrix and models, person classification is done, and each person's authentication rate is computed.

2.2.1 T-F Features extraction

This speech data as a vector for training is converted into frames of 8192 samples, with 12.5% of samples overlapping between the frames. Spectrograms with sub-bands in BARK, MEL and ERB scales are taken, and two-dimensional T-F features are extracted corresponding to the number of sub-bands and window size. Figure 2 indicates the blocks used for spectrogram and Melspectrogram T-F feature extraction. Sub-band filters are calibrated in non-linear BARK and MEL frequency scales. The conversion formulae used for HZ into BARK, MEL and ERB scale are given in (1),(2),(3).

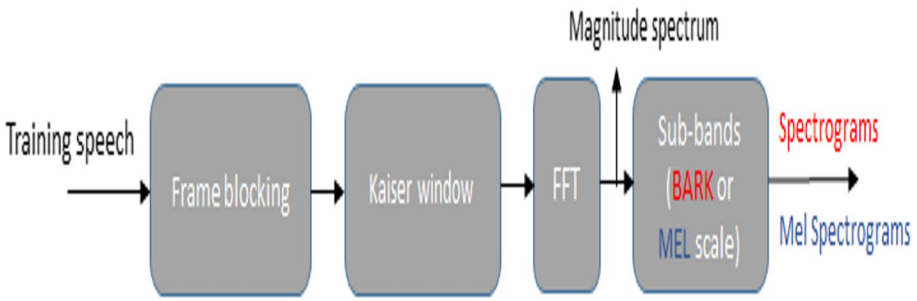

Fig. 2 Spectrogram and Mel spectrogram T-F feature extraction

$$f(Bark) = 6 * \sinh^{-1}\left(\frac{f(hz)}{600}\right) \tag{1}$$

$$f(Mel) = 2595 * \log10\left(1 + \frac{f(hz)}{700}\right) \tag{2}$$

$$f(erb) = 21.4 * \log10\left(4.37e^{-3*f(Hz)} + 1\right) \tag{3}$$

Short-time Fourier transform is computed for windowed speech segments as in (4)

$$X(n,k) = \sum_{n-N+1}^{n} x(m)w(n-m)e^{-jw_k m} \tag{4}$$

A spectrogram is computed by applying the squared magnitude of the short-time Fourier transform coefficients to the filter banks with filters calibrated in BARK, MEL and ERB scales as in (5)

$$s(n,l) = \sum_{k=0}^{L-1} |X(n,k)|^2 h_l(n-k) \tag{5}$$

Where l denotes the number of sub-bands.

Figure 3 depicts the modules used for Gammatonegram T-F feature extraction. Filters are Gammatone filters in Equivalent rectangular bandwidth (ERB) scale.

Figure 4 details the characteristics of the T-F spectrogram, Mel spectrogram and Gammatone gram features.

2.2.2 CNN Template creation

Speech-related applications [1], [19], [2], [21]] are developed by using CNN. Figure 5 depicts the CNN layered architecture for developing templates to authenticate 16 persons. T-F features from the training set of speeches to perform person authentication are given to the CNN structure as a two-dimensional image, and the structure is trained through an epoch-based iterative process.

