International Journal of Future Generation Communication and Networking
Vol. 13, No. 2s, (2020), pp. 324–329

Text-Independent Speaker Identification: A Survey

Rupesh Neve[1],N.M.Wagdarikar[2],Mohit Patil[3],Paras Agrawal[4]
[1,2,3,4] *Dept. of E & TC Engg., Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune*
[1]*rupeshneve123@gmail.com*
[2]*narendradsp@rediffmail.com*
[3]*mp351759@gmail.com*
[4]*agrawalmohit1997@gmail.com*

Abstract

Nowadays it is obvious that speakers can be identified from their voices. In this work the details of speaker identification from the real-time system point of view are discussed. The speaker identification systems can be subdivided into text-dependent and text-independent methods. Text-dependent systems require the speaker to utter a specific phrase (pin-code, password etc.), while a text-independent method should catch the characteristics of the speech irrespective of the text spoken. The system developed in this work is the latter, text-independent, meaning the system can identify the speaker regardless of what is being said. This paper presents text-independent speaker identification system, which consists of mapping a speech signal from an unknown speaker to a database of known speakers, i.e. the system hasbeen trained with a number of speakers which the system can recognize. Speaker identification has been done with hidden Markova model and vector quantization.

I.INTRODUCTION

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different . In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and soon. An important application of speaker recognition technology is forensics. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in recent years there has been increasing interest to integrate automatic speaker recognition to supplement auditory and semi-automatic analysis methods. It is also used in telephone based services with integrated speech recognition.

Speaker recognition is the state of the art technique for identifying the person based on the preferences. HMM outperforms neural networks and SVM and hence HMM is preferred for various applications. This includes voice operated home appliances, audio system for cars, wheel-chair, cordless/mobile phones, and robots. The feature extraction algorithms have greater impact on the performance of the speech recognition system. Mel frequency cepstral coefficients and the cochlear filter banks are the possible algorithms for the ASR. In order to cater to the various applications, an isolated speech recognition system is proposed using HMM in this report. Two approaches of the feature extraction are compared to suggest the best suitable feature extraction algorithm depending on the operating environment

ISSN: 2233-7857 IJFGCN

II.LITERATURE SURVEY

Speaker recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of sub word units (e.g., phonemes), words, phrases, and sentences. A text-independent method should catch the characteristics of the speech irrespective of the text spoken. The system developed in this work is the latter, text-independent, meaning the system can identify the speaker regardless of what is being said. This paper presents text-independent speaker identification system, which consists of mapping a speech signal from an unknown speaker to a database of known speakers, i.e. the system has been trained with a number of speakers which the system can recognize.

