Study of Speaker Verification Methods

Sneha M.Powar
Electronics & Telecommunication Department
Dr. J. J. Magdum College of Engineering, Jaysingpur
Kolhapur, India
Email: powarsneha27@gmail.com

Dr. V.V.Patil
H.O.D. Electronics Department
Dr. J. J. Magdum College of Engineering
Jysingpur, India
Email: vvpatil2429@gmail.com

Abstract— Speaker verification is a process to accept or reject the identity claim of a speaker by comparing a set of measurements of the speaker's utterances with a reference set of measurements of the utterance of the person whose identity is claimed.. In speaker verification, a person makes an identity claim. There are two main stages in this technique, feature extraction and feature matching. Feature extraction is the process in which we extract some useful data which can later to be used to represent the speaker. Feature matching involves identification of the unknown speaker by comparing the feature extracted from the voice with the enrolled voices of known speakers.

Keywords- *speaker verification, text dependent, text independent*

*****

I. Introduction

In our everyday lives there are many forms of communication, for instance: body Language, textual language, pictorial language and speech, etc. However amongst those forms speech is always regarded as the most powerful form of communication. From the signal processing point of view, speech can be characterized in terms of the signal carrying message information. The waveform could be one of the representations of speech, and this kind of signal has been most useful in practical applications. Extracting from speech signal, we could get three main kinds of information: Speech Text, Language**,** and Speaker Identity. Speech recognition refers to the ability of a machine or program to recognize or identify spoken words and carry out voice. The spoken words are digitized into sequence of numbers, and matched against coded dictionaries so as to identify the words. Speaker recognition maybe defined as the process of recognizing a person automatically using the information extracted from speech signal of the person. This technique uses the voice of the speaker to verify their identity to access to several services such as accessing the computer or server from remote place,
voice dialing, accessing security services, mobile banking etc. where security is the primary concern.[6]

II. SPEAKER INFORMATION IN SPEECH SIGNAL

Speech is human being's primary means of communication, and it contents essentially the meaning of information from a speaker to a hearer, individual information representing speaker's identity and gender, and also sometimes the emotions. In a speech production the properties of both articulators, which produce the sound, and auditory organs, which perceive the sound, should be involved in Speech production can be divided into three principal components: excitation production, vocal tract articulation, and lips' and/or nostrils' radiation. Excitation powers the speech production process. It is produced by the airflow from lungs, and then carried by trachea through the vocal folds, during inspiration, air is filled into lungs, and during expiration the energy will be spontaneously released. The trachea conveys the resulting provider to serve inputs to the vocal tract, and the volume of air determines the amplitude of the sound. The vocal tract works as a filter to shape the excitation sources. The uniqueness of speaker voice not only depends on the physical features of the vocal tract, but the speaker's mental ability to control the muscles of the organs in the vocal tract. It is not easy for speaker to change the physical features intentionally. However, these physical features are possible to be changed with ageing. Speaker characteristics in the speech signal are often difficult to carry out. Segmenting, labeling, and measuring specific segmental speech events that characterize speakers, such as nasalized speech sounds, is difficult because of variable speech behavior and variable and distorted recording and transmission conditions. Overall qualities, such as breathiness, are difficult to correlate with specific speech signal measurements and are subject to variability in the same way as segmental speech events. The most important analysis tool is short-time spectral analysis. It is no coincidence that short-time spectral analysis also forms the basis for most speech recognition systems. Short-time spectral analysis not only resolves the characteristics that differentiate one speech sound from another, but also many of the characteristics already mentioned that differentiate one speaker from another. There are two principal modes of short-time spectral analysis: filter bank analysis and linear predictive coding (LPC) analysis. In filter bank analysis, the speech signal is passed through a bank of band pass filters covering the available range of frequencies associated with the signal. Typically, this range is 200 to 3,000 Hz for telephone band speech and 50 to 8,000 Hz for wide band speech. A typical filter bank for wide band speech contains 16 band pass filters spaced uniformly 500 Hz apart. The output of each filter is usually implemented as a windowed, short-time Fourier transform [using fast Fourier transform (FFT) techniques] at the center frequency of the filter. LPC-based spectral analysis is widely used for speech and speaker recognition. The LPC model of the speech air stream to the larynx. Larynx refers as an energy signal specifies that a speech sample at time *t* ,*s* .*t/*, can be represented as a linear sum of the *p* previous samples plus an excitation term, as follows:

