ISSN:0975-9646

Review of Algorithms and Applications in Speech Recognition System

Rashmi C R
Assistant Professor, Department of CSE
CIT, Gubbi, Tumkur,Karnataka,India

***Abstract*- Speech is one of the natural ways for humans to communicate. Human Voice is a unique characteristic for any individual. A valuable biometric tool can be designed based on the ability to recognize a person by his/her voice and this biometric tool has enormous commercial as well as academic potential. It can be commercially used for ensuring secure access to any system. This paper delivers an overview of different algorithms that can be used in applications of speech recognition based on the advantages & disadvantages. It also helps in choosing the better algorithm based on the comparison done.**

***Keywords*- Feature extraction, Pattern matching**

I. INTRODUCTION

Speech signal conveys speaker information as well as linguistic information (e.g. Regional, physiological and emotional characteristics). This richness of information in speech has inspired many researches to develop the system that automatically process the speech, this speech technology has many applications [1]. Speech signal contains extremely rich information which exploits amplitude-modulated, time-modulated and frequency-modulated carriers (e.g. noise and harmonics, power, pitch, duration, resonance movements, pitch intonation) to convey information about words, speaker identity, style of speech, emotion, accent, the state of health of the speaker and expression. All these information are conveyed primarily within the traditional telephone bandwidth of 4 kHz. The speech energy above 4 kHz mostly conveys audio quality and sensation [2].

The information conveyed in speech includes the followings:

(a) Acoustic phonetic symbols. These are most elementary speech units from which larger speech units such as syllables and words are built. Some of the words have only two phones such as 'me', 'you',' he'.

(b) Prosody. These are rhythms of speech signal carried by changes in the pitch trajectory and stress and mostly called as intonation signals. This helps to signal such information as the boundaries between segments of speech, link sub-phrases and clarify intention and remove ambiguities such as whether a sentence is a question or a statement.

(c) Gender information. Gender information is generally communicated by the pitch (related to the fundamental frequency of voiced sounds) and the size and physical characteristics of the vocal tract. Because of the vocal anatomy differences, female voice usually has higher resonance frequencies and a higher pitch.

(d) Age. It is known by the effects of the size and the elasticity of the vocal cords and vocal tract, and the pitch. Children can have the pitch of voice more than 300 Hz.

(e) Accent. It broadly conveyed through: (i) any changes in the pronunciation that will be in the form of substitution, deletion or insertion of phoneme units in the "standard" transcription of words (e.g. US Jaan pronunciation of John or Australian todie pronunciation of today) and (ii) systematic changes in speech resonance frequencies (formants), emphasis, stress and pitch intonation, duration.

(f) Speaker's identity is known by the physical characteristics of a person's vocal folds, vocal tract, pitch intonations and stylistics.

(g) Emotion and health is known by changes in: vibrations of vocal fold, vocal tract resonance, duration and stress and by the dynamics of pitch and vocal tract spectrum.

II. REVIEW ON ALGORITHMS

In most of the Speech Processing applications, the speech signal or the audio signal has to be properly represented. Different representation formats are .MP3, .WAV, .AIFF, .AU, .RA etc. Based on the application proper format will be chosen. This representation follows feature extraction from the speech signal where only the necessary features will be extracted. Some applications like Speaker recognition systems require pattern matching (or Classification) to be done for the extracted features. Both feature extraction and pattern matching (or Classification) is necessary in many applications. These processes are carried out using different algorithms. This section provides the necessary details on the existing algorithms for feature extraction and pattern matching (or Classification) and also describes few of the applications.

A. *Feature Extraction*

According to the speaker recognition application, feature extraction is the process of retaining necessary information of the speech signal while rejecting redundant and unwanted information or we can say this process as analysis of speech signal. Sometimes while removing the unwanted information, we may also lose some useful information. Feature extraction may also involve transforming the signal into a form proper for the models used for classification. In developing a speaker recognition system, a few desirable properties of the features are:

- High discrimination between sub-word classes.
- Low Speaker variability.
- Invariance to degradations in the speech signal due to channel and noise.

The main aim is to find a set of properties of an utterance that have acoustic correlates in the speech signal, that is, parameters that can somehow be computed or estimated through processing of the signal waveform. Such parameters are called as features. Next procedure after the preprocessing of the speech signal in the signal modeling is feature extraction. The parameterization of speech signal is called as feature extraction. This is aimed to produce a meaningful representation of the speech signal. Feature extraction mainly includes the process of converting the speech signal to a digital form (i.e. signal conditioning), measuring few important characters of the signal such as energy or frequency response (i.e. signal measurement), augmenting these measurements with few perceptually-meaningful derived measurements (i.e. signal parameterization) and statistically conditioning these numbers to form observation vectors. The objective with feature extraction to attained are:

• To untie the speech signal into various acoustically identifiable components.

