ANN based efficient feature fusion technique for speaker recognition

Savina Bansal[1], R. K. Bansal [2]and Yashender Sharma[3]

[1,2,3]*Giani Zail Singh Campus College of Engg & Tech., MRSPTU, Dabwali Rd, Bathinda, Punjab,151001, India*

Abstract

Speaker recognition is an all-important research field that intent to identify speaker using his/her voice utterances. It holds utilities in authentication, forensics voice comparison, mobile banking and security authentication for access control. An effectual feature extraction technique is vital for recognition or classification. In this work, Fused Features Hybrid extraction Technique (FFHT) has been proposed. It comprises features from time, frequency and cepstral-domain. A feed-forward NN model trained by means of Gradient Descent with Momentum function is used for speaker feature classification. Initial results achieved 97.56% regression accuracy, thus show the goodness of FFHT over contemporary techniques.

Keywords

Speaker recognition, Artificial neural network, MFCC, Hybrid Features, Feature fusion, FFHT.

1. Introduction

Speaker recognition is an intriguing area that finds applications such as surveillance, voice comparison, segmenting speakers, personalized interfaces [1]. This system in general analyze characteristics or features in speech that are unique among speakers. Its objective is to acknowledge a speaker by means of analyzing, processing and identifying speaker-specific traits that are present in the particular's voice. Text-dependent (A specified phrase needs to be uttered by the speaker) and text-independent (Speaker can say anything) are two classes of speaker recognition based on ease of text-restriction imposed to the speaker [2]. On the use of datasets, speaker recognition is categorized in open-set (Test sample may also belong to an unregistered speaker) and closed-set (Test utterance needs to belong to registered speaker) [3]. Feature extraction and classification are the foremostphases involved in speaker recognition. Certain features that are essential for identification of the speaker are extracted from speech data. High inter-speaker variations, low-intra speaker variations, and easily computable features are some important qualities of an ideal speaker recognition system. Considering the traits of an ideal speaker recognition system and the need for less complex training data, a new feature fusion extraction technique termed as Fused Features Hybrid extraction Technique (FFHT) is proposed. It is a combination on the available Mel Frequency Cepstrum Coefficients (MFCC), Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE), Chroma Feature (CF), Spectral Centroid (SC), Spectral Roll-off (SR) and Spectral Bandwidth (SB) techniques. The fusion of time-domain features, frequency-domain features and cepstral-domain features results in lesser training time with better quality. While in classification phase, the features of an unknown speech are extracted and compared with the feature database of registered speakers to identify the actual speaker. In this work, recognition or classification is performed using artificial neural network. It is an effectual computing system that try to solve any given problem by impersonating human nervous system. Here, a feed-forward back-propagation neural network model is used for classification of extracted features using FFHT.

International Conference on Emerging Technologies: AI, IoT, and CPS for Science &Technology Applications, September06–07, 2021, Chandigarh, India

EMAIL: savina.bansal@gmail.com (A. 1); drrakeshkbansal@gmail.com (A. 2); yashenders@gmail.com (A. 3)
ORCID: 0000-0001-7483-1841 (A. 1); 0000-0002-2288-4965 (A. 2); 0000-0002-4978-143X (A. 3)

2. Related Work

Rafizah *et al.* [2] reviewed speaker recognition techniques and challenges. Their work presented a structure of speaker recognition along with different feature extraction methods and classifiers. Zhongxin *et al.* [3] provided a broad summary of deep learning-based speaker recognition and also investigated the relation amiddiverse sub tasks. Different feature extraction techniques from various domains, image/texture-based features and deep features are reviewed by Garima *et al.* [4] and Alías *et al.* [5]. Murty *et al.* [6] combined residual phase information with MFCC features. Researchers used 149 male speaker utterances of NIST 2003 dataset and claimed 90% classification accuracy with Auto-associative neural network model. Fong *et al.* [7] did a relative study to categorize speakers by means of time-domain statistical features and machine learning classifiers.By using the multi-layer perceptron classifier,they obtained accuracy of approximately 94%. Kharibam Jilenkumari Devi *et al.* [8] proposed automatic speaker recognition using MFCC features with multi-layered perceptron. Training of MLP is done with back-propagation. IITG Multi-variability Speaker Recognition Database was used. The researchers claimed an accuracy of 94.44%. Soley-manpour *et al.* [9] explored clustering-based Mel Frequency Cepstral Coefficients features. ANN model is used to classify 22 speakers from the ELDSR dataset. Their experimental results achieved 93% classification accuracy. Savina *et al.* [10] proposed text-independent speaker identification and verification system using multi-domain features fusion based feature extraction process. Three different supervised machine learning classifiers were used. Two-open source datasets each containing 10 and 20 speakers were used. Researchers claimed better accuracy and better training time as compared to MFCC and hybrid MFCC techniques.

