A Survey of Speech Recognition on South Indian Languages

Anand H.Unnibhavi
Dept. of Electronics and Communication Engineering
Basaveshwar Engineering College
Bagalkot, Karnataka, India
Anandhu.rampur@gmail.com

D.S. Jangamshetti
Dept. of Electronics and Communication Engineering
Basaveshwar Engineering College
Bagalkot, Karnataka, India
dsj1869@rediffmail.com

Abstract— **Automatic Speech Recognition is an active field of research to identify speech patterns for providing the equivalent text. Many types of interactive software applications are available and the uses of these applications are limited due to language barriers. Therefore development of speech recognition systems in local languages will help anyone to make use of this technological advancement. This paper presents a brief survey of Automatic Speech Recognition System of south Indian Languages and discusses the advances made in the recent years of research and compares some of the well known methods used in various stages of speech recognition system of south Indian languages.**

Keywords— Speech Recognition; Feature Extraction; MFCC; Support Vector Machine; RBF kernel; Hidden Markov Model; Neural Network; Dynamic Time Warping.

I. Introduction

Speech recognition is the translation of spoken words into text. It is also known as Automatic Speech Recognition (ASR), computer speech recognition or just Speech To Text (STT). Speech recognition system performs two fundamental operations: signal modeling and pattern matching. Signal modeling represents a process of converting voice signal into a set of parameters. The signal modeling involves four basic operations: Spectral shaping, feature extraction, parametric transformation, and statistical modeling. Spectral shaping is the process of converting the voice signal to a digital signal and emphasizing important frequency components in the signal. Feature extraction is a process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Parameter transformation is the process of converting these features into signal parameters. Statistical modeling involves conversion of parameters to observation vectors [1, 2]. The pattern matching approach involves two essential steps namely, pattern training (this is the method by which representative sound patterns are converted into reference patterns for use by the pattern matching algorithm) and pattern comparison. The essential feature 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, a word or a phrase. In the pattern comparison stage of the approach, a direct comparison is made between the unknown speech patterns with each possible pattern learnt in the training stage, in order to determine the identity of the unknown according to the goodness of match of the patterns.

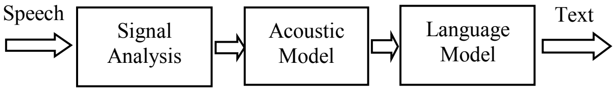

Fig.1: Basic Block diagram of speech recognition

Figure 1 shows different stages of Automatic Speech Recognition System. The microphone will pick up the analog signal and convert it into electrical signal and is passed to a processing unit called as signal analyzer (feature extractor), which processes the speech signal. The acoustic model compares the phonemes with the previous results from the trained models and highest probability match of a phoneme is selected as being the correct phoneme. Language model consist of two parts Dictionary file and Grammar file. Dictionary file matches the phoneme patterns to particular words and decides the correct pattern for the user input. Grammar file is made up of many templates which map the process of conversation that is made of limited number of paths which allows the Language Model to match a sentence to the user's response i.e Text output [3, 4].

A. *Types of speech recognition*

Speech recognition systems can be classified as:

- Isolated word recognition

Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results.

- Connected word recognition

In connected word systems separate utterances are 'run-together' with a small pause between them.

978-1-5090-4620-1/16/$31.00 ©2016 IEEE

- Continuous speech recognition

Continuous speech recognition deals with the speech where words are connected together instead of being separated by pauses. As a result unknown boundary information about words, co-articulation, production of surrounding phonemes and rate of speech affect the performance of continuous speech recognition systems.

- Spontaneous speech recognition

Spontaneous Speech recognition is natural sounding or it is unprepared speech. Spontaneous speech occurs in An interviews, debates, dialogues, etc. ASR system with spontaneous speech ability should be able to handle a variety of natural speech features.

II. Feature Extraction Techniques

Feature extraction involves analysis of speech signal. The feature extraction techniques are broadly classified as: temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis where as in spectral analysis spectral representation of speech signal is used [5].

A. Spectral analysis techniques

In speech analysis, spectral models of speech signals are useful in good understanding of the voice production process, and can be used for both speech synthesis and speech recognition. Different spectral analysis methods are discussed in this section.

- Critical band filter bank analysis

The concept of critical band is one of the successful simplifications of an aspect of our auditory functions. The critical band is dependent on the auditory periphery, particularly on the excitation patterns of the basilar membrane in the cochlea. The concept had been further simplified as a bank of band-pass filters that do not overlap each other. This model is well matched with the essential parts of the experimental facts of auditory masking and loudness perception. The bandwidths are chosen to be equal to a critical bandwidth for corresponding center frequency [6].