• To get a set of features with less rates of change in order to keep computations feasible.

Feature extraction can be categorized into three basic operations: spectral analysis, parametric transformation and statistical modeling. The complete steps are shown in figure 1[3].

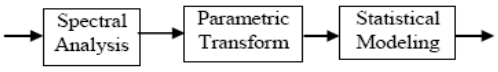

Figure 1: Feature Extraction Process

1. Spectral Analysis

When speech is produced in the sense of time varying signal, its characteristics can be represented via parameterization of the spectral activity. There are six major categories of spectral of analysis algorithms i.e. Digital filter bank (Power estimation), Fourier Transform (FT Derived Filter Bank Amplitudes, FT Derived Cepstral Coefficients), Linear Prediction (LP, LP Derived Filer Bank Amplitudes, LP Derived Cepstral Coefficients) used in speech recognition system.

2. Parameter Transforms

Signal parameters are generated from signal measurements through two fundamental operations: differentiation and concatenation. The output is a parameter vector with our raw estimates of the signal.

3. Statistical Modeling

In this step it assumed that the signal parameters were generated from few underlying multivariate random process. To study or reveal the nature of this process, it impose a model on the data, optimize (or train) the model, and then estimate the quality of the approximation. The only information about the process is its noticed outputs, the signal parameters that have been calculated. For this cause, the parameter vector output from this step of processing is called the signal observations. A statistical analysis is to be executed on the vectors to find if they are part of a spoken word or phrase or whether they are just noise. Speech sounds such as the 'ah' sound in the 'father' shows many resonance in the spectrum that usually extends 120ms. Transitional sounds, such as the 'b' in 'boy' exists may be for a brief interval of 20ms [3].

B. ***Algorithms for Feature Extraction***

The different algorithms used for feature extraction are RCC, MFCC, LPC, LPCC and PLPC. These are the most commonly used techniques in many applications for feature extraction especially in speaker recognition, speech recognition, biometric systems etc. This section provides a brief overview of the above algorithms.

1. RCC (Real Cepstral Coefficients)

To compute RCCs, the signal is transformed from the time domain to the frequency domain by applying a Fast Fourier Transform (FFT) to each frame. The log of the results and the inverse Fast Fourier transform (IFFT) is then applied to the signal to get the real Cepstrum of the signal, and can be written as follows [4]:

$$\text{Real Cepstrum} = \text{IFFT}\,(\log\,(\text{FFT}\,(s\,(n)))) \quad (1)$$

2. MFCC (Mel Frequency Cepstral Coefficients)

MFCC is one of the most popular algorithm and commonly used in most of the applications of speech signal for feature extraction. It is mainly used in speaker recognition systems or speech recognition systems. MFCC technique is based on the human peripheral auditory system. According to human perception of the frequency contents of sounds for speech signals, it does not follow a linear scale. Human perception is less sensitive at higher frequencies especially above 1000 Hz. Hence for each tone of actual frequency a subjective pitch is measured on different scale called as Mel Scale. Because of human perception behavior which does not follow linear scale that is above 1000 Hz, we take log scale above 1000Hz and call it as Mel Scale. This Mel scale specifies linearity up to 1000Hz and logarithmic above 1000Hz [5].

The human ear is sensitive to both the static and dynamic characteristic of a signal and the MFCC mainly concentrates on the static characteristics. To compensate the first order of the MFCC, the ΔMFCC can be appended to the MFCC to reflect the dynamic information of each frame [4].

3. LPC (Linear Predictive Coding)

In this method the speech signal is analyzed by estimating the formants. It also removes the effects of formants from the speech signal, and estimates the intensity and frequency of the remaining buzz. This process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC method, each sample of the speech signal is conveyed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding. The formants are characterized by the coefficients of the difference equation (the prediction coefficients) [5].

4. LPCC (LPC-Derived or Linear Predictive Cepstral Coefficients)

LPCC is also a well-known algorithm and widely used to extract feature in speech signal. LPC parameters can effectively describe energy and frequency spectrum of sound frames. The base of explaining acoustic signals spectrum, modeling and pattern recognition is set by the

result of increasing logarithm which restrains the fast change of frequency spectrum, more centralized and better for short-time character and it is because of Cepstrum derived from original spectrum. One of the common short-term spectral measurements currently used are LPC-derived cepstral coefficients (LPCC) and their regression coefficients [6].

LPCC shows the differences of the biological structure of human vocal tract and is computed through iteration from the LPC Parameters to the LPC Cepstrum [4].

