A Study of Various Speech Features and Classifiers used in Speaker Identification

Priyatosh Mishra  
Electronics & Telecommunication Department  
RCET, Bhilai  
Bhilai (C.G.) India

Pankaj Kumar Mishra  
Electronics & Telecommunication Department  
RCET, Bhilai  
Bhilai (C.G.) India

Abstract - Speech processing consists of analysis/synthesis, recognition & coding of speech signal. The recognition field further branched to Speech recognition, Speaker recognition and speaker identification. Speaker identification system is used to identify a speaker among many speakers. To have a good identification rate is a prerequisite for any Speaker identification system which can be achieved by making an optimal choice among the available techniques. In this paper, different speech features & extraction techniques such as MFCC, LPCC, LPC, GLFCC, PLPC etc and different features classification models such as VQ, GMM, DTW, HMM and ANN for speaker identification system have been discussed.

Keywords- Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM), Vector Quantization (VQ), Hidden Markov Model (HMM), Artificial Neural Network (ANN)

I. INTRODUCTION

Biometric information is very popular means to distinguish between different persons. Visually, face is one most important feature, other features, like finger-prints, iris, are often used. Another way to identify a person is from the acoustic fact that every person has different voice features, this forms one area of speech processing, automatic speaker recognition. Speaker Recognition is used in security system for controlling the access to secure information from being accessed by anyone. It is the process of automatically recognizing the speaker voice according to the basis of information in the voice waves of Individual. It is a branch of biometric authentication which is gaining popularity as means of security measures due to its unique physical characteristics and identification of individuals. Automatic speaker recognition systems are categorized as two types – speaker verification & speaker identification. In speaker verification systems, a person claimed identity and is verified against that particular person’s database. On the other hand in speaker identification systems, a person is identified by matching his attributes with each template stored in the database. It is used to control access to services such as mobile banking, voice dialing, database access services or security control to a secured system.

Speaker identification systems are classified into two classes – text-dependent and text-independent. In text-dependent system, there is constraint on what is to be spoken i.e. predetermined group of words or sentences is used to enroll the speaker to the system and those words or sentences are used to verify the speaker. In Text-independent recognition systems there are no constraints on sentence.

Rest of the paper is structured as follows. Speaker identification process is explained briefly in section 2. Literature survey is given in section 3. Section 4 explains the techniques involved in feature extraction. Section 5 discusses and compares different classifier models and section 6 concludes the paper.

II. IDENTIFICATION SYSTEM OVERVIEW

Figure 1 shows an illustrative speaker Identification system. Pre-processing is the first step to make data ready for feature extraction. Feature extraction is carried out in order to reduces the dimension of the data and makes easy to process the speech data. The mostly used techniques for feature extraction are linear predictive cepstral coefficients, mel frequency cepstral coefficients. After the features are extracted the classifiers are used to model the feature data for different users. Whenever a person speaks in real time or testing time its features are extracted and model is made. The model is then compared with the database of speaker models and on the basis of match/mismatch of the models, the speaker is authenticated as valid or invalid speaker. The classifiers are categorized on basis of different models: Stochastic model (e.g. Gaussian Mixture Model (GMM), Hidden Markov Model (HMM)), Template model (Dynamic Time Warping (DTW), Vector Quantization (VQ)) Deterministic model (Support Vector Machines (SVM)) and Neural Networks (NN). The classifier algorithms can also be classified as supervised and unsupervised. In supervised learning algorithm, labels which are assigned to data are known before whereas in unsupervised learning algorithm, labels are not known.
III. REVIEW OF LITERATURE

- D.A. Reynolds [1] in year 1994 has compared MFCC, LPCC, LFCC, PLPC techniques with each other. MFCC proved to be a better technique than any other techniques for lower filter orders. PLPC and LPCC gave better performance with increasing filter order but performance degraded in linear coefficients (LFCC) because it gives equal detail to entire band of the signal and hence highlights the redundant information.

- Soma Swamy et al. [2] had implemented Speaker Identification using the Mel Frequency Cepstrum Coefficient feature and the Vector Quantization for Pattern Matching respectively and then it recognizes the spoken word using Linear Predictive Coding and Neural networks. The efficiency was 91% for speaker Identification and 98% for 3 words. LPC and Back Propagation with sigmoid function gave the efficiency of only 81% for three speakers.

- Lu Xiao-chun et al. [3] in 2012 introduced a new feature transformation method based-PPCA for text-independent speaker identification assuming that a subspace made with some of the principal components with larger contribution rates is recognized as phoneme dependent subspace and a subspace made without these components is considered as speaker-dependent subspace.

