IJMTST is UGC Approved Journal with Journal Id: 43137

International Journal for Modern Trends in Science and Technology, 8(S05):109-114, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST08S0518
Available online at: https://www.ijmtst.com/vol8si05.html

Text-Independent Automatic Speaker Recognition using Machine Learning

N.M.Ramalingeswara Rao | G.Jagadeesh Chandra | K. Jayaram | N. Navya Nissi | D. John Jesu Ratnam

Department of Electronics and Communication Engineering, Godavari Institute of Engineering and Technology(A), JNTUK, Kakinada.

To Cite this Article
N.M.Ramalingeswara Rao, G.Jagadeesh Chandra, K. Jayaram, N. Navya Nissi and D. John Jesu Ratnam. Text-Independent Automatic Speaker Recognition using Machine Learning. International Journal for Modern Trends in Science and Technology 2022, 8(S05), pp. 109-114. https://doi.org/10.46501/IJMTST08S0518

Article Info
Received: 26 April 2022; Accepted: 24 May 2022; Published: 30 May 2022.

ABSTRACT

In the past decades, security is the main for everyone, and processing of security by the voice control. In this condition, security is designed by speaker voice command and speaker recognition for a short duration of text speech samples. In speaker recognition systems, the processing by Gaussian mixed models is impaired by low quality and short duration of the speech. We are proposing this project for forensic-based voice and speaker recognition and that way we are taking the voice and comparingit with the recorded voice. The voice matched and speaker recognition by preprocessing and recognized by machine learning. In this project, a large number of best material selection criteria were described, suitable for the scoring stage in forensic automatic speaker recognition systems. An application of quality-based speaker features performs outperforms forensic speaker recognition systems that assume the uniform quality of speech during model training and scoring. The speech(or)speaker recognition was described by the combination of discrete wavelet transform (DWT) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP) for feature extraction. This process of speaker recognition is enhancing the performance of more features from the speech signals and applying other computation techniques to lead to the improvement of recognition rate and computational technique if noisy speech signal is present, then separating/extracting the original by DWT and Mel
-frequency cepstral coefficients (MFCCs)

Keywords:Speaker recognition, voice comparison, Gaussian mixture model, machine learning,discrete wavelet transform (DWT), MFCC

1.INTRODUCTION

In past years, an increasing interest in security systems has arisen. These systems are very useful as they allow managing security in a very efficient way, reducing human resources. Most security systems were implementedby an access control system. In this way,a vast number of security resources by voice commands and speaker recognition in this process of verifyingpeople's identity by their voice. For security purposes, the voice recognition systems are used as a biometric system performance that will allow the control to access in fast response way and low intrusive way and reduced collaboration of people voice samples in comparison

The voice/sound generated by humans is peculiar or different from each other. the voices are generated by vocal tracks. The vocal tracks are consisting of the Oral cavity, nasal cavity, and pharyngeal cavity.

The oral cavity outsourcing the voice by the movement of jaws, tongue, and lips and it was formed by bony structure by both palate and soft palate and oral cavity muscles.

Nasal Cavity for voice formation by soft palate by isolated and it was generated the voice by passing air for voice transformation.

The pharyngeal cavity presents at the bottom of the throat and it can be stretching the base of the tongue towards the sides of the pharynx. In the lower part that is enlargedby the vocal cordsare a couple of fleshy membranes traversed by the air coming from the lungs Whilethe production of a sound, the gap between the fleshy membranes (glottis) can be completely constrained by opening and partial closing. Due to the peculiarity of the voice formation and it can be possible to recognize a particular individual from their voice. This operation can be described in a way of an automaticSpeaker recognition ASRas a hybrid biometric recognition approach and it has two componentsthe physical one related to the formation of the vocal apparatus, and the behavioral componentpertinent to the mood of the speaker just in the recording. There are several approaches to ASR based on features, vector quantization, score normalization, pattern matching, etc., and we propose a text-independent ASR system using Machine Learning based on Mel-Frequency Cestrum Coefficients (MFCC) and Gaussian Mixture Models (GMM). This model parameteris estimated with the maximum similarity ovulation use of the Expectation and Maximization (EM) algorithm

