SPEAKER IDENTIFICATION SYSTEM BUILT ON A HYBRID MODEL THROUGH DIFFERENT FEATURE EXTRACTION TECHNIQUES-A REVIEW

[1]N.KALADHARAN and [2]Dr.R.ARUNKUMAR

[1]Lecturer,
Department of Computer Engineering,
Government Polytechnic College, Theni
[2]Associate Professor,
Department of Computer Science and Engineering,
FEAT, Annamalai University

ABSTRACT

This paper aims to provide a brief outline in to the area of speaker identification techniques. Speech is the most common way for human to communicate. The speech recognition permits system to interact and process the provided verbally by the user. Speech recognition can be defined as process of recognising the human voice to generate commands or word string. Speaker recognition can be divided in to dual techniques. Speaker identification and speaker verification. Speaker identification refers to the process of identifying human voice by means of artificial intelligence techniques. Speaker identification technologies are extensively useful in security and surveillance, voice authentication, electronic voice spying and identify verification. In the speaker identification process, extracting noticeable features from speaker utterances is an important task to accurately identify speakers. Deep leaning algorithms have been mostly used to further enhance the capabilities of computers excellently.

KEYWORDS

MFCC, Spectrogram, GMM, CNN, DNN

1. INTRODUCTION

Speech signal are dominant broadcasting of communication that always convey rich and useful information, such as emotion, gender, accent and other unique characteristics of a speaker[1]. Speaker recognition is the process of finding a person based on the voice of the speaker [2]. Speaker recognition is attractive widely used for different application such as security, audio indexing, and forensic application [3]. Speaker recognition is a method to extract the features from the speech related to speaker and classify the speakers. Feature extraction is a route to first identify the features required for different speech processing tasks and then excerpt the features [4]. Speech features extraction techniques commonly used in speaker recognition system include LPCC, MFCC, first and second order coefficient Cepstrum and RASTA filters [5]. Common speech recognition method include Speaker identification and speaker verification are two chief processing in speech recognition. Speaker identification involves the identification of a presenter words from a cluster of trained speaker sounds. Then, the speaker with a high likelihood of test utterance is identified as the speaker. Generally speaker identification system starts the feature extraction and then utilizes a large scale of unlabelled speech data to train a model to capture .speaker characteristics in and supervise way, finally training a classifier for the speaker classification [6]. Speaker identification can be divided in to text independent, dependent and prompted identification [7]. The method for speaker identification task can be divided in to two mainly process, which are training process and the testing (identification or matching) process [8]. Recently, with increase in deep learning in the speech recognition community a number of various deep neural networks have successfully applied to speaker recognition [9].

K.S.Babai,P.K.Suresh,K.UmaRani et al. (Eds): ICARTEC-2021
pp. 11 - 15, 2021.IJCI – 2021

DOI:10.5121/ijci.2021.100302

This paper is organized as follows, Section 2 describes literature study ,Section 3 represent types of speaker recognition , Section 4 approaches to speaker identification, Section 5 deals the speech feature extraction techniques, Section 6 denotes modelling techniques, Section 7 deals the speech datasets and finally section 8 reveals the conclusion.

2. LITERATURE STUDY

Speaker recognition is a generic term for related to speaker deification and verification based on the information contained in to acoustic signal. Speaker recognition under noisy and unconstrained condition in extreme challenging tasks [10]. Speaker recognition is that the systems are created to detect, validate and segregate the individual speaker. SR systems cab be modelled either as text dependent or text independent system. Text dependent system can be used only for cooperative users and user needs bring prompted by the system. In text independent system, these are no constraints on the words that can use comparison. Speaker recognition is the method used to identify when the speaker speak such word or speech. The techniques for speaker recognize attempt to cover the different aspects for recognizing persons from their voice. Since can speaker has his or her characteristics means of speaking, including the use of particular accent, rhythm, intonation style, pronunciation pattern, choices of vocabulary and so on. Speaker identification consists of recognition a speaker between a fixed set of speakers by comparing his or her speaker expression with known references. Common speaker recognition methods include HMM, GMM, VQ, DTW, SVM, and ANN. For more than decades, Gmm-Ubm has become widely used paradigm in speaker recognition system because of the good performance. Speaker identification has gained increasing attention from the academic and industry communication in recent years and it is speech widely used in application, including survileince, and discriminative speaker embedding learning and speaker diarization. The principal goal of SI is to automatically infer the identity of speaker from an input utterance given a closed set of known voice models[11].In SI, text dependent mechanism to work, each speaker mist speak the same prescribed text where as in text independent system speaker are allowed to lead different texts during training and testing.