3. Proposed Methodology

Speech feature extraction is an important step in speaker recognition. Various speech feature extraction techniques are available in literature. Proposed FFHT is a two-step approach. In Step 1, it employs a fusion of features from time, frequency and cepstral-domainfor extracting salient features from the test speech sample. In Step 2, the feature fusion set is passed on to an ANN classifier for feature classification to recognize the speaker class.

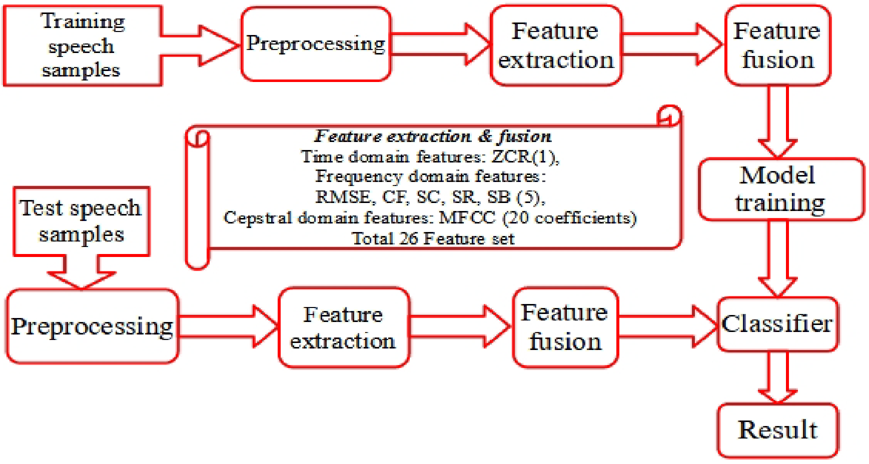


Figure 1: Proposed Speaker Recognition System

- Step 1: The FFHT employs a fusion of speech features from different domains viz. Zero Crossing Rate (from time-domain features), Root Mean Square Energy, Chroma feature, Spectral Centroid, Spectral Roll off and Spectral Bandwidth (from Frequency-Domain features) and Mel Frequency Cepstral Coefficients (from Cepstral Domain Features). In total, 26 features are thus used in the proposed technique. The pseudo code for proposed technique is given below.
- Step 2: It is the feature classification phase, wherein the features extracted in Step 1 are next used as input for Multi-Layer Perceptron (MLP) feed-forward neural network (Figure 2). In a feed-

forward neural network where values are transferred in a forward manner from input layer up to the output layer. The path of passing the inputs is always fed forward. Any output error is back-propagated to hidden inner layers for weight modifications so as to reduce the error. Gradient descent with momentum [11] is used as the acceleration of back-propagation method.