In the speech/speaker recognition security method we are using voice noise cancellation because the recorded speaker signal could be corrupted by environmental additive noise like background sounds and a spectral subtraction algorithm is also used to get a clear voice. Comparisons of the voices with stored data the state of the art demonstrate the effectiveness of the proposed approach inaccuracy rate. The data acquisition will be performed through simple microphones which are well spread and their cost is negligible. Cheap instrumentation may be more affected by disturbances like background noise and could be no more sufficient for efficient noise suppression

Text-Independent Automatic speaker recognition systems use the source of the low-level acoustic features of a speech signal. Speaker recognition is an important topic in voice/speech signal processing and has a variety of applications,especially in security systems and AI. Voice-controlled systems and devices are heavily on speaker recognition. Applications of speaker recognition security control for confidential information, verifying customers for bank transactions, forensics, and remote access to computers.The development of an automatic speaker recognition system that incorporates classification and recognition of separate home language speakers. The system uses machine learning algorithms that learn features extracted from the Separate speech data to train the classifier model. The system can be used to automatically authenticatespeaker identities using their voices to allow only the identified persons an access right to information systems or to facilities that provide to be protected from the unauthorized person commands.

Fundamental tasks of Speaker Recognition

Speaker recognition has two fundamental tasksnamely speaker verification and speaker identification. Speaker verification is the task of determining whether an unknown voice is from a particular enrolled speaker and the speaker has to claim an identity and the system validates the claimed identity. Some Applications of speaker verification aretelephone banking, computer login, cellular telephone fraud prevention, and calling cards. Speaker identification is the task of comparing an unknown voice with one from a set of enrolled speakers. The speaker provides a voice sample (without claiming an identity) and the system determines to which of the known set of enrolled speakers doesthe voice sample belongs Potential speaker identification applications include automatic speaker labeling of recordedmeetings for speaker-dependent audio indexing and intelligent answering machines with personalized caller greetings.

Classification of Speaker Recognition Systems

Speaker recognition systems are further classified by the constraints placed on the text of the speech used in the system, the classification can either be text-dependent or text-independent. In the text-dependent case, the spoken text or phrase used to train and test the system is fixed for each speaker Text-dependent speaker recognition systems are used mostly in services are access control and telephone-based services, where users are considered to

be cooperative. In the text-independent, the spoken phrase or text used to train and test is not fixed. Text-independent speaker recognition systems are the most flexible and widely used in events where speakers can be considered non-cooperative users, as they do not specifically wish to be recognized such as forensic analysis and surveillance procedures. Text-dependent recognition achieves higher recognition performance than text-independent recognition, However, due to the flexibility that the text-independent recognition provides, the increasing development trend is in the building of the text-independent recognition systems.

Phases of Speaker Recognition

A speaker recognition system using machine learning are composed of two different phases, a training phase and a test phase. In the training phase, a speaker's voice is recorded and a number of audio features are extracted to form a unique (voice-print) model that uniquely identifies the speaker. In the testing phase, (also known as the recognition phase) the voice sample provided is compared against the recorded created model.

Voice Comparison

In speaker recognition, one more operation is called voice comparison.

Voice comparison is a task to analyses two recordings of the speaker and make a decision whether the voices belong to the same speaker or to different speakers. The input to voice comparison is two voice recordings and the output is similarity score in the range 0 to 1. Voice comparison are done, based on the text or words used in voice, types of system that address the problem of speaker recognition (identification and verification) and comparison can be classified into two types:

1) text-dependent and (2) text-independent.

Text-dependent system is predefined by connecting to text used for training and testing and the text-independent system will be capable of using any text.