- Cepstral analysis

The objective of cepstral analysis is to separate the speech into its source and system components without any a priori knowledge about source and or system. According to the source filter theory of speech production, voiced sounds are produced by exciting the time varying system characteristics with periodic impulse sequence and unvoiced sounds are produced by exciting the time varying system with a random noise sequence. The resulting speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics.

If $e(n)$ is the excitation sequence and h(n) is the vocal tract filter sequence, then the speech sequence s(n) can be expressed as

$$s(n) = e(n) * h(n) \tag{1}$$

The frequency domain representation is

$$S(w) = E(w).H(w) \tag{2}$$

- Mel cepstrum analysis

An audio signal is constantly changing, so to simplify things it is assumed that audio signal does not change much on short time scales. This is why the audio signal is framed into 20-40 ms frames. The next step is to calculate the power spectrum of each frame. The periodogram estimates frequencies present in the frame. Clumps of periodogram bins are taken and it is summed up to get an idea of how much energy exists in various frequency regions. This is performed by Mel filter bank, and then followed by logarithm of Mel filter bank. The logarithm allows using cepstral mean subtraction, which is a channel normalization technique. The final step is to compute the Discrete Cosine Transform (DCT) of the log filter bank energies, the filter bank energies are quite correlated with each other. The DCT de correlates the energies which mean diagonal covariance matrices can be used to model the features.

- Linear predictive coding analysis(LPC)

One of the most powerful signal analysis techniques is the method of linear prediction. LPC of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients [6].

- Perceptually based linear predictive analysis (PLP)

Perceptual linear prediction is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short-term spectrum of the speech by several psychophysically based transformations. The PLP cepstral coefficients are computed using the PLP functions defined in the analysis library. Just like most other short-term spectrum based techniques, this method is vulnerable when the short-term spectral values are modified by the frequency response of the communication channel [7].

B. Temporall analysis techniques

It involves processing of the waveform of speech signal directly. It involves less computation compared to spectral analysis but limited to simple speech parameters.

- Power estimation

The use of some sort of power measures in speech recognition is fairly standard today. Power is rather simple to compute. It is computed on frame by frame basis as

$$P(n) = \frac{1}{N_s} \sum_{m=0}^{N_s - 1} (w(m) s(n - N_s / 2 + m)) \qquad (3)$$

Where,

N_s : Number of samples used to compute the power,

$s\ (n)$: Signal,

$w\ (m)$:Window function,

n :Sample index of center of the window

In most speech recognition system Hamming window is almost exclusively used. The major significance of *P(n)* is that it provides basis for distinguishing voiced speech segments from unvoiced speech segments. The values of *P(n)* for the unvoiced segments are significantly smaller than that for voiced segments. The power can be used to locate approximately the time at which voiced speech becomes unvoiced and vice versa.

- Fundamental frequency estimation

Fundamental Frequency (f_0) or pitch is defined as the frequency at which the vocal cords vibrate during a voiced sound. Fundamental frequency has long been difficult parameter to reliably estimate from the speech signal. It is useful in speech recognition of tonal languages (e.g. Chinese) and languages that have some tonal components (e.g. Japanese). Fundamental frequency is often processed on logarithmic scale, rather than a linear scale to match the resolution of human auditory system.

III. A Review on Speech Recognition

Hemakumar and Punith [8] have implemented continuous Kannada speech recognition system that works in speaker dependent mode using HMM method. The proposed method works through the following steps: The pre-processing of original Kannada speech database is done and then the speech signal is divided into segments called frames of 20 msec with an overlapping of 6.5 msec. In the second step the voice part is detected from the speech signal by computing the short time energy and magnitude of signal. In the third step the LPC coefficients are extracted from voiced part of signal using LPC feature extraction method and they are converted into real cepstrum coefficients. In the fourth step, real cepstrum coefficients are passed into k-means clustering algorithm and then passed into Baum-Welch decoding algorithm. Then 3 state HMM model is designed for each syllables/sub-words/sentence. The experiment is conducted for 20 unique sentences and each of these sentences recorded for 10 times for training and 3 times for testing of one male speaker. Experiment is performed using MATLAB and it shows that recognition accuracy rate of 87.76 % for individually uttered sentences.

In a work carried out by Anusha and Katti [9], the statistical approach is used to remove the silence region from the speech signal. The separation of silence region from the speech signal is done by the end point detection method. The silence region does not contain any information and it increases the storage area of the memory. Hence in the isolated word recognition system, accurate detection of endpoints of spoken word is important because reliable word recognition is critically depends on accurate end detection. Vector quantization technique is used to identify minimum speech patterns which are required in creating the training set of speech samples. Different approaches have been used in the speech recognition such as template approach, dynamic time warping, vector quantization etc based on the requirements and application.