$S(t)= a_1s(t – 1) +a_2s(t – 2)+…….a_p(t – p)+ G\ u(t)$

The LPC coefficients a_i are computed by solving a set of linear equations resulting from the minimization of the mean-squared error between the signal at time *t* and the linearly predicted estimate of the signal. Two generally used methods for solving the equations, the autocorrelation method and the covariance method. [4]

III. SPEAKER RECOGNITION

A. *Speaker Identification*

Speaker identification (SI) is the process of finding the identity of an unknown speaker by comparing his/her voice with voices of registered speakers in the database. It's a one-to-many comparison. The basic structure of *SI* system (SIS) is shown in Figure. We notice that the core components in SIS are the same as in SVS. In SIS, M speaker models are scored in parallel and the most-likely one is reported. The core components in *SIS* are the same as in SVS. In SIS, M speaker models are scored in parallel and the most-likely one is reported, and consequently decision will be one of the speaker's ID in the database, or will be 'none of the above' if and only if the matching score is below some threshold and it's in the case of a open-set SIS.

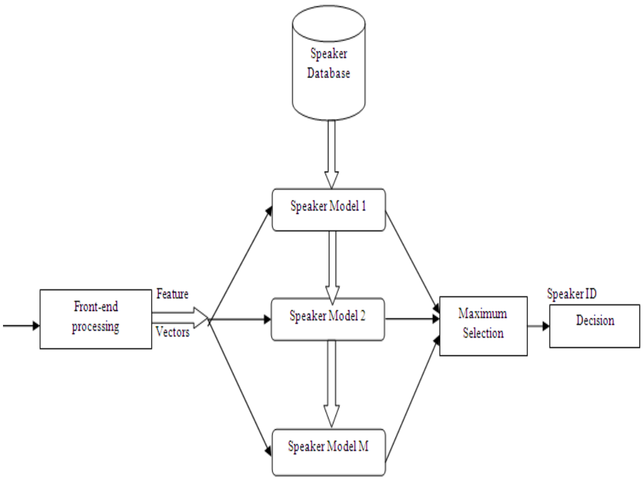

Fig 1. Basic structure of Speaker Identification

B. Speaker verification

Speaker verification (SV) is the process of determining whether the speaker identity is who the person claims to be. Different terms which have the same definition as SV could be found in literature, such as voice verification, voice authentication, speaker/talker authentication, talker verification. It performs a one-to-one comparison (it is also called binary decision) between the features of an input voice and those of the claimed voice that is registered in the system. Three main components shown in this structure are: Front-end Processing, Speaker Modeling, and Pattern Matching. To get the feature vectors of incoming voice, front-end processing will be performed, and then depending on the models used in Pattern Matching, match scores will be calculated. If the score is larger than a certain threshold, then as a result, claimed speaker would be acknowledged. There are three main components: Front-end Processing, Speaker Modeling, and Pattern Matching. Front-end processing is used to highlight the relevant features and remove the irrelevant ones. After the first component, we will get the feature vectors of the speech signal. Pattern Matching between the claimed speaker model registered in the database and the imposter model will be performed then, if the match is above a certain threshold, the Identity claim is verified. Using a high threshold, system gets high safety and prevents impostors to be accepted, but in the mean while it also takes the risk of rejecting the genuine person, and vice versa. [6]