Different approaches of voice comparison

In voice comparison there are 4 types of approaches and they are (1) auditory, (2) spectrographic, (3) acoustic, and (4) automatic approach.For all these approaches we comparing a voice, at least two recordings of a speaker are needed.

In auditory approach, the results will be the experts' (machines') subjective judgment on the basis of listening of speech recording.

speech recordings are converted into speech images called a spectrogram.The spectrogram reflects the frequency spectrumwhich spectrographic approach is a image-based approach is also known as "voiceprints".

In Acoustic-phonetic approach, It needs making quantitative estimates of the acoustic properties like (pitch, formant, fundamental frequency, and HNR) on equivalent phonetic units in both recordings of the speakers.

. In an automatic approach, speech features are automatically extracted by Frame-wise. The automated approach does not use different acoustic features on a specific part of the signal. Examples of automatic approaches are MFCC, GMM, etc

2. METHODOLOGY

Speaker Recognition and Voice Comparison System in light of Deep learning is asustainable strategy for Machine Learning. Deep-learning procedures have been effective in perceiving speakers. The DNN is prepared to group speakers with acoustic attributes at the edge level. The typical elements of these speakers, or called d-vector features, are then used to verify and confirm different speakers. The CNN comprises of a few such layers of convolution that apply a wide scope of channels to ensuing little neighborhood/local input areas. That convolution layer is trailed by a maximum pooling layer, which creates a lower resolution of the activities of the convolution layer by eliminating the absolute channel initiation from e.g., a 2X2 window. At the end, completely associated layers in the end coordinate all results of the last max-pooling layer to order speakers

3. PROPOSED METHOD

A sensor which makes obtaining of information and its resulting testing in the particular case the sensor is a microphone existing in device (laptop, mobile), it

perhaps with a high Signal to Ratio (SNR) value. the sampling rate is 8 kHz.

In Data Acquisition, Forensic scientists or researchers generally avoid modelling raw audio because it ticks so often. Generally, the text-independent voice comparison system requires a type of datasets that contain audios of the same voices and subjects having different dialogues. Some datasets are available for recognition and comparison tasks such datasets include microphone audio data, telephone speech, age-wise speech corpus.

A step of pre-processing that in the voice context is constituted by the signal cleaning and simply denoising algorithm can be applied to recorded data after a normalization procedure. In order to clean recorded speech signal from environmental additive noise, a spectral subtraction algorithm is used in this method.

Pre-processing of audio data is a very important step after data acquisition from sensor because real-world audio data is noisy. VAD (Voice Activity Detection) is used to separate voiced data and unvoiced data, i.e., VAD is used to find out the presence and absence of a human in speech. The VAD strategies utilize the prompt proportions like pitch, frequency etc., of the dissimilarity separation among speech and noise. VAD was based on extracting features such as short-time energy, zero-crossing rate, and pitch analysis. The classification of voiced and unvoiced segments is done based on cepstral coefficients, and wavelet transforms. Feature extraction is one of the most important aspects of speaker recognition and generates a vector that represents the speech signal. We extract features using pyAudioAnalysis, an open-source comprehensive package developed in Python. A total of 34 short-term features are implemented by pyAudioAnalysis. The Raw Audio samples are extracted by Time-domain features (Zero Crossing Rate (ZCR), Energy and Entropy of Energy). Frequency domain features are then based on the magnitude of the Discrete Fourier Transform (DFT). The cepstral-domain features (Mel Frequency Cepstral Coefficients or MFCCs) result after the Inverse DFT is applied on the logarithmic spectrum shows the time domain (ZCR), frequency domain (Spectral Centroid) and cepstral domain (MFCCs) features extracted from a single audio file of one speaker. MFCCs are popular audio features extracted from speech signals for use in recognition tasks and widely used for speaker and speech recognition