3. SPEAKER RECOGNITION

Speaker recognition is a practice of identifying individual persons from their voice. Owing to different voice making organs such as vocal tract forms, larynx shape, glottis type and some other parts, no individuals have identical voice. SI is categorized in to two divisions.

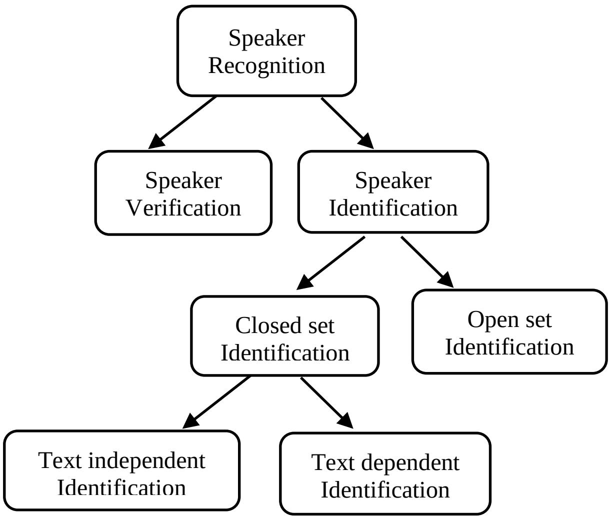
Speaker Recognition
Speaker Verification
Speaker Identification
Closed set Identification
Open set Identification
Text independent Identification
Text dependent Identification

Fig.1. Order of Speaker Recognition Model

They are open set and closed set. In situation of closed set speaker identification, the authorized person of the unknown voice sample is one of the known speakers. In open set speaker identification, it is not known that the authorized speaker of the unknown voice trial is extant in the mention or not. Speaker recognition methods can similarly split in to text dependent and text independent methods. In case of text dependent methods a speaker is essential to utter and planned set of words or sentences. In case of text independent method, there is no determined set of words or sentences and the speaker's may not level be alert that they are being verified. Features of speech are extracted from the alike utterances.

4. SPEAKER IDENTIFICATION

Speaker identification is the task to identify an unknown speaker from a set of already known speakers .The traditional speaker identification task usually consists of three statges: Pre-processing, feature extraction and speaker modelling. Speech pre-processing comprises pre emphasis, framing, and windowing and so on. Feature extraction is to extract the characteristic limits that effectively characterize the speaker’s traits from the pre-processed speech signal. The extracted features are than sent to model for training or identification [12]. Speaker identification system agrees a human being according to his or her speech pattern there are two research themes in the area thereof. It is features extraction from the voice signal and their comparison [13]. Find the speaker who sounds closest to the test sample. When all speakers within a given set are known, it is called closed set. Speaker identification task can be divided in to two mainly process, which are the training and the testing process. The training process can also be separated in to four stages as.1.Inputting the speech signal, 2. Pre-processing of the speech, 3.normalization and 4. Feature extraction. Feature extraction is the most significant stage for speaker identification. Speaker identification systems can also be categorized into speech dependent systems and speech independent systems. Speech dependent system is a structure in which the identification tasks place on the basis of a particular text, where the people are required to read a particular text and on the basis of which the identification of the person is done [14]. In speech independent systems the speaker can say anything but the system still identifies the particular user. Developing a speech independent system is much challenging than a speech dependent system. Two leading stages are involved in speaker identification and they are feature extraction and feature classification. In feature extraction, certain feature crucial for identification of the person is extracted from the speech data. While in feature classification, the features of an unknown person is taken and is compared with the features of unlike speaker and thus identifies the exact speakers [15].