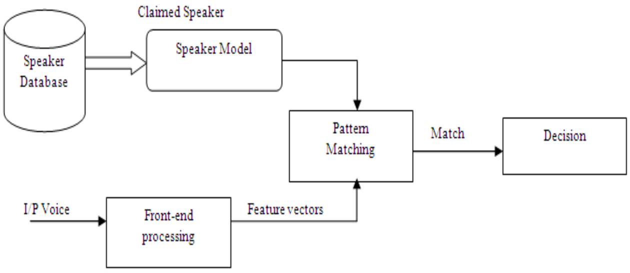

Fig. 2 Basic Structure of Speaker Verification

IV. METHODS OF SPEAKER VERIFICATION

Speaker verification systems typically operate in one of two input modes: text dependent or text independent. In the text-dependent mode, speakers must provide utterances of the same text for both training and recognition trials. In the text-independent mode, speakers are not constrained to provide specific texts in recognition trials. Since the text-dependent mode can directly exploit the voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent mode.

A. *Text-Dependent (Fixed Passwords)*

The structure of a system using fixed passwords is rather simple; input speech is time aligned with reference templates or models created by using training utterances for the passwords. If the fixed passwords are different from speaker to speaker, the difference can also be used as additional individual information. This helps to increase performance. The most common approach to automatic speaker recognition in the text-dependent mode uses representations that preserve temporal characteristics. Each speaker is represented by a sequence of feature vectors (generally, short-term spectral feature vectors), analyzed for each test word or phrase. This approach is usually based on template matching techniques in which the time axes of an input speech sample and each reference template of registered speakers are aligned, and the similarity between them accumulated from the beginning to the end of the utterance is calculated. Trial-to-trial timing variations of utterances of the same talker, both local and overall, can be normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a dynamic programming (DP) time warping algorithm or DTW. Since the sequence of phonetic events is the same for training and testing, there is an overall similarity among these sequences of feature vectors. Ideally the intra-speaker differences are significantly smaller than the inter-speaker differences.

B. Text Independent (No Specified Passwords)

There are several applications in which predetermined passwords cannot be used. In addition, human beings can recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have recently been actively investigated. Another advantage of text-independent recognition is that it can be done sequentially, until a desired significance level is reached, without the annoyance of having to repeat passwords again and again. In a text-independent system, the words or phrases used in recognition trials generally cannot be predicted. Therefore, it is impossible to model or match speech events at the level of words or phrases. Classical text-independent speaker recognition techniques are based on measurements for which the time dimension is collapsed. Recently text-independent speaker verification techniques based on short duration speech events have been studied. The new approaches extract and measure salient acoustic and phonetic events. The bases for these approaches lie in statistical techniques for extracting and modeling reduced sets of optimally representative feature vectors or feature vector sequences or segments. These techniques fall under the related categories of vector quantization (VQ), matrix and segment quantization, probabilistic mixture models, and HMM.

A set of short-term training feature vectors of a speaker can be used directly to represent the essential characteristics of that speaker. However, such a direct representation is impractical when the number of training vectors is large, since the memory and amount of computation required become prohibitively large. Therefore, efficient ways of compressing the training data have been tried using VQ techniques. In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific feature. A speaker specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker, and the VQ distortion accumulated over the entire input utterance is used in making the recognition decision.

A five-state ergodic linear predictive HMM is used for broad phonetic categorization. After identifying frames belonging to particular phonetic categories, feature selection is performed. In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores for each category. The weights are chosen to reflect the effectiveness of particular categories of phonemes in discriminating between speakers and are adjusted to maximize the verification performance. The performances of speaker recognition based on a VQ-based method and that using discrete/continuous ergodic HMM-based methods have been compared, in particular from the viewpoint of robustness against utterance variations. It was shown that a continuous ergodic HMM method is far superior to a discrete ergodic HMM method, and that a continuous ergodic HMM method is as robust as a VQ-based method when enough training data is available. However, when little data is available, the VQ-based method is more robust than a continuous HMM method.[6]

V. APPLICATION

A. Secure Access via telephone

The most straightforward way to employ SV is in the cases when one has to gain access to some secure place via telephone. Voice is completely compatible with the existing transmission protocols via telephone channels; therefore no special adaptations of the system (besides the installment of a SV system) are necessary.