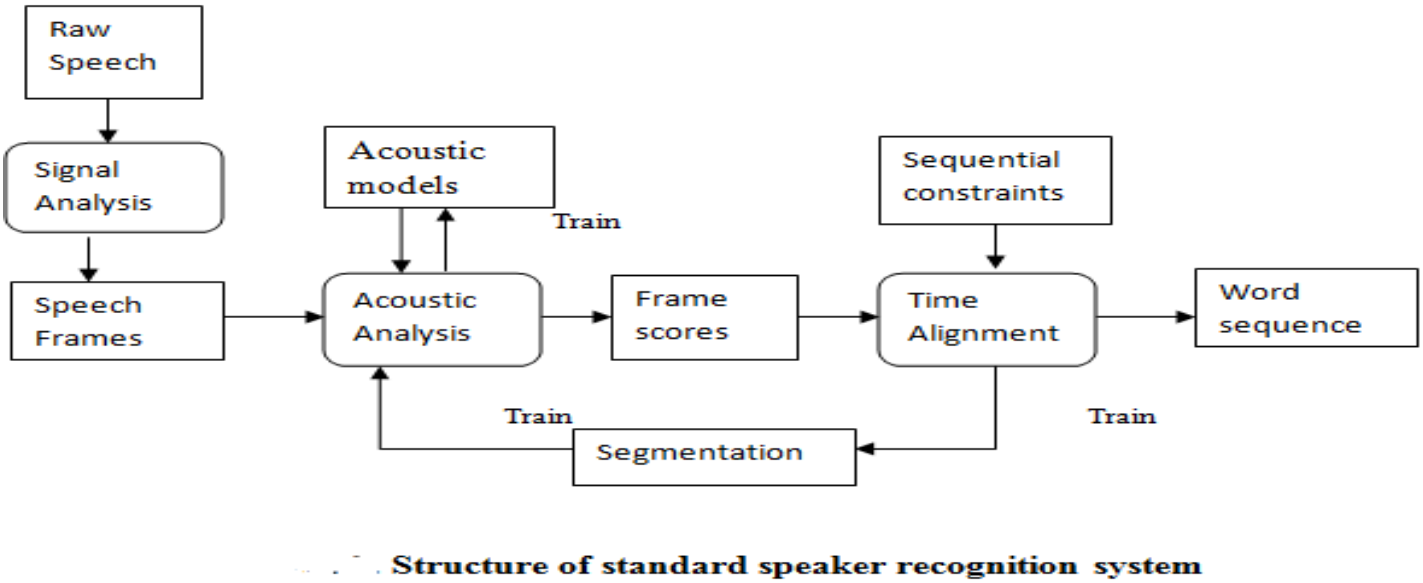

Figure 1: Standard speaker recognition system

- Raw speech**:-** Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time.
- Signal analysis**:-** Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information [11].
- Speech frames:- The result of signal analysis is a sequence of speech frames, typically at 10 msec intervals, with about 16 coefficients per frame. These frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics; this typically leads to improved performance. The speech frames are used for acoustic analysis [12].
- Acoustic models**:-** In order to analyze the speech frames for their acoustic content, we need a set of acoustic models. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties

A. Feature Extraction Techniques

Feature extraction is one of the most important steps required for representation and recognition of the speech signals. Feature extraction using Linear Predictive Coding (LPC) is based on the speech

ISSN: 2233-7857 IJFGCN

production model. Feature extraction using Mel frequency cepstral coefficients (MFCC) are based on crude approximation of human peripheral auditory.

Mel Frequency Cepstral Coefficients

Speech in the time-domain is obtained by sampling the real signal wave-form. The waveform can be converted to a spectrogram by taking a Fourier transform. For a sampling frequency of 8000 this results in 129 frequency bins for each frame. Each complex number is converted to a real number by taking the magnitude. The spectrogram represents the power of different frequency bands over time (usually on a log scale). The spectrograms extracted in this work have a window size of 256 samples which corresponds to about 32 ms (256 divided by a sampling frequency of 8000). The windows are half-overlapping and the 'Hamming' windowing function is used [15].

Linear Predictive Cepstral Coefficients

The basic idea behind the LPC analysis is that a speech sample can be approximated as linear combination of past speech samples. The procedure for LPC feature extraction of the input speech is explained in this section. The basic steps for feature extraction includes pre-emphasis, frame blocking, windowing and autocorrelation

B. Feature Classification Techniques

There are a lot of approaches to speech recognition. Algorithms and feature extraction are based on the acoustic-phonetic approach. Algorithms such as template matching come under the pattern recognition approach, while algorithms that depend on knowledge sources, stochastic of speech signals and neural networks are based on the artificial intelligence approach

a) Acoustic Phonetic Approach

The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach, which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterizedby a set of acoustics properties that are manifested in thespeech signal over time. Even though, the acoustic propertiesof phonetic units are highly variable, both with speakers andwith neighboring sounds (the so-called co articulation effect),it is assumed in the acoustic-phonetic approach that the rulesgoverning the variability are straightforward and can be readily learned by a machine.

b) Pattern Recognition Approach

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns [21].

ISSN: 2233-7857 IJFGCN

A. Template Based Approach

Template-based speech recognition systems have a database of prototype speech patterns (templates) that define the vocabulary. The generation of this database is performed during the training mode. During recognition, the incoming speech is compared to the templates in the database, and the template that represents the best match is selected. Since the rate of human speech production varies considerably, it is necessary to stress or compress the time axes between the incoming speech and the reference template. This can be done efficiently using Dynamic Time Warping (DTW). In a few algorithms, like Vector Quantization (VQ), it is not necessary to vary the time axis for each word, even if any two words have different utterance length. This is performed by splitting the utterance into several different sections and coding each of the sections separately to generate a template for the word.

B. Stochastic Approach

Stochastic modeling entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability s, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state Markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variability's, while the parameters in the output distribution model, spectral variability's. These two types of variability's are the essence of speech recognition.

C. Hidden Markov Model

The basic theoretical strength of the HMM is that it combines modeling of stationary stochastic processes (for the short-time spectra) and the temporal relationship among the processes (via a Markov chain) together in a well-defined probability space. This combination allows us to study these two separate aspects of modeling a dynamic process (like speech) using one consistent framework. Another attractive feature of HMM's comes from the fact that it is relatively easy and straightforward to train a model from a given set of labeled training data (one or more sequences of observations).

III.METHODOLOGY

A. Database

In this system we are going to use database available in TIMIT only for testing the accuracy purpose and for recognition purpose we are going to use real time database such as utterances of different peoples who can utter any word. Some of available database used for testing is as follows. The systems are evaluated on four publically available speech databases: TIMIT, NTIMIT, Switchboard and YOHO. The different levels of degradations and variability found in these databases allow the examination of system performance for different task domains.