- Jianglin Wang et al. [4] had introduced two speaker-specific features for speaker identification. The paper introduced usage of two new features for speaker identification- Glottal Flow Cepstrum Coefficients (GLFCC) and Residual Phase Cepstrum Coefficients (RPCC). The standard GMM-UBM was used for modeling. The result of experiment shows that the proposed features give information about speaker characteristics that is notably different in nature from the phonetically-focused information exists in features such as MFCCs. These two new features give better results with lower model complexities.

- B. G. Nagaraja et al. [5] in year 2013 had compared the performance of four different windowing methods using MFCC for monolingual and cross-lingual speaker identification. The results indicate that window 3 and window 4 based system can be used for improving the identification performance.

- S. G Bagul et al. [6] 2013 had used Mel Frequency Cepstral Coefficients (MFCC’s) from the speech signal as feature extraction method and Gaussian mixture model (GMM) having Estimation and Maximization algorithm is used for modeling. The Gaussian mixture speaker model maintained very high identification performance with increasing population size. Their results indicate that Gaussian mixture models provide a powerful speaker representation for the difficult task of speaker recognition using corrupted, unconstrained speech.

- Shahzadi Farah et al. [7] implemented a speaker recognition system using MFCC, LPC as feature extraction and Vector Quantization as speaker classification technique. The performance of text-independent case was lower than text dependent. In the same manner the results of MFCC’s text-independent speaker reorganization are lower than the text-independent case. MFCC and VQ showed better accuracy as compared to system with LPC and VQ. The accuracy of speaker recognition decreased with increase of noise. The effect of pitch alteration results in lower classification accuracy.

- Sourjya Sarkar et al. [8] in year 2013 examined the performance of multilingual speaker recognition. Speaker identification and Speaker verification experiments were individually performed on 13 widely spoken Indian languages. The author focused on the effects of language mismatch in the recognition performance by working on individual languages and by taking all languages together. The standard GMM speaker recognition framework was used. The average language-independent speaker identification rate was as high as 95.21% and an average equal error rate of 11.71% showed good scope for improvement in speaker verification performance further.

- Sumit Srivastava et al. [9] in 2014 proposed Formant Based Linear Prediction Coefficient features for speaker identification for all type of environment. Gaussian Mixture Models (GMMs) was used for classification of speakers. The performance of Linear Prediction Coefficient features was computed and compared with the identification rate of Formant Based Linear Prediction Coefficient features. The performance of FBLPC features was found superior in both clean and noisy environment.
IV. FEATURES EXTRACTION TECHNIQUES

Feature extraction is the main part of speaker identification process. It is the speech features which distinguish a speaker among the other speakers. The different speech features & their extraction techniques are discussed below:

A. Pitch frequency & contour

Fundamental frequency of a speech signal is very important features, referred as pitch. Fundamental frequency (F0) refers to the rate at which the human vocal cords vibrate. Pitch detector provide necessary information about the nature of source of excitation for speech coding & the pitch contour of an utterance is very useful for recognizing speakers, emotion state determination, for voice activity detection etc. The variations of the fundamental frequency (pitch) over the duration of the utterance provide the contour that can be used as a feature for speaker identification. The speech utterance is then normalized and the contour is determined. The normalization of the speech utterance is required because the accurate time alignment of utterances is crucial; else the same speaker utterances could be treated as utterances from two different speakers. The contour is split into a set of segments and the measured pitch values are averaged over the whole segment. The vector that contains the average values of pitch of all segments is then used as a feature for speaker identification.

B. Formant Frequencies

Periodic excitation is seen in the spectrum of sounds, especially in sound of vowels. The speech organs form certain indefinite shapes to produce the vowel sound and so regions of resonance and anti resonance are developed in the vocal tract. Location of these resonances in the frequency spectrum depends on the shape of the vocal tract. As the physical structure of the speech organs is a representative of every individual speaker, differences among the speakers can also be found in the position of their formant frequencies. The resonance heavily affects the overall spectrum shape and is referred as formants. Some of these formant frequencies can be sampled at convenient rate and used for speaker recognition. These features are commonly used in combination with other features.

C. Average Zero Crossing Rate

A zero crossing is occurred in a signal when its waveform changes its algebraic sign. For a discrete time signal with zero crossing rates (ZCR) in sample and a sampling frequency of Fs, the frequency F0 is given by

\[ F_0 = \frac{ZCR \times F_s}{2} \]

The speech contains most of its energy in voiced signals at lower frequencies. For the unvoiced sounds, the broadband noise excitation takes place at the higher frequencies due to short dimension of the vocal tract. Therefore a high and a low ZCR is related to unvoiced and voiced speech respectively.