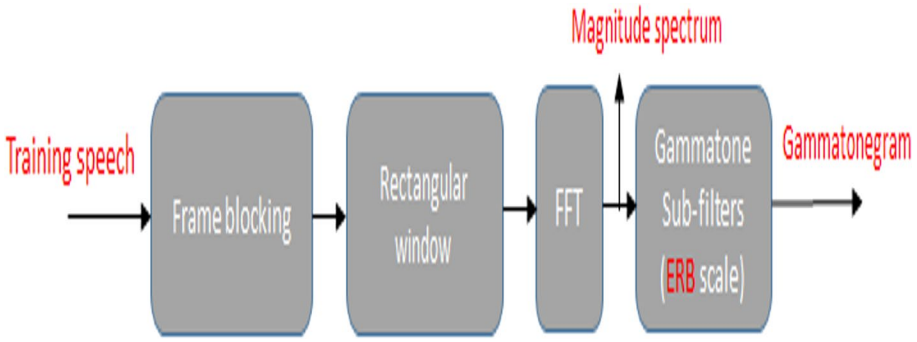

Fig. 3 Gammatonegram T-F feature extraction

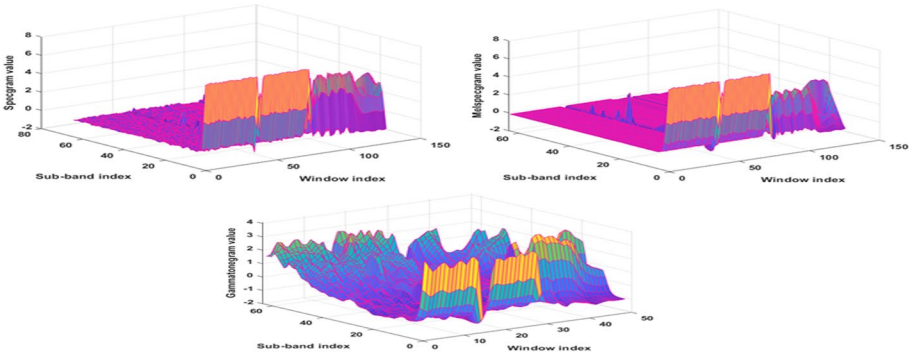

Fig. 4 T-F spectrogram, Mel spectrogram and Gammatonegram feature characteristics

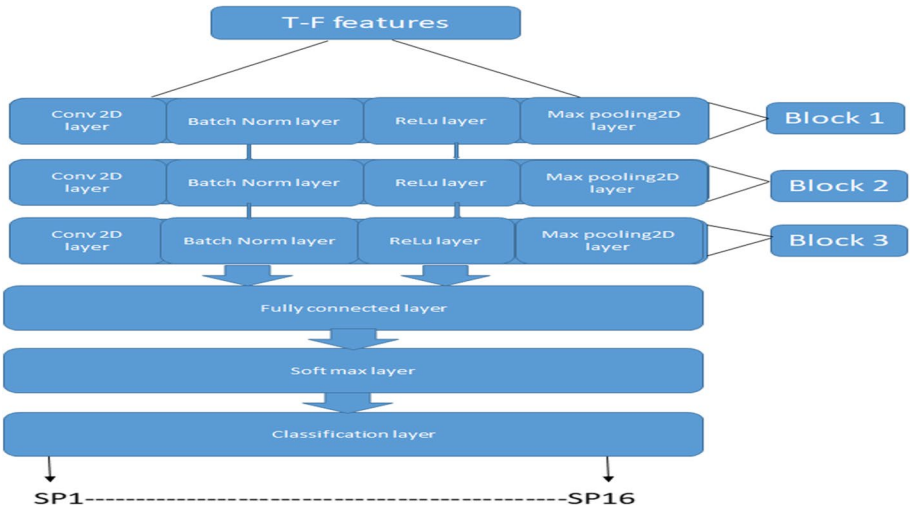

Fig. 5 CNN layered architecture – Person authentication system

Table 1 indicates the CNN layered structure used for the spectrogram and CNN- based speaker authentication system.

Figure 6 indicates the network architecture and trained CNN network details for T-F Gammatonegram features.

2.3 Results and discussion based on experimental analysis

This person authentication using T-F features and CNN is implemented by T-F feature extraction, CNN template creation and testing of test T-F features for performing authentication of persons by using speech as a biometric like other biometrics. Figure 7 indicates the performance assessment of the system for the spectrogram T-F feature. This figure illustrates that the 1-8 indices correspond to female and 9-16 to male speakers.

Table 1 CNN model – layered structure

```
net.Layers

ans =

  15×1 Layer array with layers:

     1   'imageinput'   Image Input          65×127×1 images with 'zerocenter' normalization
     2   'conv_1'       Convolution          7 3×3×1 convolutions with stride [1  1] and padding 'same'
     3   'batchnorm_1'  Batch Normalization  Batch normalization with 7 channels
     4   'relu_1'       ReLU                 ReLU
     5   'maxpool_1'    Max Pooling          2×2 max pooling with stride [2  2] and padding [0  0  0  0]
     6   'conv_2'       Convolution          14 7×7×7 convolutions with stride [1  1] and padding 'same'
     7   'batchnorm_2'  Batch Normalization  Batch normalization with 14 channels
     8   'relu_2'       ReLU                 ReLU
     9   'maxpool_2'    Max Pooling          7×7 max pooling with stride [7  7] and padding [0  0  0  0]
    10   'conv_3'       Convolution          28 3×3×14 convolutions with stride [1  1] and padding 'same'
    11   'batchnorm_3'  Batch Normalization  Batch normalization with 28 channels
    12   'relu_3'       ReLU                 ReLU
    13   'fc'           Fully Connected      16 fully connected layer
    14   'softmax'      Softmax              softmax
    15   'classoutput'  Classification Output  crossentropyex with '1' and 15 other classes
```