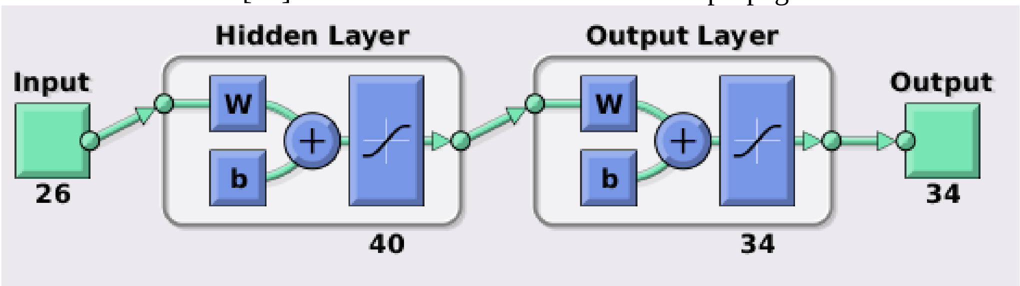


Figure 2: Structure of Multi-layered Feed-forward neural network

Table 1

Proposed Feature Fusion Technique

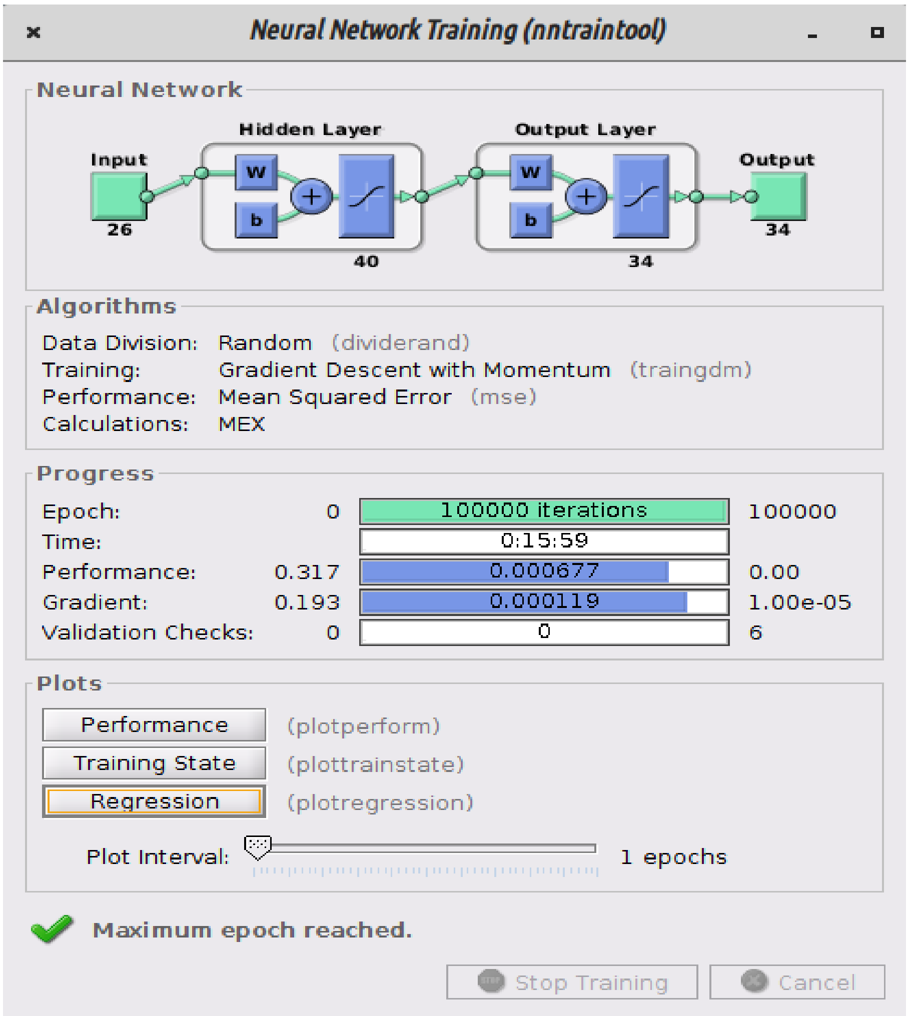
Result: *26 feature set*

- Input: Audio signal y(n), sampling rate s, time window N
- Create .csv file with “filename, ZCR, RMSE, CR, SC, SR, SB, MFCC(i)s” labels
- ZCR = $\sum(|diff(y_i(n) > 0)|)/length(y_i(n))$
- RMSE = $\mu(\sqrt{(\frac{1}{N}\sum_n(|(y_i(n)|))^2}))$
- Convert $y_i(n)$ into frequency-domain
- chroma = compute logarithmic STFT of the sound signal
- CR = mean(chroma)
- SC =$\mu((\sum_{j=b}^{c} f_j S_j)/(\sum_{j=b}^{c} S_j))$, where f_j is frequency corresponding to j, S_j is spectral value at j, b and c are band edges
- spec_roll = r such that$\sum_{j=b}^{r} S_j = 0.95(\sum_{j=b}^{c} S_j)$
- SR = mean(spec_roll)
- spec_bw = compute second order statistical value
- SB = mean(spec_bw)
- Compute 20 DCT coefficients,and MFCC(i) = DCT_i coefficients, where i = 1 to 20
- Repeat the same technique for feature extraction for each audio sample in the whole dataset.

4. Results and discussion

- Dataset Used:VoxForge dataset consists 34 speakers each having 10 voice samples of approximately 4 sec duration. All 340 speech samples are randomly divided into validation, training and testing datasets.

Simulations for FFHT are done on Matlab R2021a. Firstly, the input audio sample is fed and then 26 featuresare extracted individually for every speech sample. Lastly, extracted fusion based features are given as input to feed-forward back propagation neural network for classification. The back-propagation algorithm performs the Gradient Descent with Momentum based learning function on the FFNN. The MLP modelexhibits 34 outputs for the reason that of 34 speakers which are represented by binary bits i.e., 0 and 1 for speakers. For speaker 1, the n bits are 1 and other bits are 0, where n denotes total number of audio samples related to speaker 1. For speaker 2, the n+1 bit is 1 and other bits are 0. Same procedure is applied for 34 speakers. The system has been trained with 340 samples for 34 speakers at 100000 epochs, the error drops to 0.000677 as shown in Figure 3. This system accomplishes an overall accuracy of 97.564%.



