Classification Techniques for Speech Recognition: A Review

Mayur R Gamit¹, Prof. Kinnal Dhameliya², Dr. Ninad S. Bhatt³

¹M.Tech student, ²Assistant Professor, E&C Department, C.G.P.I.T, Bardoli, India
³Associate Professor, E&C Department, CKPCET, Surat, India

Abstract—Speech Processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Speech recognition is the process of automatically recognizing the spoken words of person based on information content in speech signal. This paper introduces a brief survey on Automatic Speech Recognition and discusses the various classification techniques that have been accomplished in this wide area of speech processing. The objective of this review paper is to summarize some of the well-known methods that are widely used in several stages of speech recognition system.

Keywords—Feature Extraction, Acoustic phonetic Approach, Pattern Recognition, Artificial intelligence Approach, Speech Recognition.

I. INTRODUCTION

Automatic speech recognition (ASR) has been the most investigated topic in speech processing since early 1960s. Speech recognition is a popular and active area of research, used to translate words spoken by humans so as to make them computer recognizable. It usually involves extraction of features from speech signal and representing them using an appropriate data model. ASR system involves two phases. Training phase and Testing phase. In training phase, known speech is recorded and parametric representation of the speech is extracted and stored in the speech database. In the testing phase, for the given input speech signal the features are extracted and ASR system compares it with the reference templates to recognize the utterance.

There are usually two categories for isolated and continuous speech recognition: Speaker dependent and Speaker independent [2]. Speaker dependent method involves training a system that recognize each of the words uttered single or multiple times by specific set of speakers, while for speaker independent training is not applicable and words are recognized by analyzing their inherent acoustical properties. The main challenges of speech recognition involves modeling the variation of the same word as spoken by different speakers depending on speaking styles, accents, social dialects, gender, vocabulary size, recognition environment etc.

Some typical applications of such speech recognition are voice-recognized passwords, voice repertory dialers, automated call-type recognition, call distribution by voice commands, directory listing retrieval, credit card sales validation, speech to text processing, automated data entry etc.

Figure 1, shows basic representation of speech recognition system which consist of pre-processing, feature extraction, classification block as shown below.

A fundamental distinctive unit of a language is a phoneme. Different languages contain different types of phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables, concatenated to form phrases and sentences. One broad classification for English is in terms of vowels, consonants, diphthongs, affricates and semi-vowels [3]. The speech recognition system can be classified by the type of speech. They are continuous speech, isolated word, connected word and spontaneous speech.

II. PRE-PROCESSING

At the time of recording speech the interference due to noise mainly occurs. The performance can be degraded mainly due to noise. Before feeding the speech signal to feature extraction block the noise contained in speech signal must be removed. Preprocessing does this task. It removes the noise based on zero-crossing rate and energy. The separation of voiced and unvoiced speech based on both energy and zero-crossing rate gives the best result [4]. The start point and end points are determined based on energy and zero-crossing rates. The output speech contains the information and noise is eliminated.
III. FEATURE EXTRACTION TECHNIQUES

In speech recognition, feature extraction is the most important phase and the systems performance mainly depends on this block. The main task of feature extraction phase is to extract the feature from speech signal and representing them using an appropriate data model of the input signal. Various feature extraction techniques that are commonly used in speech recognition are as follows [7] [9] [11] [4].

A. Mel-Frequency Cepstrum Coefficients (MFCC)
B. Linear Predictive Coding (LPC)
C. Linear Prediction Cepstral Coefficients (LPCC)
D. Perceptual Linear Prediction (PLP)
E. Linear Discriminant Analysis (LDA)
F. Discrete Wavelet Transform (DWT)
G. Relative Spectral (RASTA-PLP)
H. Principal Component Analysis (PCA)

IV. CLASSIFICATION TECHNIQUES

In speech recognition there are three approaches [19][21].

A. Acoustic Phonetic Approach
B. Pattern Recognition Approach
C. Artificial Intelligence Approach

A. Acoustic Phonetic Approach

In Acoustic Phonetic approach the speech recognition were based on finding speech sounds and providing appropriate labels to these sounds [2] [21]. This is the basis of the acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units called phonemes and these units are broadly characterized by a set of acoustics properties present in speech.

B. Pattern Recognition Approach

The Pattern Recognition approach involves two essential steps namely, pattern training and pattern testing [19]. 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. This approach contains many techniques such as HMM, DTW, SVM, VQ etc.