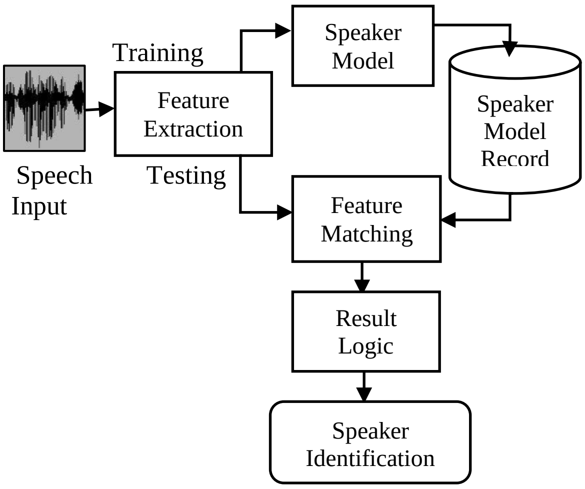

Fig.2.Block diagram of Speaker Identification model

5. FEATURE EXTRACTION

Feature extraction is the most essential process performed in all the SR systems in which unique vector are formed by sampling of speech waveform in time domain. Feature extraction plays a vital part in extracting the features from the boundless information containing voice signal which can be used for identifying the speaker among a group of N number of speakers. Feature extraction a crucial role in the success of any machine learning model. Any feature of speech that comprises segments layer than the phonetic segments is called a suprasegmentally feature. The feature extraction can be watchful as a data decrease process that tries to capture the essential characteristics of speaker with a small data rate. There are numerous techniques for extracting speech feature in the form of coefficients such as the mfcc, nonlinear energy operator, etc. Features including prosodic feature like pitch, energy zero crossing, spectral features also. Feature used for SR systems are mostly categorized in to acoustic, linguistic, context information and hybrid features and other features. The subsequent points should be considered while selecting features to be able to identify a speaker. a. Features should not be affected by a speaker's health and aging such as: Cold. Features should be difficult to mimic by others. c. Features should be independent of noise. The following features are act as a lead role in the speaker recognition systems. i. Frequency band analysis, ii. Formant frequencies, iii. Pitch contours, iv. Coarticalution etc. The significant features are: an MFCC-Mel frequency cepstral coefficient is developed by David and mermelstain. MFCC imitated the human speech production and reception system. MFCC is founded on the notorious deviations of ears of human. Critical band width with frequency, linearly spaced filters and log are at low and high frequencies respectively are used to phonetically key features of speech. Mel-Spectrogram-A spectrogram is the pictorial representation of a signal strength over time at different frequencies present in certain waveforms. It is a two dimensional view along with horizontally time and vertically with frequency. PNCC-Power normalized cepstral coefficients is great accuracy of ASR system straight in high noise background. PNCC is an acoustic feature which achieves the calculation by means of online algorithms in real time and delivers high accuracy even in noisy surroundings. Short Time Fourier Transform-STFT is a sequence of Fourier transforms of a windowed signal. It permits us to attain time frequency analysis. It is used to produce illustrations that capture mutually the local time and frequency content in the signal.