B. Feature extraction technique

ISSN: 2233-7857 IJFGCN

Mel Frequency Cepstral Coefficients: Speech in the time-domain is obtained by sampling the real signal wave-form. The waveform can be converted to a spectrogram by taking a Fouriertransform. For a sampling frequency of 8000 this results in 129 frequencybins for each frame. Each complex number is converted to a real numberby taking the magnitude. The spectrogram represents the power of differentfrequency bands over time (usually on a log scale). The spectrograms extracted in this work have a window size of 256 samples which corresponds toabout 32 ms (256 divided by a sampling frequency of 8000). The windowsare half-overlapping and the 'Hamming' windowing function is used [15].

Linear Predictive Cepstral Coefficients: The basic idea behind the LPC analysis is that a speech sample can be approximated as linear combination of past speech samples. The procedure for LPC feature extraction of the input speech is explained in this section. The basic steps for feature extraction includes pre-emphasis, frame blocking, windowing and autocorrelation. LPCC are calculated as follows.

C. Feature classification techniques

Hidden Markov Model

The basic theoretical strength of the HMM is that it combines modeling of stationary stochastic processes (for the short-time spectra) and the temporal relationship among the processes (via a Markov chain) together in a well-defined probability space. This combination allows us to study these two separate aspects of modeling a dynamic process (like speech) using one consistent framework. Another attractive feature of HMM's comes from the fact that it is relatively easy and straightforward to train a model from a given set of labeled training data (one or more sequences of observations).
As mentioned above the technique used to implement speech recognition is Hidden Markov Model (HMM).The HMM is used to represent the utterance of the word and to calculate the probability of that the model which created the sequence of vectors. There are some challenges in designing of HMM for the analysis or recognition of speech signal. HMM broadly works on two phases under which phase I is Linear Predictive Coding and phase II consists of Vector Quantization, training, and recognition phases.

IV.CONCLUSION

Speaker recognition is more challenging due to dynamic and complex structure of the speech signal. Presence of different noises may worsen the performance of the speech recognition system. Also, various approaches of the feature extraction and feature classification affect the recognition accuracy of the speech recognition system.The comparative analysis of the Mel frequency cepstral coefficients and LPCC approach of the feature extraction is done in the report. The system gives the satisfactory performance for the MFCC in clean environment The testing of the system with the different databases including Digit and Alphabet database shows the robust nature of the system with different operating environment.The system is text independent. The whole system can be transferred to the real time application with much better performances for the noisy conditions.

REFERENCES

ISSN: 2233-7857 IJFGCN

[1] Yu Shao, Chip-Hong Chang, "Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition", *IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, March 2011, Vol. 41, No.2, pp. 284-293*

[2] Mr.P.Dileep Kumar, Dr.G.N.Kodanda Ramaiah Mr.A.Subramanyam, Mrs.M.Dharani,"A Solar powered Hybrid helmet with Multifeatures" International journal of Engineering Inventions e-ISSN:2278-7461,p-ISSN:2319-6491 Volume 4,Issue 10[June 2015]PP:06-11

[3] Marc Ferras**,** Cheung-Chi Leung**, "**Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition**",** *IEEE Transactions on Audio, Speech, and Language Processing, August 2010, Vol.18, No.6, pp.1366-1378.*.

[4] George Papandreou, Petros Maragos "Adaptive Multimodal Fusion by Uncertainty Compensation With Application to Audiovisual Speech Recognition", *IEEE Transactions on Audio, Speech, and Language Processing, March 2009, Vol.17, No.3, pp.423-435*.

[5] Samuel Thomas, Sriram Ganapathy, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction", *IEEE Signal Processing Letters, June 2008, Vol.15, pp. 681-684.*

[6] Ramón Fernández Astudillo, "An Uncertainty Propagation Approach to Robust ASR Using the ETSI Advanced Front-End *IEEE Journal Of Selected Topics In Signal Processing, October 2010, Vol. 4, No. 5, pp. 824-833*.

[7] Bhiksha Raj and Richard M. Stern, "Improving recognition accuracy in noise by using partial spectrographic information", *IEEE Signal Processing Magazine, September 2005, pp.101-116.*

[8] Ramalingam Hariharan, Imre Kiss, "Noise Robust Speech Parameterization Using Multiresolution Feature Extraction" *IEEE Transactions on Speech and Audio Processing, November 2001, Vol. 9, No. 8, pp.856-865.*

ISSN: 2233-7857 IJFGCN