1) Hidden Markov Model (HMM)

Hidden Markov model (HMM) is the most powerful parametric model at the acoustic level. The HMM is popular statistical tool for modeling a wide range of time series data [10]. An HMM is a doubly stochastic process with an underlying Markov process that is not observable.

The semantic of the model is usually encapsulated in the Hidden part for instance in ASR, an HMM can be used to model a word in the task-dependent vocabulary, where each state of the hidden part represents a phoneme [1]. The HMM can be described as follows:

- A set S of Q states, S = {S₁, S₂,...Sₜ}, which are distinct values that the discrete hidden stochastic process can take.
- An initial state probability distribution is given by the \( \pi = \{ P (S_i | t=0), S_i \in S \} \), where \( t \) is a discrete time index.
- An observation or the feature space F of states, which is \( \gamma_{t} = \{ P (S_j = q_{j}) \} \), it is a last state probability.
A probability distribution that characterizes the state transition probabilities which is \( a_{ij} = \{S_j | S_i\} \).

A set of probability distribution that describe the statistical properties of the observations for each state model: \( b(k) = \{P(O_t | q_t)\} \).

The summation of all these probability must be equal to one.

The sum of all the transition probability, initial state, last state and observation of states probability is equal to zero.

The most popular algorithms are the forward-backward and the Viterbi algorithms [1]. These belong to the class of unsupervised learning techniques, since they perform unsupervised parameter estimation of probability distribution. This technique is widely used because of its high recognition accuracy [10].

2) **Dynamic Time Warping (DTW)**

The time alignment of different utterances is the core problem for distance measurement in speech recognition. A small shift leads to incorrect identification. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed.

DTW is a method that finds an optimal match between two given sequences with certain restrictions. The sequences are warped nonlinearly in the time dimension [19]. DTW was recognized as the most suitable method for speech recognition because of its capability to cope with different speaking speeds. The Euclidean can be calculated as follows [17]:

\[
d(i, j) = (x_i - y_j)^2
\]  

(1)

Various sections of the utterances are stretched and compressed so as to find alignment that result in best possible match between test and reference vectors feature by feature basis[6].The local distance measure is the distance between features at a pair of frames while the global distance from beginning of utterance until last pair of frames.

3) **Vector Quantization (VQ)**

Vector Quantization is the pattern classification technique applied to speech data to forms a representative set of feature vectors. In the training phase, a speech specific VQ codebook is formed for each speech uttered by clustering the speech training acoustic vectors. Each group is represented by its centroid point. Vector quantization is a lossy data compression method based on principle of block coding. It is a fixed-to-fixed length algorithm [18].

It is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all codeword is called a codebook. The Euclidean distance is calculated from the input vectors by using the equation:

\[
d(x, y) = \sum_{i=1}^{M} (x_i - y_i)^2
\]  

(2)

Vector quantization involves extraction of features from training and testing data and VQ codebook model is built for all speech samples [5]. The distance between the input feature vectors and the code words are calculated and those having minimum distance can be selected as the recognized word.

4) **Support Vector Machine (SVM)**

Support vector machine is a simple and effective algorithm for classification of speech or speaker recognition [14]. SVM is a binary nonlinear classifier capable of guessing whether an input vector \( x \) belongs to a class-1 or class-2 category[15]. The margin known as soft margin is determined. The distance (\( R \)) between the sample points and margin are calculated.

As shown in above figure 3, the feature space consists of the features extracted from the speech signal. The decision boundary or soft-margin is determined to classify the problems. The distance between the margin and samples are computed. The sample that is nearer to margin or having least distance is chosen.

5) **Polynomial Classifier**

Polynomials have the excellent properties as classifiers. Polynomial classifiers are universal approximators to the bayes classifiers because of weierstrass approximation theorem[13].
The structure of classifiers is shown in figure 4. It consists of various blocks. The input features vectors $X_1, X_2, \ldots X_N$ are given to the classifier. The speaker model is given by $W_{spk}^T$ and $P(x)$ is the vector of the polynomial basis terms of the input feature vectors [12]. For each input feature vectors $X_i$, the score is produced by $W_{spk}^T P(x_i)$. The output is averaged to obtain the score.

$$S_t = \frac{1}{N} \sum_{i=1}^{N} W^T p(x_i)$$  \hspace{1cm} (3)

Where, $x_i$ = $i^{th}$ input test feature, $w$ = speaker model, $p(x_i)$ = vector of polynomial basis terms of the input test vectors.