Spectral contrast- This feature provides a more detailed spectral spoof of a sound with respect to mfcc and spectrograms. Tonnetz- It represents a sound is similar to the Chroma grams with respect to the depiction of harmony and pitch classes. The method measures the tonal centroids of a sound, pitch space is called tonal centroid space. Linear prediction coefficients- To reduce the deconvolution complexity from the convolution process in speech signal, to find the time domain parameters, the linear prediction coefficient analysis is established. Teager energy operator-TEO provides a measure of the energy of a speech signal. Linear predictive cepstral coefficients-LPCC are used to capture emotion specific information revealed through vocal tract features. Cepstrum may be obtained using linear prediction analysis of a speech signal. Perceptual linear prediction-PLP is an extended version of linear prediction coefficients till fast Fourier transformation same as MFCC method. PLP is combination of concepts of LPCC and MFCC method. Bark frequency cepstral coefficients is the method to calculate band of frequency are almost 500 Hz and which is also in log illustration. Wavelet decomposition- The speaker signal features are decomposed by DWT features. It creates lower and higher frequency sub bands. The set of discrete value rules and translation scales are designed by wavelet transform. Power normalized cepstral coefficient –PNCC is one of the most recent technique with high accuracy. The high accuracy achieved by pncc is mainly attributed to the key features such as power law nonlinearity, symmetric noise suppression and temporal masking etc. Relative spectral transform Perceptual linear prediction- The basic principal behind the RASTA processing of speech is that human auditory system is relatively indifferent towards slowly varying spurs. Line Spectral frequencies-LSF is the individual lines of the line spectral pairs are known as LSF. It defines the two resonance situations taking place in the inter connected tube model of the human vocal tract. The model takes to consider the nasal cavity and mouth shape which gives the prediction illustration.

6. MODELLING TECHNIQUES

Gaussian mixture model- GMM is combination of Gaussian probability density function that are commonly used to multivariate data. Applying GMM to speaker modelling provides the speaker pdf, from which probability score can be obtained. Hidden markov model -HMM is models that is constructed on probability and uses markov processes that generates hidden and unknown parameters and reuse the parameters for analysis. DTW -Dynamic time warping is a method to identify a resemblance between two speech signals. It allows a machine to find out the best match between two signals and create a third signal. DTW is that it compares speech signals that are based on time. Vector quantization-VQ is a model that represents a larger amount of data into smaller data in the vector space and each region is called a cluster and then represented as a code word. VQ calculates the codebook with the smallest distortion and then identifies a speaker. i vector- The inconsistency space and hidden variables are calculated called total factors. The total factors are not observable, but can be estimated using factor analysis. These total factors than can be used as features to a classifier, and came to be known as I vector or identity vector. PLDA-PLDA was applied to compare i vectors. PLDA is a capable to be applied to any vectors.

Deep neural network - DNN are used from extracting features to complete end to end system speaker verification work. DNN involves of numerous hidden layers and fully connected feed forward network. These hidden layers are used to convert the input features vector in to probability distribution to estimate the output classes.

Convolution neural network -The name tells this network use the convolution operation as the key processing operation. It is a hierarchical neural network which consists of a variety of layers in sequence. CNN network came from three of concepts, which are sparse communication, limit sharing and equal representation. CRNN- Convolutional recurrent neural network model is a combination of a CNN and LSTM that exploits the spatial and temporal features of both networks. SVM-Support vector machine is belongs to supervised machine leaning. SVM can be used for classification and regression analysis. SVM are simply the coordinates of individual observation. SVM is a frontier which best segregates the two classes line and hyper plane. TDNN- Time delay neural network is introduced for ASR. It is purely DNN based type to recognize sequence of input vectors. The output of a unit at a time step depends on the previous layer with time interval including time step, with different time steps previous layer unit. Decision tree – Decision trees, or classification trees and regression trees, predict responses to the given data. To predict a response, we follow the decisions in the tree from the root. K-nearest neighbours – This is sometimes called straightforward extension of singleNN. Calculate the majority voting from the k nearest neighbour. FFNN-If there is no feedback from the outputs of the neurons towards the inputs throughout the network, then the network is referred as a feed forward neural network. MLP-Multilayer perceptron is the most known and most frequently type of neural network. On most occasions, the signals are transmitted within the network in one direction: from the input to output. There is no twist, the output of each neuron ensures not disturb the neuron itself.