Neural Network Training (nntraintool)

Neural Network

Algorithms

Data Division: Random (dividerand)
Training: Gradient Descent with Momentum (traingdm)
Performance: Mean Squared Error (mse)
Calculations: MEX

Progress

```
Epoch:               0   100000 iterations   100000
Time:                    0:15:59
Performance:     0.317   0.000677            0.00
Gradient:        0.193   0.000119            1.00e-05
Validation Checks:   0   0                   6
```

Plots

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Plot Interval: 1 epochs

✓ Maximum epoch reached.

Figure 3: Training progress of MLP at 100000 iterations

Figure 4 shows the neural network performance of the system and the best validation performance obtained is 0.001923 at 100000 epochs.

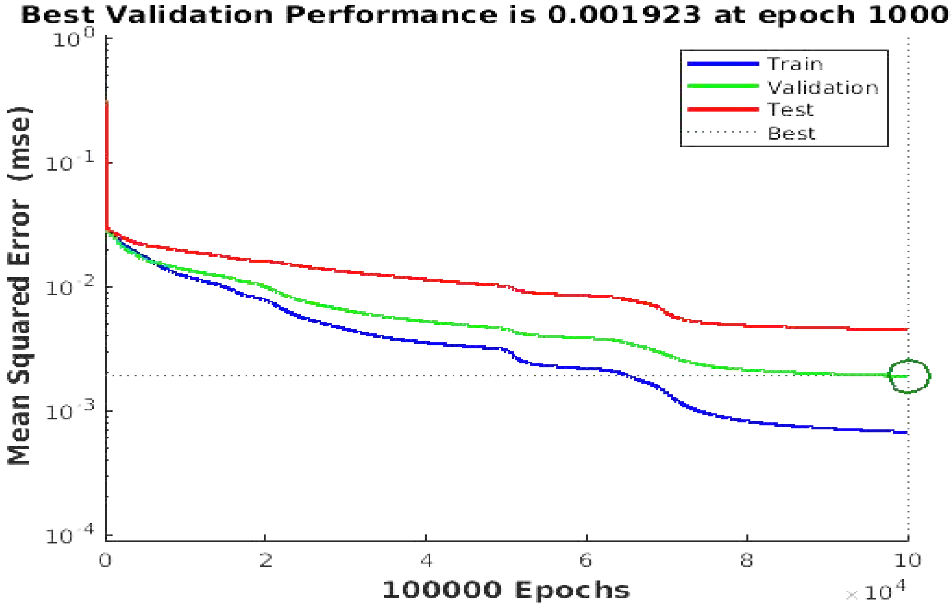


Figure 4: Validation performance of ANN

Figure 5 shows the neural network regression plots.