C. Artificial Intelligence Approach

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing and finally making a decision on the measured acoustic features. The artificial intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach [7] [19]. The hybrid concept of both Hidden Markov Model and Artificial Neural Network is also applied in speech recognition [1]. The various methods in artificial intelligence are Multi-layer Perceptron (MLP), Self-Organizing Map (SOM), Back-propagation Neural Network (BPNN), Time Delay Neural Network (TDNNs) [1][20].

1) Time Delay Neural Network (TDNNs)

Time Delay Neural Network is an advanced version of the artificial neural network.

Time-delay neural networks (TDNNs) represent an effective attempt to train a static multilayer perceptron (MLP) for time sequence processing. An example of a TDNN is shown in figure 5.

The input layer has been enlarged to accept as many input patterns as the fixed sequence length to be processed at each time step. There are three layers input, hidden and output layers. The input vector enters the network from the leftmost set of input neurons. At each time step the inputs are shifted to right through the unit time delay. The outputs of the input layers are feed to the right most of the hidden layer and procedure for all the subsequent layers. Finally the output layer gives the output. This technique gives the ability to deal with more complicated time dependences. The back propagation algorithm can be used to train such a network. Recurrent Neural Networks (RNNs) is a better concept of artificial neural network.

2) Self-Organizing Map (SOM)

Kohonen proposed a neural network architecture which can be automatically generate self-organizing properties during unsupervised learning called Self-Organizing Map (SOM) [8].
SOM uses Euclidean distance to measure the distance between data vectors. The input vectors are normalized between -1 and +1 before it is feed into the network.

- The learning algorithm is as follows[8]:
- Each nodes weight is initialized between 0 to1.
- A vector is chosen at random from a set of training data.
- Every node is examined to calculate which nodes weights are most likely the input vectors.
- The winning node is commonly known as Best Matching Unit (BMU).
- The radius of neighborhood of BMU is calculated.
- More the node closest to BMU more the weights are updated.
- This procedure repeats until it is closer to the input vectors.

This technique is used to classify the features to reduce the feature vectors and complexity in speech recognition.

3) **Multilayer Perceptron (MLP)**

This MLP is a type of neural network which has one input layer, more than one hidden layer and one output layer that contains neurons [4].

As shown in figure 6, there are three input neurons, four hidden neurons and three output neurons. Weights associated to input and hidden layer are \( w_{ij} \) and \( u_{ij} \) are hidden and output weights. The input is given from left side of input neurons. The output of input neurons are multiplied with weights and feed as input to hidden neurons. Same operation is performed for hidden and output layer. This technique is commonly used in speech recognition systems [16].

![Figure 6. Multilayer perceptron](image)

**V. PERFORMANCE MEASURING PARAMETERS**

The performance of the speech recognition system is often described in terms of accuracy and speed. Accuracy may be measured in terms of Word Error Rate (WER), where speed is measured with the real time factor.

Single Word Error Rate (SWER) and Command Success Rate (CSR) are other measures of accuracy [2] [3].

**A. Word Error Rate (WER)**

Word error rate is a common metric of the performance of speech recognition. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence. The WER is derived from the Levenshtein distance working at the word level instead of the phoneme level. The word error rate can be computed as [2].

\[
\text{WER} = \frac{(S + D + I)}{N} \tag{4}
\]

Where,
\( S \) is the number of substitutions,
\( D \) is the number of deletions,
\( I \) is the number of insertions,
\( N \) is the number of words in the reference.

When reporting the performance of a speech recognition system, sometimes Word Recognition Rate (WRR) is used instead of Word Error Rate (WER).

\[
\text{WRR} = 1 - \frac{(N - S - D)}{N} = \frac{(H - I)}{N} \tag{5}
\]

Where, \( H = (N - S - D) \) is the correctly recognized words.

**VI. CONCLUSION**

In this paper different classification technique are introduced. Several advanced concept of classification techniques have been used recently in speech recognition systems. The problem always arises due to variation of the speech in time and environmental noise makes the recognition accuracy difficult. This paper describes the different classification techniques that can be helpful in speech recognition approach.

**REFERENCES**