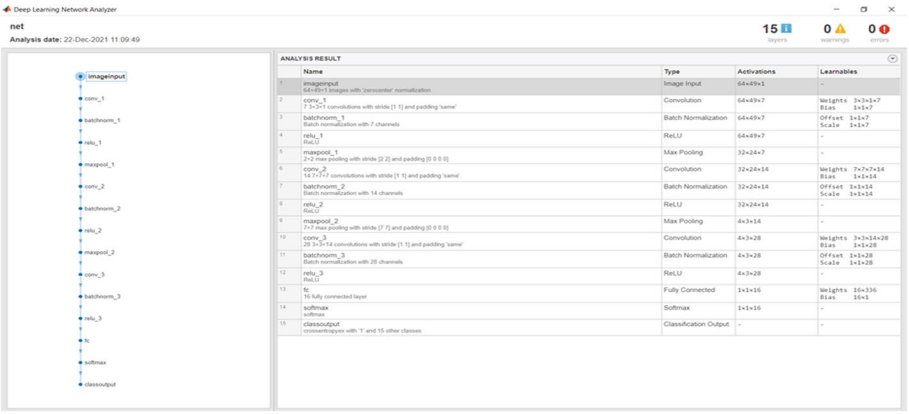

Fig. 6 CNN network structure – Analysis of the trained network

From Fig. 7, it is understood that the misclassification of a person is most probably among the female set or male set of speakers. The male's Speech features are misclassified with the male group of speakers, and the female's speech features are misclassified with the female speakers. Overall average accuracy is 88%. Figure 8 depicts the system's performance for Mel spectrogram T-F feature, and the overall average accuracy is 83%.

Figure 9 indicates the system's performance for the Gammatonegram T-F feature, and the overall accuracy is 84.5%.

Figure 10 indicates the overall comparative performance of the person authentication system for the individual T-F features with a decision-level fusion of features with CNN for modelling and authentication of persons.