In this Kannada speech recognition, the MFCC is used for the feature extraction. Speech database is constructed for training and testing. An adult female speaker uttered Kannada words from 1 to 10 each word 10 times (100) for training and 3 times (30) for testing purpose. For speaker dependent speech recognition test utterance were taken with the training set and for speaker independent speech recognition test utterance were taken with out of training set. After the feature extraction, features are fed into Vector Quantization (VQ) algorithm to form a group of clusters for each word. To reduce the problem of feature vectors for speaker dependent/independent recognition task, two clustering algorithm in VQ is used namely VQ_1 and VQ_2. The standard Euclidean distance measure is used to find the distance between the test signal and reference template of speech signal, when new silence removal algorithm is used, recognition error rate has been decreased from 2.59 to 1.56 in speaker dependent mode and 2.5 to 1.45 for speaker independent mode.

In another work carried out by Rohini *et al.* [10] a speaker-independent speech recognition system for Marathi language is presented. Marathi speech database is created by proper recording of speech utterances for training and testing purpose. For the generation of database the continuous speech and isolated words are recorded from the different age group (from 18 to 35 years) and total 21 speakers are considered. Feature extraction is done by LPC. Steps involved in the feature extraction are: acquiring the speech signal, digitization, pre-emphasis, framing and windowing, FFT computation and then finally DCT is computed for coefficients calculation. In the second phase the DTW technique is used to compare the similarities between the series of feature vectors which are computed from the every 10 msec of speech signal. DTW algorithm is based on the dynamic programming that is used to find the optimal alignment between two time series and finally the performance of LPC and DTW for same speakers and different speakers are compared.

In another work implemented by Gajanan Pandurang *et al.* [11], automatic speech recognition is used to identify each Marathi spoken word. In this work the features of the Marathi speech signals are extracted from the MFCC feature extraction method. Vector quantization (VQ) is used for training the features of speech signal. The training process of the VQ codebook applies an important algorithm known as the LBG VQ algorithm, which is used for clustering a set of L training vectors into a set of M codebook vectors. The matching of an unknown Marathi word is performed by measuring the Euclidean distance between the feature vectors of the unknown Marathi word to the model of the known Marathi words in the database. The goal is to find the codebook that has the minimum distance measurement in order to identify the unknown word. The result analysis shows that database1, database2 and database3 containing 63, 90 and 36 unknown speech samples were tested and yielding an accuracy rate of nearly about 85.71 %, 31.11 % and 88.88 % respectively.

In the study carried out by Vijai Bhaskar and Rama Mohana Rao [12] the speaker independent Telugu isolated word speech recognition system is developed. Sphinx4 is a speech recognition system written entirely in the java programming language. Sphinx4 was used to train and decode for recognizing isolated Telugu words. Isolated word speech recognition system for Telugu, based on Hidden Markov Models (HMM) and MFCC has been developed using Sphinx4. A HMM consists of a number of states; each state is associated with a probability density function. The HMM parameters comprises of the parameters of the set of probability density functions, and a transition matrix that contains the probability of transition between states. The MFCC feature vectors extracted from speech signals and their associated transcriptions are used to estimate the parameters of HMMs is called ASR system training. Over 29 context-dependent Telugu phonemes used for HMM Tool Kit, HTK-3.4 in the chosen application and basic acoustic units are context dependent phonemes, i.e. tri-phones modeled by left-to-right, 5-state HMMs. Two hundred and fifty Telugu words recorded 5 times by 25 speakers were used for training. Then 10 different speakers spoken some list of words for performance evaluation. An average of 91 % accuracy was achieved.

Sigappi and Palanivel [13] introduced the strategy for recognizing a preferred vocabulary of words spoken in Tamil language. The hidden Markov models (HMM) and auto associative neural networks (AANN) models are used in this speech recognition task. The HMM is used to model the temporal nature of speech and the AANNs to capture the distribution of feature vectors in the feature space. The created models provide a way to investigate an unexplored speech recognition arena for the Tamil language. The performance of the strategy is evaluated for a number of test utterances through HMM and AANN and the results project the reliability of HMM for emerging applications in regional languages. A text and speaker dependent medium-sized vocabulary speech recognition mechanism is designed with an ability to recognize one hundred railway station names uttered in Tamil language. The human ear resolves frequencies non-linearly across the audio spectrum and it is thus desirable to obtain the nonlinear frequency resolution. Auto Associative Neural Network models (AANNs) are feed forward neural networks bestowed with an ability to perform identity mapping of the input space and are increasingly used in speech processing applications. The speech database is created which includes the names of different railway stations of Tamil Nadu and 39 dimensional MFCC feature vectors extracted from the non silence frames of the speech signal corresponding to each word are given as input to estimate the parameters of HMM and is implemented using HTK toolkit. HMM with 5 states and 4 mixtures in each state yields a recognition rate of 95.0 % similarly the recognition rate using AANN is 90.0 %.