In Deep learning model, after pre-processing, the inputs are fed into the model. The Siamese Neural Network (Siamese NN) is well adapted for comparison. Siamese NN learns a similarity function that takes two inputs with similar weights and length (i.e., spectrogram or voiceprint) as input and shows how identical the both two inputs are shown in fig no:01. The Siamese NN's are acquiring the similar data through loss function and shows the score of how much both input objects are identical. In past days the Siamese NN was used for Signature verification. Deep neural networks (DNNs) are artificial neural networks with multiple hidden layers between the input and output layers. In Deep Neural Networks a classifier is used as MLP with more than two hidden layers that typically uses random initialization and stochastic gradient descent to initialize and optimize the weight. DNNs can handle extremely complex tasks and they require more time to prepare and are computationally costly

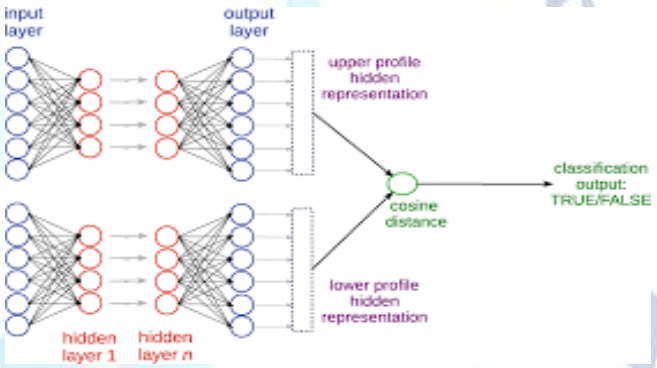

Fig no :01 Siamese Neural Network

The process of generating a specific template for every speaker in this process we are using the Gaussian Mixture Models (GMM) where model parameters are estimated with the maximum similarity and EM algorithm (Expectation and Maximization) is used.

In case of the user is registering (enrolment) for the first time to the system, this template will be added to the database, using some database programming techniques. The test among users already presents in the database, a comparison (matcher) determines which profile matches the generated template of the test speech. The matcher utilizes a similarity test over similar weighted speech/voice prints or spectrograms are obtaining by a ratio value that can be accepted if it is higher than a 0.5. in Fig 2: Proposed model of Text-Independent Automatic Speaker verification. The technologies are used for the development of the biometric system are the MMFCC for the extraction of

the characteristics and the GMM used for the statistical analysis of the data matched and obtained, for comparison of voice prints and generation of template speechanalyzed by the GMM. The template is employed when a speaker details and voice were added into the system or for the test step among the users already registered. The public voice database was used in order to validate the system. The present aims are to support who intends to realize and test an automatic speaker recognition system, a speech recognition engine, or any application related to analysis, to the recognition and more to the study of the human voice as security purpose. Anyone can register on the website and send his own voice recordings to be made available at our database. For this operation we are storing recording of persons voice of speaker utterances were randomly extracted. For each speaker two speeches were employed: the first one in order to perform the training and second one is in testing.

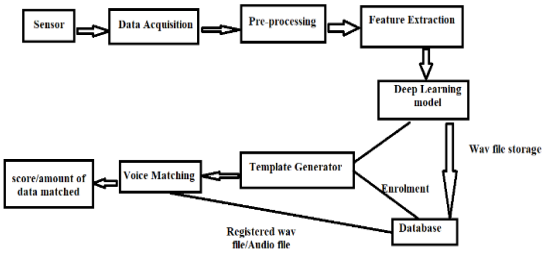

Fig 02: Proposed model of Text-Independent Automatic Speaker verification

In the pre-processing phase, the signal has been improved using spectral subtraction and segmented into frames partially overlying (50%) and relatively small. Frames not containing voice were skipped. The size of each frame is less than 20ms in order to make the contained wave stationary. The discontinuities at the edges of the frame are minimized by Hamming window. For each frame 20 MFCC were calculated. The obtained data represents the characteristics of a speaker. A vector of Mel-Cepstral coefficients for each frame for

5. RESULTS

Checking of two speaker audios in the proposed method having one completed detail in database and another not having the details in database. After operation the results are in table 1.