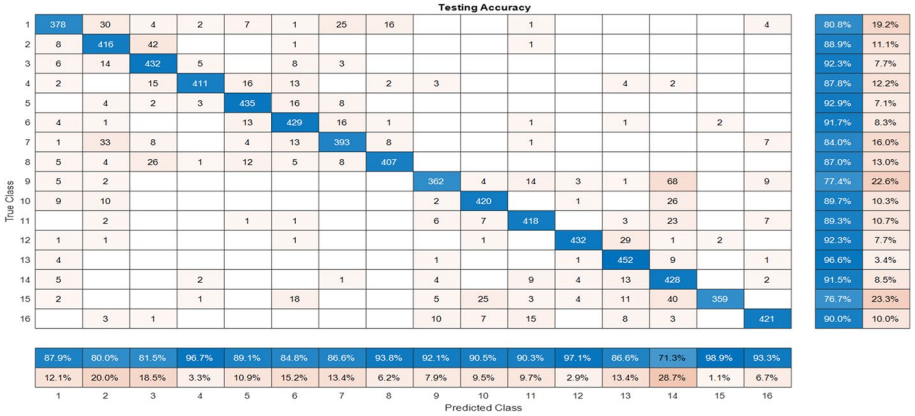

Fig. 7 Performance of the system- Spectrogram with CNN

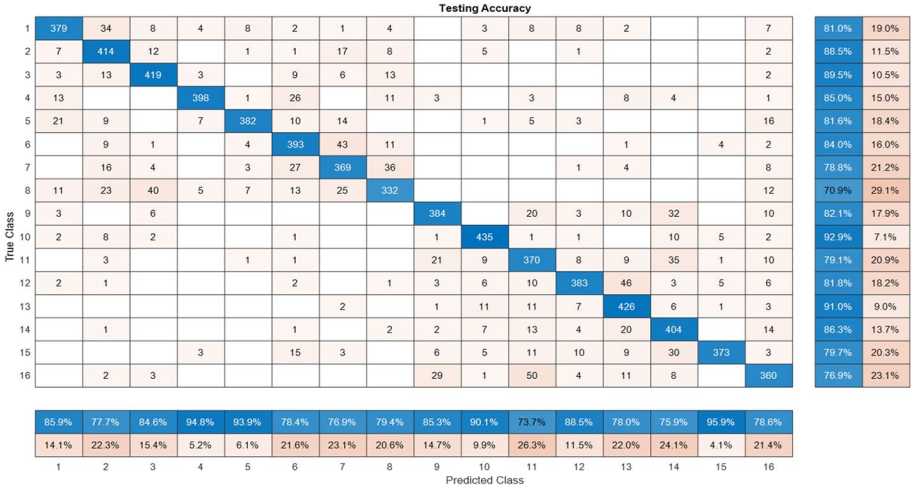

Fig. 8 Performance evaluation – Mel spectrogram T-F feature with CNN

Let us consider the word "Yes" uttered by two female and two male speakers to make effective and quantitative comparisons in time and frequency domains. Figure 11 indicates the comparative analysis between the test speeches uttered by two female speakers in time and frequency domains.

Similarly, analysis is done between the test speeches uttered by two male speakers. Figure 12 indicates the speech signals of two male speakers analyzed in time and frequency domains.

Figure 13 indicates the correlation analysis between the speeches of two female and male speakers.

Figure 14 depicts the correlation analysis done between the test speeches uttered by one male and female speaker.

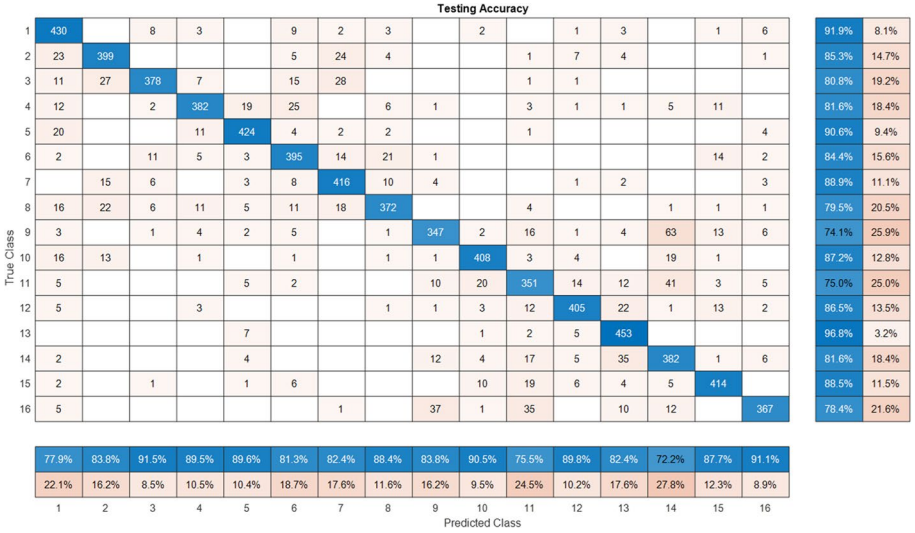

Fig. 9 Performance evaluation – Gammatonegram T-F feature with CNN

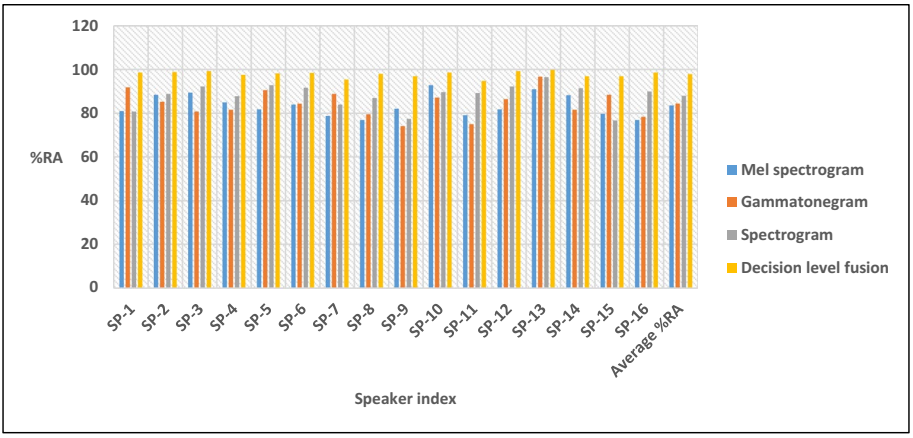

Fig. 10 Performance assessment - Comparative analysis of features with a decision-level fusion of features for CNN models

Correlation analysis reveals the fact that there is a similarity between the speeches uttered by two male or two female speakers. There is a difference in characteristics between the speeches spoken by one male and one female speaker.