7. SPEECH DATASETS

There are numerous speech databases that are created and used in the speaker recognition, identification and verification. The available corpuses are as followed. The most often used speech datasets for speaker recognition are: TIMIT Dataset-TIMIT contains broadband recordings of 630 speakers of eight major dialect of American English. Voxceleb-VoxCeleb is a large scale speaker identification dataset. It covers around 100000 utterances by 1251 peoples extracted from YouTube videos. LibriSpeech- LibriSpeech is a corpus of approximately 1000hrs of 16 kHz open English speech dataset. AudioMNIST- The dataset consists of 30000 audio samples of spoken digits of 60 different speakers.

VoxForge- VoxForge was set up to collect recorded speech for use with free and open source speech recognition engines. Common Voice- This dataset comprises hundreds of thousands of voice samples for voice recognition. It includes over 500 hrs of speech recordings. The Fisher- This dataset is used as the training corpora. It consists of 13k utterances with an average duration of utterances of around 5 mins. ELSDSR-English language speech database for speaker recognition database contains voice messages from 22 speakers with age of 25 to 65. TIDIGITS-TI digits database was developed for particularly designing and evaluating algorithms for speaker independent recognition of connected digit sequences. There are 326 speakers, each pronouncing 77 digit sequences. Berlin Dataset-The data set consists of expert annotated speech data from four different users. Every audio file is annotated using one of the seven different emotions including fear, anger, neutral, happy, sadness, boredom and disgust.

8. CONCLUSION

This paper concentrates on analysing and studying, various speaker identification modelling methods, feature extractions techniques and speech datasets. Further, in future work mainly focus on to incorporate some new skills with each other to expand the accuracy of speaker identification system such as LPC and DWT, MFCC with LPC and RNN were the growth can be occurs in this juncture that concentrated on shrinks the noise, number of features, removes irrelevant signals and redundant data and results in satisfactory recognition accurateness. Several datasets are used in speaker identification often. Mostly TIMIT and Librispeech are widely used in numerous identification systems with fine results.

REFERENCES

[1]. Rashid Jahangir, et.al "Text independent speaker through feature fusion and deep neural network", IEEE Access, Vol 8

[2]. Ting Lin and Ye Zhang, "Speaker recognition based on long term acoustic features with analysis sparse representation", IEEE Access, Vol 7.

[3]. Ahmed Isam Ahemd.et.al, "Speaker Recognition using PCA based feature transformation", Speech Communication 110, Elsevier.

[4]. Manish Gupta, et.al, "Gender based speaker recognition from speech signals using GMM model", Modern physics letters B, Vol 33.

[5]. Yanjie Jia, et.al, "Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network", Complex and intelligent systems, Springer.

[6]. Nguyen Nang An, et.al, "Deep CNNs with self-attention for speaker identification", IEEE access, Vol 7.

[7]. Xingmei Wang, et.al, "A network model of speaker identification with new feature extraction methods and symmetric BLSTM", Neuro computing 403, Elsevier.

[8]. Supapon Bunrit, et.al, "Text independent speaker identification using deep learning model of convolution neural network", IJMLC, vol9.

[9]. Ali Bou Nassif, et.al, "Speech recognition using deep neural networks: A systematic review", IEEE Access, Vol 7.

[10]. Arsha Nagrani.et.al, "Voxceleb: Large scale speaker verification in the Wild", Computer speech and language 60, Elsevier.

[11]. Zheli liu, et.al, "GMM and CNN hybrid method for short utterance speaker recognition", IEEE Transactions on industrial informatics.

[12]. Chao Zhang, et.al, "Depth wise separable convolutions for short utterance speaker identification", ITAIC, IEEE.

[13]. Yedilkhan amirgaliyev, et.al, "Development of speaker voice identification using main tone boundary statistics for applying to robot verbal systems", Intljournal of electronics and telecommunications.

[14]. Liyang Chen, et.al, "SpeakerGAN: Speaker identification with conditional generative adversarial network", Neuro computing 418.

[15]. Anett Antony, R.Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features", ICACC-2018, Elsevier 143.