D. Frequency Band Analysis

Filter banks were initially used to gather information about the spectral structure of signal. The filter banks consist of number of filters where each filter covers one group of frequencies. Bandwidths of filters could be chosen to be equal, Logarithmic or may correspond to certain critical intervals. The output of such filter bank offers largely depends upon the number of the filters being used, which normally varies from 10-20 and thus this technique represents an approximate representation of the actual spectrum. The output of the filter bank is sampled (usually 100 Hz) and the samples of the output indicate the amplitude of the frequencies from a particular bandwidth. The output is thus used as the feature vector for speaker identification.

E. Mel Frequency Cepstral Coefficients (MFCCs)

MFCCs are considered to be low-level information. This feature is based on spectral information derived from a short time-window segment of speech of about 20 millisecond. MFCCs are represented by a real valued N-dimensional vector. The coefficients are parameters of the spectrum which have some dependency on the physical characteristics of the speaker. The MFCC feature is derived directly from the FFT power spectrum. The mel-scale filter bank centers and bandwidths are fixed to follow the mel scale, giving more detail to low frequencies.
G. Linear Prediction-based Cepstral Coefficients (LPCCs)

LPCCs are often used in speaker recognition systems, although their sensitivity to noisy environments have made them more undesirable as speaker recognition systems are applied to the more challenging channels. Just like MFCCs, the LPCC processing uses a fixed analysis window of 20 milliseconds and are of the continuous measurement type. LPCCs are dependent on the spectral envelope and are considered to be low-level information. Linear prediction based features use an all pole model to represent the smoothed spectrum. The LPCC can be considered as having adaptive detail in that the model poles move to fit the spectral peaks wherever they occur. The details are limited by the number of poles available. However LPCC has also inheriting the disadvantages from LPC.

H. Perceptual linear prediction cepstral coefficient (PLPC)

The PLPC features are hybrid between the all-pole model and filter bank spectral representation. The spectrum is first passed through a bark-spaced trapezoidal-shaped filter bank and then adapted with an all pole model. The detail of the PLPC representation is determined by both the filter bank and all-pole model order.

I. Glottal Flow Cepstral Coefficients (GLFCC)

The glottal flow is the airflow arising from the trachea and passing by the vocal folds. There are many reasons for the glottal flow to be speaker specific. Videos of vocal fold vibration show large variations in the movement of the vocal folds from one speaker to another. For some individuals the vocal folds never close fully and in other cases vocal folds close completely and rapidly. The closing manner of vocal fold vibrations is also speaker dependent. The closure of vocal folds for some individuals shows zipper-like patterns, while others close on the length of the vocal folds about the same time. In addition, the configuration of the opening area shows differences for different individuals. The glottal opening for some individuals is nearly equal in width along the glottis length, such as pressed phonation. For some individuals, a more triangular shaped opening will occur according to their own anatomical structure of vocal folds. Because of this the glottal flow contains speaker specific information. The features derived from glottal flow are useful for speaker identification.

V. MODELS AND CLASSIFIERS

There are two types of modeling methods in the realm of speaker recognition: The deterministic methods (Vector Quantization and Dynamic Time Warping) and statistics methods (Gaussian Mixture Model, Hidden Markov Model and Artificial neural network).

A. Vector quantization

Vector quantization (VQ) is a classical quantization technique in signal processing. It was used for data compression. It works by dividing a large set of points called vectors into groups having approximately the same
number of points nearest to them. Each group is characterized by its centroid point. The density matching property of VQ is very powerful for high-dimensioned data. Hence VQ is suitable for lossy data compression. It can also be used for density estimation and lossy data correction. The method most commonly used to generate codebook is the Linde-Buzo-Gray (LBG) algorithm.

B. Dynamic Time Warping
This algorithm specifically deals with variance in speaking rate and variable length of input vectors. It calculates the similarity between two sequences which may differ in time or speed. To normalize the differences between timing of test utterance and the reference template time warping is executed non-linearly in time dimension. Time normalized distance is computed between the patterns after time normalization. The speaker with minimum time normalized distance is identified as authentic speaker.

C. Hidden Markov Modeling (HMM)
A Hidden Markov Model (HMM) consists of a set of transitions between a set of states. Two sets of probabilities are defined for each transition: a transition probability and the output probability density function (PDF). The output probability density function is the probability of emitting each of the output symbols from a finite vocabulary. The transitions are allowed to the next right state or the same state, thus the model is named left to right model. The HMM parameters are generated from the speech during the training phase and for verification, the likely hood of the input feature sequence is computed with respect to the speakers HMMs. In case of finite vocabulary being used for speaker identification, each word is modeled using multiple states left to right HMMs. Therefore in case of large vocabulary, larger number of models is required. The first step in HMM modeling is to form a representation of the impostors. Here the concept of the background model is to form a model of the world of all possible speakers. HMM background models can then be trained through the use of a full large vocabulary continuous speech recognition.