5. PLPC (Perceptual Linear Predictive Cepstral Coefficients)

PLPCC is based on the magnitude spectrum of the speech analysis window. Like MFCC and LPC which are cepstral methods, the PLPCC is a temporal method. The steps followed to calculate the coefficients of the PLPCC are described. First, compute the power spectrum of a windowed speech. Second, for sampling frequency of 8 kHz perform grouping of the results to 23 critical bands using bark scaling. Thirdly to simulate the power law of hearing carry out loudness equalization and cube root compression. Fourth, perform inverse Fast Fourier Transform (IFFT). Fifth, perform LP analysis by Levinson- Durbin algorithm. Lastly, convert LP coefficients into cepstral coefficients [4].

The relationship between frequency in Bark and frequency in Hz is specified as in

$$f(\text{bark}) = 6*\text{arcsin}\, h(f(\text{Hz})/600) \quad (2)$$

6. PCA (Principal Component analysis)

It is a nonlinear feature extraction method. It is linear map, fast, eigenvector based method also known as karhuneu-Loeve expansion. This method is good for Gaussian data [18].

7. LDA (Linear Discriminant analysis)

It is a nonlinear feature extraction method. It is supervised linear map, fast, eigenvector based method. Better than PCA for classification [18].

8. ICA (Independent Component Analysis)

It is a nonlinear feature extraction method. It is linear map, iterative non-Gaussian method. The procedure used for implementation is blind course separation, used for demixing non-Gaussian distributed sources (features) [18].

C. *Algorithms for Pattern Matching or Classification*

After the extraction of necessary features from the speech samples, pattern matching or classification will be done in most speech processing applications. Some of the algorithms are VQ, HMM, GMM, SVM, MLP, DTW etc. This section gives brief overview of different pattern matching techniques.

1. VQ (Vector Quantization)

It is the classical quantization technique of signal processing which permits the modeling of probability density functions by the dividing of prototype vectors. The process starts by splitting a large set of points into clusters or groups having approximately the same number of points nearest to them. Centroid point represents each group. The density matching property of this quantization technique is very powerful, mainly in the density of large and high-dimensional data identification. Data points are shown by their closest centroid indexing, frequently occurring data have low error, and rare data high error. Hence, this method is also appropriate for lossy data compression [5].

2. HMM (Hidden Markov Models)

HMM is a popular statistical tool for modeling a huge range of time series data. In the context of natural language processing (NLP), HMMs are usually applied with great success to problems such as part-of-speech tagging and noun-phrase chunking. This powerful statistical tool is also used for modeling generative sequences that can be distinguished by an underlying process generating an observable sequence. HMMs have wide range of applications especially interested in signal processing which is more particular about speech processing and it has been used with success in low level NLP processes such as phrase chunking, extracting necessary information from documents and part-of-speech tagging. Andrei Markov gave his name to the mathematical theory of Markov processes in the early 12th century, but theory of HMMs was developed by Baum and his colleagues in 1960s [7].

This method can be called as a doubly stochastic process which has first-order Markov chain whose states are hidden from the spectator. Each state has a random process to produce observation sequence. The temporal structure of the data is captured by hidden states of model. There are 5 elements in HMM and they are number of hidden states, number of observation symbols per state, state transition probability distribution, observation symbol probability distribution in each state and initial state probability distribution [8].

3. GMM (Gaussian Mixture Model)

GMM is a probabilistic model used for density clustering and estimation. This model can be regarded as a special continuous HMM which has single state. GMMs are very effective in modeling multi-modal distributions. GMMs training and testing requirements are very less compared to the requirements of a general continuous HMM. GMM is based on the hypothesis that all the vectors are independent [8].

4. SVM (Support Vector Machines)

SVMs have proven to be most powerful method for pattern classification. SVM maps the input into a high-dimensional space and then distinguishes the classes with a hyperplane. A crucial aspect of opting SVMs successfully is due to the design of the inner product, the kernel, caused by the high dimensional mapping. The application of SVMs can be speaker and language recognition. It is a two-class classifier or also called as binary classifier. SVM can be considered as competitive and complimentary system to other approaches, such as Gaussian mixture models (GMMs) [9].

5. MLP (Multi-Layer Perceptrons)

MLPs are neural network based classifiers. They are used mainly for the powerful structure in classifying complex, nonlinear instants and in regression. Critical parameters such as learning rate, size of hidden layer, transfer functions in both hidden and output layers can be

well optimized to get best results for the specific purpose [10].

6. DTW (Dynamic Time Warping)

DTW is one of the Dynamic Programming technique based algorithm. This algorithm is mainly used for measuring similarity between two time series which may change in time or speed. This method is also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can further be used to determine similar regions between the two time series or to find the similarity between the two time series [11].

D. *Comparison of different methods*

In most of the applications like speaker recognition and speech recognition systems, MFCC is used for the feature extraction process rather than LPC or LPCC. MFCC is said to have better performance results. According to the review made in [1] for different feature extraction techniques, MFCC has better success rate than LPC and LPCC for speaker recognition application. The error rate for speaker recognition application is less when a combination of MFCC and VQ is used as quoted in [12] when compared with LPCC and VQ. According to the comparison of different feature extraction methods in [4], ΔMFCC and MFCC has better recognition rate.