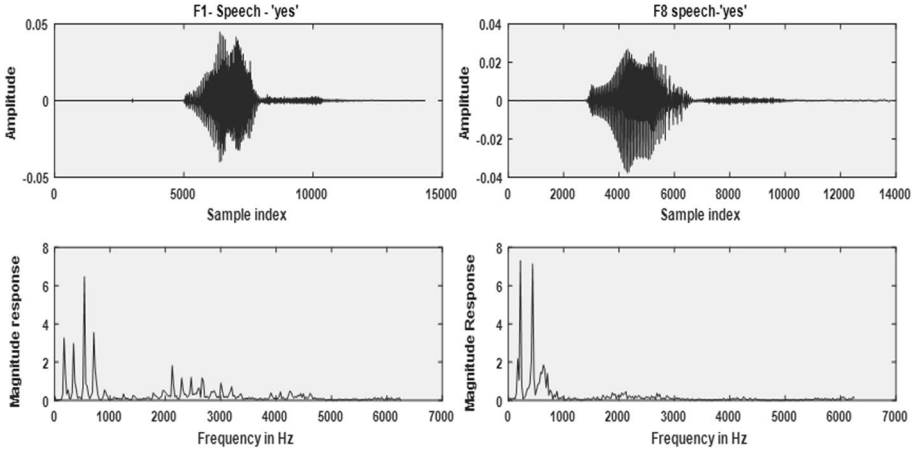

Fig. 11 Analysis- Female test speech – Time and frequency domain

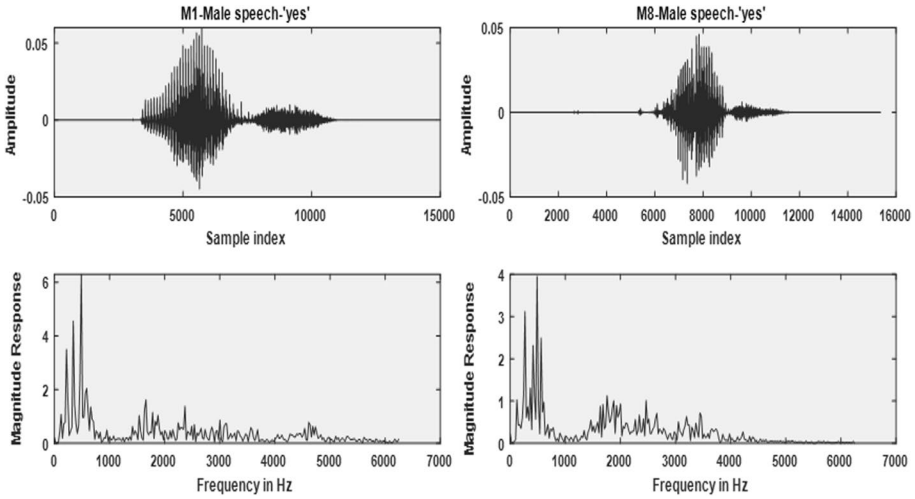

Fig. 12 Analysis- Male test speech – Time and frequency domain

3 Implementation of person authentication using Raspberry Pi hardware

Checking the voice characteristics of the input utterance using automatic speaker recognition can add an extra level of security. The CNN model ported on the raspberry pi achieves speaker recognition. The CNN model used for porting in the Raspberry Pi is being trained with a dataset from "TIMIT". The following methodology has been used for deploying feature extraction and a convolution neural network (CNN) for person authentication using speech on Raspberry Pi [20], [22], [10]. The two-dimensional spectrogram is used as a feature to train models and authentication of persons subsequently.