B. Home Banking

Home banking is another application where SV can be applied. For the time being such a service is restricted to operations within the accounts maintained by a single individual. One can for e.g. check the status of their account, transfer money between one’s own saving accounts, etc. The security is pretty low in these cases, the users are verified only by saying their PIN and FR almost never occurs (after all who wants to play \robber" with his own savings!). Still, however, it is being researched how secure it could be to use SV for transactions including a second and third party (i.e. the so called high-risk bank transactions). It is always noted that the security measures should be proportional to the value that could be obtained by this service.

C. *Home shopping*

Home shopping (see for e.g. http://www.hsn.com) is the service that is most uninteresting to an imposter. SV is here being employed, though backed up by a human operator. In this service people ring to order products that are later on shipped to their home addresses. In cases when all lines are busy, a customer can always choose to use the automatic service. They just have to speak their telephone number and if their identity is successfully verified they can start ordering products. If they are rejected, they are redirected to a human operator. But even if their identity is mistaken for someone else and some products are sending to another customer, there is no harm because these products cannot go to an unauthorized party (i.e. a criminal).

D. *Forensics and Survelliance*

Detection of speakers in forensic cases boils down, in most situations, to deciding whether a given recording is really from a suspect or not. This is exactly the case of the Hypothesis Test. Leaving aside legal issues SV can help police discover how many different individuals are involved in a conversation on a tape.

VI. DISCUSSION

The challenges for implementing practical and uniformly reliable systems for speaker verification, are rooted in problems associated with variability and insufficient data. As described earlier, variability is associated with trial-to-trial variations in recording and transmission conditions and speaking behavior. The most serious variations occur between enrollment sessions and subsequent test sessions resulting in models that are mismatched to test conditions. Most applications require reliable system operation under a variety of environmental and channel conditions and require that variations in speaking behavior will be tolerated. Insufficient data refers to the unavailability of sufficient amounts of data to provide representative models and accurate decision thresholds.

REFERENCES

[1] Matsui, T. and Furui, S., Concatenated phoneme models for text-variable speaker recognition, *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing,* II, 391–394, 1992.

[2] Matsui, T. and Furui, S., Speaker adaptation of tied-mixture-based phoneme models for text prompted speaker recognition, *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, I, 125–128, 1994

[3] Higgins, A.L., Bahler, L.andPorter, J., Speaker verification using randomized phrase prompting, *Digital Signal Processing*, 1, 89–106, 1991.

[4] Rabiner, L.R. and Juang, B.-H., *Fundamentals of Speech Recognition*,Prentice-Hall Englewood Cliffs, NJ, 1993.

[5] Rosenberg, A.E., Lee, C.-H. and Gokcen, S., Connected word talker verification using whole word hidden Markov models, *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Toronto, 381–384, 1991.

[6] Technical University of Denmark Informatics and Mathematical Modelling Master thesis by Ling Feng IMM-THESIS: ISSN 1601-233X

[7] R. A. Cole and colleagues, “Survey of the State of the Art in Human Language Technology”, National Science Foundation European Commission.

BIOGRAPHY

Miss S.M.Powar received the BE degree in 2011 in Electronics from Bharti Vidyapeeths College Of Engg Kolhapur & persuing ME Electronics & Communication in Dr. J.J.Magdum College of Engg. Jaysingpur.

Dr.Mrs.V.V.Patil received the BE degree in1994 & ME degree in 2004 in Electronics Engg. Dept. of Walchand College of Engg. sangli & PHD degree from Electrical Engg. Dept. of I.I.T Bombay in 2014.
She is currently professor & Head in Electronics Engg. Dept. of Dr. J.J.Magdum College of Engg. Jaysingpur.
Her research interests are in the area of speech & signal processing applications.