For applications like emotion recognition, GMM gives better performance result compared to HMM as in experimental result of [8]. In gender recognition application, MLP, GMM and VQ have better accuracy [10]. The combination of MFCC and VQ gives good result as in [13] for speaker recognition. DTW is used only for application related to time series [11]. SVM with MFCC also gives good performance for automatic speaker recognition system [14]. Based on application proper algorithm can be selected for better performance.

E. Speech Processing Applications

Speech processing has five categories as mentioned in previous study of speech in chapter 1. Applications by considering the different categories of speech processing as mentioned in [15] are as follows:

1. Speech Recognition
   - Isolated Word Recognition (IWR)
   - Continuous Speech Recognition (CSR)
   - Key-Word Recognition (KWR)
   - Speech Understanding (SU)
2. Speaker Recognition
   - Speaker Verification (SV)
   - Speaker Identification (SI)
   - Language Identification (LI)
3. Speech Coding and Digitization
   - Waveform Coding
   - Source Coding Using Analysis/Synthesis
   - Vector Quantization (VQ)
   - Multiplexing
4. Speech Enhancement
   - Noise Reduction
   - Interference Reduction
   - Speech Transformations (Rate and Pitch)
   - Distortion Compensation
5. Speech Synthesis
   - Synthesis from Coded Speech
   - Synthesis from Text

Some of the applications in military areas are as follows [15]:

1. Speech Communications (Speech Coding, Speech Enhancement)
   - Secure Communications
   - Bandwidth Reduction
2. Speech Recognition Systems for Command and Control (C^2) (IWR, CSR, KWR, SU, Synthesis)
   - Avionics
   - Battle Management
   - Resource and Data Base Management
   - Interface to Computer and Communication Systems

3. Speech Recognition Systems for Training (IWR, CSR, SU, Synthesis)
4. Processing of Degraded Speech (Enhancement)
5. Security Access Control (SV)

III. CONCLUSION

In this review, there is a discussion on various algorithms that can be used for many speech processing applications. Most commonly used feature extraction algorithms and pattern matching algorithms are discussed. With the comparison we have found that MFCC is the commonly used algorithm for feature extraction of speech & the techniques like VQ, HMM, SVM for pattern matching.

REFERENCES

[1]. Bansod N.S.*, Seema Kawathekar and Dabhade S.B, "Review of different techniques for speaker recognition system", Volume 4, Issue 1, 2012, pp.-57-60.

[2]. dea.brunel.ac.uk/cmsp/Home.../Chapter13-Speech%20Processing.pdf.

[3]. Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur, "The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review", IJARCSSE, Volume 2, Issue 2, February 2012.

[4]. Genevieve I. Sapijaszko, Wasfy B. Mikhael, "An Overview of Recent Window Based Feature Extraction Algorithms for Speaker Recognition", IEEE, pp 880-883, 2012.

[5]. Vibha Tiwari, "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies, pp 19-22, 2010.

[6]. Hai-yan Yang, Xin-xing Jing," Performance Test of Parameters for Speaker Recognition System based on SVM-VQ", IEEE, pp 321-325, 2012.

[7] http://digital.cs.usu.edu/~cyan/CS7960/hmm-tutorial.pdf.

[8]. Manav Bhaykar, Jainath Yadav, and K. Sreenivasa Rao, "Speaker Dependent, Speaker Independent and Cross Language Emotion Recognition from Speech Using GMM and HMM", IEEE, 2013.

[9]. W.M. Campbell *, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", pp 210-229, Elsevier Ltd, 2005.

[10] Rafik Djemili, Rocine Bourouba, Mohamed Cherif Amara Korba, "A Speech Signal Based Gender Identification System Using Four Classifiers", IEEE, 2012.

[11] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, pp 138-143, VOLUME 2, ISSUE 3, MARCH 2010.

[12] Jorge MARTINEZ*, Hector PEREZ, Enrique ESCAMILLA, Masahisa Mabo SUZUKI, "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques, IEEE, pp 248-251, 2012.

[13] Fatma zohra .Chelali, Amar. DJERADI, "MFCC and vector quantization for Arabic fricatives Speech/Speaker recognition", IEEE, 2012.

[14] W. Astuti, A.M Salma, A.M. Aibinu, R. Akmeliawati, Momoh Jimoh E.Salami, "Automatic Arabic Recognition System based on Support Vector Machines (SVMs)", IEEE, 2011.

[15] Clifford J. Weinstein, "Opportunities for Advanced Speech Processing in Military Computer-Based Systems", Department of the Air Force and the Defense Advanced Research Projects Agency.

[186 review on Speech recognition technique", International Journal of Computer Applications, Volume 10-No. 3, November 2010.