Speakers	Database recording	Language Granules	Accuracy	Status
Speaker1	Yes	Matched	85%	Recording matched with speaker
Speaker2	No	Unknown	0%	Store the recording in database
Speaker2 (After Store in Database)	Yes	Unmatched (Check For Pitch, Pronunciation)	50%	Checking for pitch and pronunciation
Speaker2	Yes	Matched	85%	Recording matched with speaker

6. CONCLUSION

We are concluded that the project of Text-Independent Automatic Speaker Recognition using machine learning are useful in identification and voice comparison of speaker.in this paper we are considering the voice database as storage of voice data and compared with input through deep learning model.in deep learning model ,we having the algorithms of neural networks like Convolution neural networks for getting the similar weighted voice/wav file and after CNN the similar

weighted file and input will compared by Siamese neural network and score will represents the how much amount of voice are matched.in this project various datasets used for automated voice processing.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] A Maesa, F Garzia, M Scarpiniti, R Cusani - Journal of Information, 2012. This shows the accuracy and time results of a text independent automatic speaker recognition (ASR) system, based on Mel-Frequency cepstrum.

[2] J Wang, MT Johnson - 2012 International Conference. This paper describes a unique cross-phoneme speaker identification experiment, using deliberately mismatched phoneme sets for training and testing.

[3] S Bhardwaj, S Srivastava, IEEE transactions on 2013.This paper presents three novel methods for speaker identification of which two methods utilize both the continuous density Hidden Markov model (HMM) and the generalized fuzzy model.

[4] S Nandyal, SS Wali, SM Hatture - International Journal of Signal, 2015. This paper shows how speech processing has emerged as one of the important application areas of digital signal processing.

[5] SS Nidhyananthan, RSS Kumari - WSEAS Transactions on Signal, 2013. This paper propels the utilization of Dynamic Mel-Frequency Cepstral Coefficient (DMFCC) component and mix of DMFCC and MFCC highlights for hearty language and text.

[6] S Farah, A Shamim - 3rd IEEE International Conference,2013. This paper shows personality confirmation is vital issue in current period of data innovation. Customary method for character confirmation utilizing keys or individual distinguishing proof numbers.

[7] A Nasef, M Marjanovic-Jakovljevic, A Njegus - Analog Integrated Circuits, 2017. This paper shows performance optimization in speaker recognition is a challenging task in the field of vocal based human-computer interaction.

[8] V Vasilakakis, S Cumani, P Laface, P Torino, biometric technologies, 2013. This paper shows most cutting-edge speaker acknowledgment frameworks depend on Gaussian Mixture Models (GMMs), where a discourse section is addressed by a minimal portrayal.

[9] B Ayoub, K Jamal, Z Arsalane - World Congress, 2015. This paper is to assess, examine and think about the presentation of the most famous MFCC variations for highlights extraction in message free speaker recognizable proof.

[10] I Shahin, MN Ba-Hutair -12th International Conference, 2014.In this paper we focus on Emarati speaker identification systems in neutral talking environments based on each of Vector Quantization (VQ), Gaussian Mixture Models.

[11] G Nijhawan, MK Soni - International Journal of Image, Graphics and Signal, 2013.This paper presents a new approach for designing a speaker recognition system based on Mel frequency cepstral coefficients (MFCCs) and voice activity detector (VAD)

[12] AH Mansour, GZA Salh, KA Mohammed - International Journal, 2015. In this voice acknowledgment is a significant and dynamic exploration region of the new year's. This exploration expects to fabricate a framework for voice acknowledgment utilizing dynamic time wrapping calculation.

[13] D Desai, M Joshi - Recent Advances in Intelligent Informatics, 2014. In this speaker recognition is widely used for automatic authentication of speaker's identity based on human biological features.

[14] CR Rashmi - International Journal of Computer Science and, 2014. In this human voice is an interesting trademark for any person. A significant biometric apparatus can be planned in view of the capacity to perceive an individual by his/her voice.