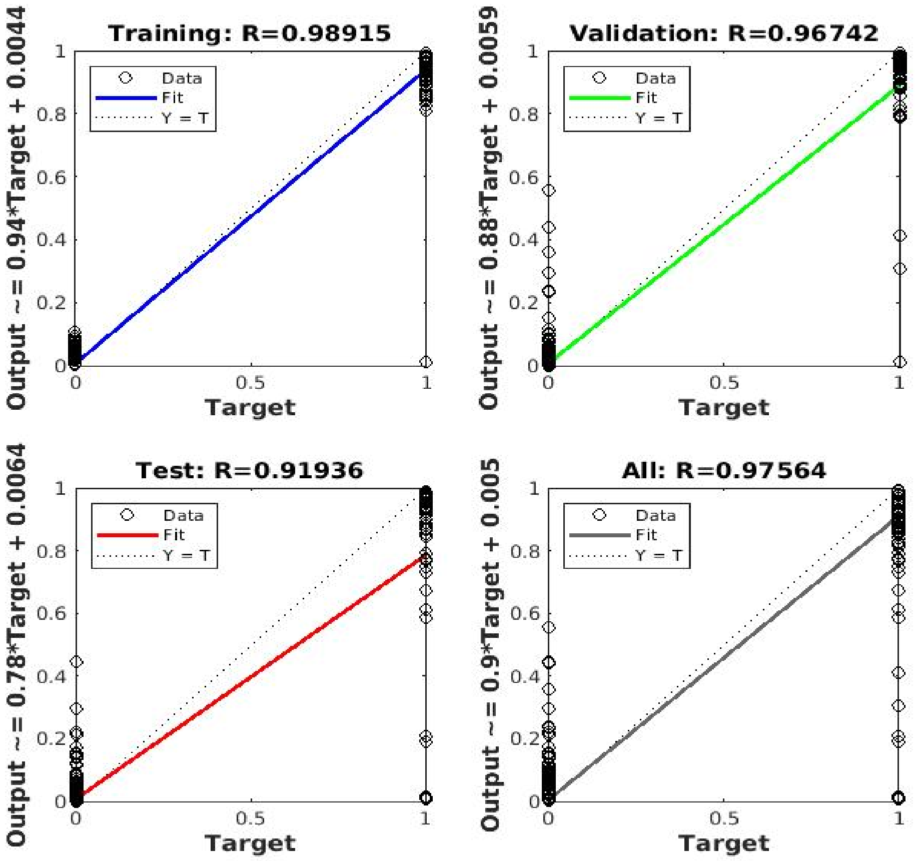


Figure 5: Regression Plots

Table 1 shows the comparison of our FFHT with other state of-art-methods [6], [8] and [9]. It has been found that using fusion of multi-domain features, better performance/recognition rate is accomplished by the proposed technique.

Table 2

Table title

S No	Methodology	Accuracy
1	Residual phase information + MFCC features with Auto-associative neural network[6]	90.00%
2	MFCC features with MLP[8]	94.44%
3	Clustering based MFCC features with ANN[9]	93.00%
4	Proposed FFHT(Fused Features Hybrid extraction Technique)	97.56%

5. Conclusion

In this work, a fusion based speech feature extraction technique FFHT for speaker recognition is proposed and evaluated on limited dataset of 34 speakers. Furthermore, results of the proposed work have been achieved with lesser number of training samples per speaker. For the tested dataset, initial results achieved 97.56% accuracy and are clear indication of better accuracy and performance as compared to experimental results on MFCC and hybrid versions of MFCC. In future, we will be realizing Automatic Speaker Recognition system using deep learning, more exhaustive results shall be presented in our future work.

6. References

[1] Hitesh Garg, R K Bansal & Savina Bansal, “Improved Speech Compression using LPC and DWT Approach”, International Journal of Electronics, Communication & Instrumentation Engineering Research and Development (IJECIERD), Volume 4, Issue 2, Pages 155-162, (2014).

[2] Rafizah Mohd Hanifa, Khalid Isa, Shamsul Mohamad, “A review on speaker recognition: Technology and challenges”, Computers & Electrical Engineering, Volume 90, 107005, (2021) https://doi.org/10.1016/j.compeleceng.2021.107005

[3] Zhongxin Bai, Xiao-Lei Zhang, “Speaker recognition based on deep learning: An overview”, Neural Networks, Volume 140, (2021), Pages 65-99, https://doi.org/10.1016/j.neunet.2021.03.004

[4] Garima Sharma, KartikeyanUmapathy, Sridhar Krishnan, “Trends in audio signal feature extraction methods”, Applied Acoustics, Volume 158, 107020, (2020) https://doi.org/10.1016/j.apacoust.2019.107020

[5] Alías F, Socoró JC, Sevillano X. “A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds”, Applied Sciences. 2016; 6(5):143. https://doi.org/10.3390/app6050143

[6] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," in IEEE Signal Processing Letters, vol. 13, no. 1, pp. 52-55, Jan. 2006, https://doi.org/10.1109/LSP.2005.860538

[7] S. Fong, K. Lan, and R. Wong, ‘‘Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection,’’ BioMed Research International, vol. 2013, Oct. 2013, Art. no. 720834, https://doi.org/10.1155/2013/720834

[8] Kharibam Jilenkumari Devi, Ayekpam Alice Devi and KhelchandraThongam, “Automatic Speaker Recognition using MFCC and Artificial Neural Network,” International Journal of Innovative Technology and Exploring Engineering (IJITEE) Vol. 9, SI. 1 39-42 (2019) https://doi.org/10.35940/ijitee.A1010.1191S19

[9] Soleymanpour, M., Marvi, H. Text-independent speaker identification based on selection of the most similar feature vectors. Int J Speech Technol 20, 99–108 (2017). https://doi.org/10.1007/s10772-016-9385-x

[10] Savina Bansal, R K Bansal, Yashender Sharma, “An efficient feature fusion technique for text-independent speaker identification and verification”, International Conference on Advances in Data Computing, Communication and Security (I3CS-2021), 2021, NIT Kurukshetra

[11] Amit Bhaya, Eugenius Kaszkurewicz, Steepest descent with momentum for quadratic functions is a version of the conjugate gradient method, Neural Networks, Vol. 17, Iss. 1, Pages 65-71 (2004) https://doi.org/10.1016/S0893-6080(03)00170-9