In a work carried out by Thushara and Gopakumar [14], a speaker independent speech recognition system for limited vocabulary Malayalam Words is developed. MFCCs are used for feature extraction. The classifier used for training and testing is Support Vector Machine (SVM). Six Malayalam words are used. The samples stored in the database are recorded by using a high quality studio recording microphone at a sampling rate of 16 kHz. Malayalam numerals from one to six are chosen to create the database. Twelve speakers are selected to record the words. Each speaker uttered six words with thirty samples each. The database is created for six male and female speakers. Thus the database consists of a total of 2160 utterances of the spoken words. The experimental results prove that training with Radial Basis Function (RBF) kernel gives better accuracy in recognition than with linear kernel. System trained with linear kernel has got an average accuracy of 75.5 % whereas for RBF kernel it is 91.8 %.

IV. Discussion and Conclusion

Speech has the potential to become an important mode of interaction with computer when such an interaction becomes a reality. The most important activity of speech recognition at present is to take sound and to translate this information into text and commands. In this paper classification of speech recognition system, applications of feature extraction techniques are discussed. Also recent developments of speech recognition system of south Indian languages such as Kannada, Marathi, Malayalam, Telugu and Tamil languages are reviewed. Through this review it is found that using the combination of HMM, MFCC, and Support Vector Machine (SVM), MFCC accuracy of recognition has been improved from 91 % to 91.8 % compared to existing methods such as Wavelet coefficients, ANN.

Acknowledgment

We are thankfull to the World Bank assisted TEQIP Phase-II of Basaveshwara Engineering College Bagalkot in the process of partial fulfillment of our research work.

References

[1] J Bilmes, <bilmes@ee.washington.edu> Lecture 2: Jan 5, EE516 Computer Speech Processing winter 2005.

[2] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition," Proceedings of the IEEE, Vol. 81(9), September 1993.

[3] http://www.matchproject.org.uk/resources/tutorial/Speech_Language/Speech_Recognition/Rec_4.html

[4] Vimala C, V. Radha, "A Review on Speech Recognition Challenges and Approaches," World of Computer Science and Information Technology Journal (WCSIT) Vol. 2, No. 1, 1-7, 2012.

[5] Shanthi Therese S, Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition," International Journal of Scientific Engineering and Technology, ISSN: 2277-1581, Vol. 2(6), pp 479-484, June 2013.

[6] Akansha Madan, Divya Gupta, "Speech Feature Extraction and classification A comparative Review", International Journal of Computer Applications (0975 – 8887) Vol. 90 No 9, March 2014.

[7] Lawrence Rabinenr, Biing Hwang Juang, "Fundamental of Speech Recognition".year 2009, First edition.

[8] Hemakumar G. , Punitha P, "Speaker Dependent Continuous Kannada Speech Recognition using HMM," International Conference

on Intelligent Computing Applications 2014.

[9] M. A. Anusha, S.K. Katti, "Speaker Independent Kannada Speech Recognition using Vector Quantization", International Journal of Computer Applications (IJCA) ISSN: 0975 – 8887. 7-8 April, 2012

[10] Rohini. B. Shinde, V. P. Pawar, "Marathi Isolated Word Recognition System based on LPC and DTW Technique," International Journal of Computer Applications (0975 – 8887) Vol. 59(6), December 2012.

[11] Gajanan Pandurang Khetri, Satish L. Padme, Dinesh Chnadra Jain, "Automatic Speech Recognition for Marathi Isolated Words," International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 1(3), November 2012.

[12] P. Vijai Bhaskar, S.Rama Mohana Rao, "Telugu Speech Recognition System development using MFCC based Hidden Markov Model technique with Sphinx-4," ISSN: 2347-9329 (Online) IJECEAR, Vol. 2 ISSUE 2, Feb. 2014.

[13] A.N. Sigappi, S. Palanivel, "Spoken Word Recognition Strategy for Tamil Language," International Journal of Computer Science Issues (IJCSI), Vol. 9(1), No 3, January 2012.

[14] Thushara P. V., Gopakumar. C, "An SVM Based Speaker Independent Isolated Malayalam Word Recognition System," International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS). ISSN (Print): 2279-0047 ISSN (Online): 2279-0055 March-May, 2043, pp. 281-285.