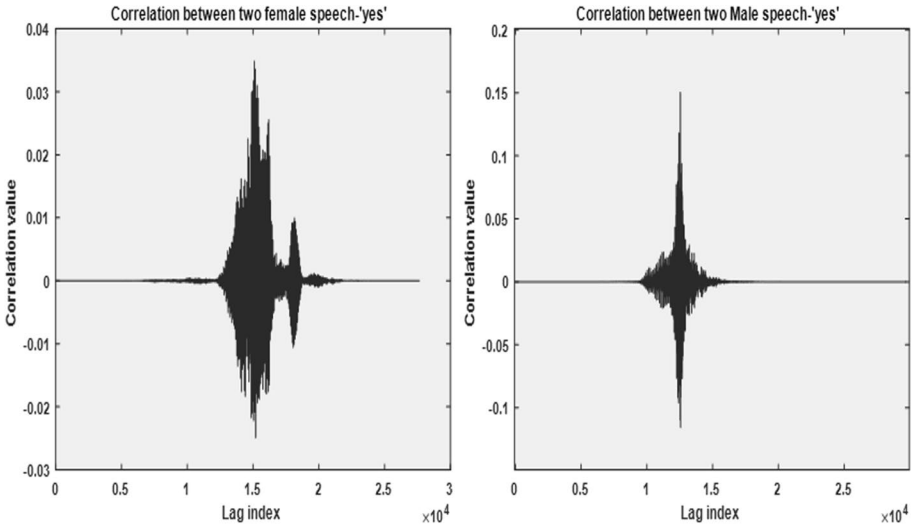

Fig. 13 Correlation analysis between test speeches of two male and female speakers

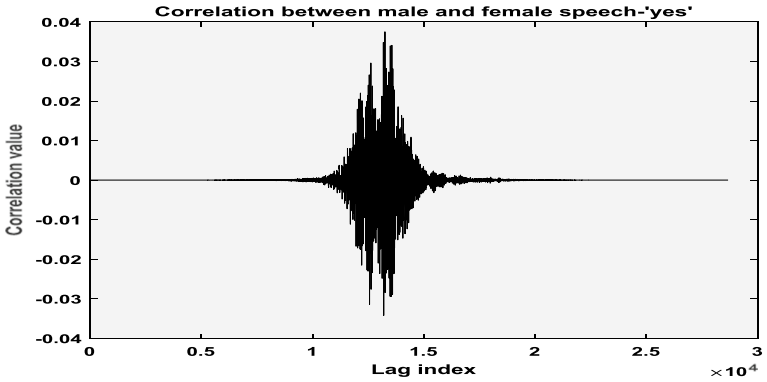

Fig. 14 Correlation analysis- test speech of a male and female speaker

Step 1: Connect a microphone to your Raspberry Pi board and use the list of AudioDevices; ALSA Audio Capture (Simulink Support Package for Raspberry Pi Hardware) block captures the audio signal from the default audio device on the Raspberry Pi hardware with the sampling frequency of 12500Hz and 512 samples per frame.

Step 2: Configure Code Generation Settings

```
Set parameter (model, SystemTargetFile=" ert.tlc")
Set parameter (model, TargetLang="C++")
Set parameter (model,TargetLangStandard="C++11 (ISO)")
```

Step 3: Select a solver that supports code generation

```
Set parameter (model,SolverName="FixedStepAuto")
Set parameter (model SolverType="Fixed-step")
```

Step 4: set the Hardware board to Raspberry Pi and enter your Raspberry Pi credentials in the Board Parameters
Step 5: set the External mode's Communication interface to XCP on TCP/IP
Step 6: Check Signal logging in Data Import/Export to enable signal monitoring in External Mode.
Step 7: Deploy the Model on Raspberry Pi and Perform Speech Command Recognition

For the Raspberry Pi-based person authentication system, two-dimensional spectrogram features are extracted from the speech commands of a pertinent speaker. These features are given to the CNN layered architecture to generate models for authenticating speakers. First, the network is trained to recognize the following speakers (F1-F8, M1-M8) for the training set of utterances. Then, the speech command of a speaker from Raspberry Pi board memory is given to the feature extraction block, and the extracted features are given to the trained CNN models to check the speaker's and model's association. Figure 15 indicates the Raspberry Pi board used for person authentication.

Figure 16 shows the Raspberry Pi board-based person authentication results using voice commands.