D. Gaussian Mixture Models GMM
The distribution of feature vectors extracted from a speaker’s speech is modeled using a GMM i.e. Gaussian Mixture Model. The output density of GMM is a linear combination of M components called mixtures.

\[ p(x | \lambda) = \sum_{i=1}^{M} W_i \times g(x | \mu_i , \Sigma_i) \]  

(5)

Where \( x \) is a D-dimensional continuous-valued data vector (or features), \( W_i, i = 1, \ldots, M \), are the mixture weights, and \( g(x | \mu_i , \Sigma_i) \), \( i = 1, \ldots, M \), are the Gaussian densities components. Each component density is a D-variate Gaussian function of the form,

\[ g(x | \mu , \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

(6)

with mean vector \( \mu \) and covariance matrix \( \Sigma \). The mixture weights satisfy the constraint that \( \sum_i W_i = 1 \) each mixture is characterized by weight of mixture, a mean vector and covariance matrix. These parameters are collectively represented by the notation \( \lambda \).

\[ \lambda = \{ W_i, \mu_i, \Sigma_i \} \quad i = 1, \ldots, M \]

A speaker identification system must be capable to estimate probability distributions of the computed feature vectors. The storage of every single vector that generates from the training mode is impossible, as these distributions are defined on a high-dimensional space. Vectors from a large vector space can be mapped to a finite number of regions in the space known as clusters. Cluster analysis creates various clustering models to form an efficient speaker model. The best known models for speaker identification are centroid models that involve clustering algorithms like k-means algorithm, LBG algorithm, fuzzy c-means algorithm and expectation maximization algorithm.

E. Artificial Neural Networks
An Artificial Neural Network is mathematical model that tries to simulate the structure and functions of biological neural networks. The basic building block of every Artificial Neural Network is an artificial neuron which is nothing but a mathematical model. Such a model has three simple sets of rules: multiplication, summation and activation. At the input of artificial neuron the inputs are weighted which means that every input value is multiplied with individual weight. The middle section of artificial neuron is sum function that sums all weighted inputs and bias. At the output of artificial neuron the sum of weighted inputs and bias passes through activation function which is also called as transfer function.
Back Propagation Neural Networks

Back propagation network is the most widely used artificial neural networks. It is multilayered network which includes at least one hidden layer. First the input is propagated forward through the network to get the output from the output layer. Then, the sensitivities are propagated backward to decrease the error. During this process, weights in all hidden layers are modified. As the propagation continues, the weights are constantly adjusted and the precision of the output is improved.

1) Radial Basis Function (RBF) Neural Networks

RBF networks have three layers:

a) Input layer – There is one neuron in the input layer for each of the predictor variable. In the case of categorical variables, N-1 neurons are used where N is the number of categories. The input neurons normalize the range of the values by subtracting the median and dividing by the interquartile range. The input neurons then pass the values to each of the neurons in the hidden layer.

b) Hidden layer – The number of neuron in this layer are variable (the optimal number is determined by the training process). Each neuron consists of a RBF which is centered on point with as many dimensions as there are predictor variables. The spread (radius) of the RBF function may be different for each dimension. The centers and radius are determined by the training process. When presented with the x vector from the input layer, a hidden layer neuron computes the Euclidean distance of the test case from neuron’s center point and then the Radial Basis Function kernel function is applied to this distance using the spread values. The resulting output value is passed to the the summation layer.

c) Summation layer – The output of a neuron in the hidden layer is multiplied by the weight related with the neuron and passed to the summation which sum up the weighted values and presents this sum as the output of the network. A bias value of 1.0 that is multiplied by a weight W0 is fed into the summation layer. For the classification problems, there is one output and independent set of weights and summation unit for each target category.

VI. CONCLUSIONS

In this paper we have shown an overview of speaker identification techniques. In section 4, we presented some of the features of speech that can be extracted to identify the speaker. More precise measurements and methods for extracting these features will lead to more accurate and diverse speaker recognition systems. In Section 5, we described classification methods that are currently being studied in research and used in application.

ACKNOWLEDGMENT

I would like to thank my supervisor Mr. Pankaj Kumar Mishra, Associate & Head of Department of Electronics & Telecommunication Engineering of RCET, Bilai (C.G.) for his immense support and enlightened guidance for this work. I am very grateful for the inspiring discussions with all my faculties. Their valuable support and path-guiding suggestions have helped a lot. I am thankful to Prof. Nitin Nair (PG Coordinator) for giving thoughtful suggestions during my work.

REFERENCES