From the tests and the results, it can be deduced that the speaker recognition module performs better when the speaker's voice is loud, and there is a silence where the module performs computation. Moreover, the response time of this module is relatively fast. So, it can be used to securely log the data of the persons who have access to a particular place. So, getting the speech command from the customer/client is possible, and a Raspberry Pi board system receives the speech command at the receiving end. Speech command is taken to extract the spectrogram features and deployment of components in a Raspberry Pi-based receiving system [7], [9], [23]] with the models for checking and authenticating persons. Thus, the Raspberry Pi-based voice communication will be utilized for developing IoT based system for person authentication using speech commands.

This Raspberry Pi-based system would require the hardware and the networking for remote-based person authentication. However, this system can be extended to authenticate speech-based person authentication in real-time in banks and forensic sectors. Table 2 indicates the computational and time complexity of the work concerning the feature size and time taken for training and testing.

Fig. 15 Raspberry Pi board

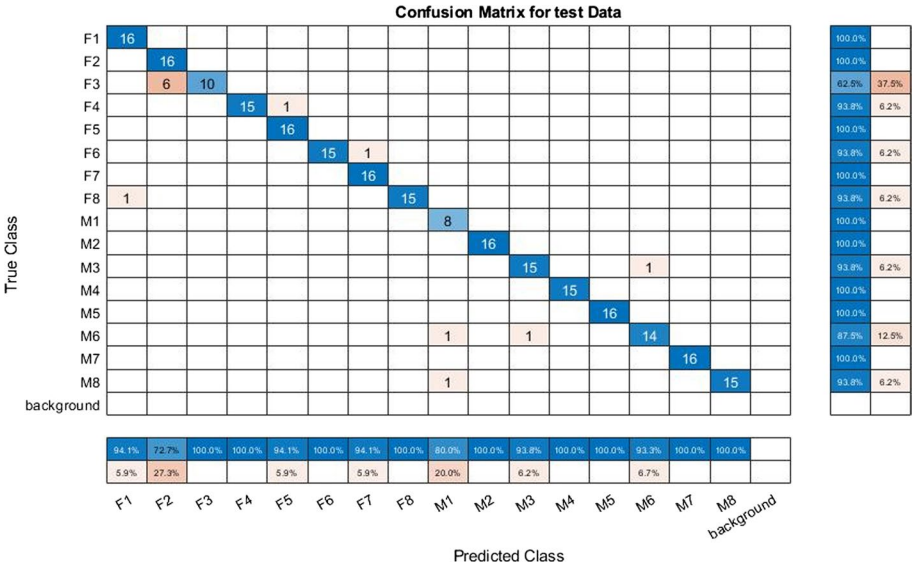

Fig. 16 Performance evaluation of the Raspberry Pi-based person authentication system

4 Conclusion

This paper mainly illustrates developing a person authentication system using T-F features and a convolutional neural network with the same set of speech utterances considered for training and testing. The training system is implemented by deriving T-F features with filters calibrated in BARK, Mel and ERB scales and applying features to the CNN layered architecture for creating templates or representations for the persons. The testing phase is implemented by extracting features and using features to the CNN-trained models and, based on the match test frame, is classified to be one among the set of enrolled speakers. Decision-level fusion of correct indices of test frames for each T-F feature has provided more than 95% individual accuracy for authenticating speakers, and overall average accuracy is 98% even though the same set of utterances is used for training and testing the system to authenticate speakers. This automated system emphasizes the importance of

Table 2 Analysis of the work- computational and time complexity

Feature	Spectrogram	Melspectrogram	Gammatonegram
Size of the feature	64X49	64X127	65X127
Time for training 76160 segments (10 epochs)	70 minutes	70 minutes	70 minutes
Time for testing 7488 segments	2.43 second	2.72 seconds	2.54 seconds

using speech as a biometric to ascertain speakers even if the voice password is the same for speakers. This system is extended to perform person authentication using speech in

real-time by utilizing the Raspberry Pi hardware. A Raspberry Pi-based system would be deployed to perform IoT assistive person authentication using speech command as a biometric. This automated system would be helpful in defence and forensic sectors where person authentication could be done by using speech as a biometric and other conventional biometrics.

Acknowledgements Authors wish to express their sincere thanks to the SASTRA Deemed University, Thanjavur, India, for extending infrastructural support to carry out this work.

Data availability All relevant data are within the paper and its supporting information files.

Declarations

Ethical approval This article does not contain any studies being performed with human participants or animals

Conflict of interest The authors have no relevant conflicts of interest to disclose.

References

1. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio, Speech, Lang Process 22(10):1533–1545
2. Albuquerque RQ, Mello CA (2021) Automatic no-reference speech quality assessment with convolutional neural networks. Neural Comput Appl 33:9993–10003
3. Bigun J, Fierrez-Aguilar J, Ortega-Garcia J, Gonzalez-Rodriguez J (2003) Multimodal biometric authentication using quality signals in mobile communications. In 12th International Conference on Image Analysis and Processing, 2003. Proceedings. (pp 2-11). IEEE
4. Das RK, Jelil S, Mahadeva Prasanna SR (2017) Development of multi-level speech based person authentication system. J Signal Process Syst 88:259–271
5. Dey S, Barman S, Bhukya RK, Das RK, Haris BC, Prasanna SM, Sinha R (2014) Speech biometric based attendance system. In 2014 twentieth national conference on communications (NCC) (pp 1-6). IEEE
6. Duc B, Bigün ES, Bigün J, Maître G, Fischer S (1997) Fusion of audio and video information for multi modal person authentication. Pattern Recog Lett 18(9):835–843
7. Gonzalez-Huitron V, León-Borges JA, Rodriguez-Mata AE, Amabilis-Sosa LE, Ramírez-Pereda B, Rodriguez H (2021) Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. Comput Electronics Agric 181:105951
8. Gunawan TS, Mokhtar MN, Kartiwi M, Ismail N, Effendi MR, & Qodim H (2020) Development of voice-based smart home security system using google voice kit. In 2020 6th International Conference on Wireless and Telematics (ICWT) (pp 1-4). IEEE
9. Hu F, Li Z, Yan L (2020) CNN and raspberry PI for fruit tree disease detection. In Intelligent Computing, Information and Control Systems: ICICCS 2019 (pp 1-8). Springer International Publishing
10. Johnston SJ, Cox SJ (2017) The raspberry Pi: A technology disrupter, and the enabler of dreams. Electronics 6(3):51
11. McCool C, Marcel S, Hadid A, Pietikäinen M, Matejka P, Cernocký J, ... Cootes T (2012) Bi-modal person recognition on a mobile phone: using mobile phone data. In 2012 IEEE international conference on multimedia and expo workshops (pp 635-640). IEEE
12. Pal M, Saha G (2015) On robustness of speech based biometric systems against voice conversion attack. Appl Soft Comput 30:214–228
13. Ramos-Lara R, López-García M, Cantó-Navarro E, Puente-Rodriguez L (2013) Real-time speaker verification system implemented on reconfigurable hardware. J Signal Process Syst 71:89–103
14. Rani R and Sachdeva R (2016) Genetic algorithm using speech and signature of biometrics. International Research J Eng Tech

15. Safavi S, Gan H, Mporas I, Sotudeh R (2016) Fraud detection in voice-based identity authentication applications and services. In 2016 IEEE 16th international conference on data mining workshops (ICDMW) (pp 1074-1081). IEEE
16. Sanderson C, Paliwal KK (2004) Identity verification using speech and face information. Digital Signal Process 14(5):449–480
17. Sarria-Paja M, Senoussaoui M, Falk TH (2015) The effects of whispered speech on state-of-the-art voice based biometrics systems. In 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE) (pp 1254-1259). IEEE
18. Suri M, Parmar V, Singla A, Malviya R, Nair S (2015) Neuromorphic hardware accelerated adaptive authentication system. In 2015 IEEE Symposium Series on Computational Intelligence (pp 1206-1213). IEEE
19. Telmem M, Ghanou Y (2021) The convolutional neural networks for Amazigh speech recognition system. TELKOMNIKA (Telecommun Comput Electro Control) 19(2):515–522
20. Vashistha P, Singh JP, Jain P, Kumar J (2019) Raspberry Pi based voice-operated personal assistant (Neobot). In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp 974-978). IEEE
21. Vázquez-Romero A, Gallardo-Antolín A (2020) Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy 22(6):688
22. Yamanoor NS, Yamanoor S (2017) High quality, low cost education with the Raspberry Pi. In 2017 IEEE Global Humanitarian Technology Conference (GHTC) (pp 1-5). IEEE
23. Yang S, Gong Z, Ye K, Wei Y, Huang Z, Huang Z (2020) EdgeRNN: a compact speech recognition network with spatio-temporal features for edge computing. IEEE Access 8:81468-81478