International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Review on Different Approaches for Continuous Speech Recognition System

Yin Win Chit, Dr. Renu

University of Technology (Yatanarpon Cyber City), Information and Communication Technology Department
Pyin Oo Lwin, Myanmar

Abstract: ***Speech is the most natural form of communication and interaction between humans. In computer system, the text and symbols are the most common form of translation. Speech Recognition is an important application that enables interaction of human being with machines. The various stages in the speech recognition system are pre-processing, segmentation of speech signal, feature extraction of speech and recognition stage. Among many speech recognition systems, continuous speech recognition system is very important and most popular system. This paper presents the basic idea of speech recognition, proposed types of speech recognition techniques, issues in speech recognition, different useful approaches for noise filtering, features extraction of the speech signal with its advantages and disadvantages and various pattern matching approaches for recognizing the speech of the speakers. Now day's research in speech recognition system is motivated for ASR system with a large vocabulary that supports speaker independent operations and continuous speech in different language.***

Keywords: Segmentation, Feature extraction, Pattern Matching

1. Introduction

When Speech is the most basic, common and efficient form of communication method for people to interact with each other. Speech is a powerful, flexible and familiar interaction modality. Speech Recognition (is also known as Automatic Speech Recognition ASR or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program. Research in speech processing and communication for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human verbal communication capabilities. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech recognition can also be isolated word recognition or continuous word recognition.

Since, 1960s computer scientists have been researching various ways and means to make computer record, interpret and understand human speech [1]. The fundamental aspect of speech recognition is the translation of sound into text and commands. Speech recognition is the process by which computer maps an acoustic speech signal to some form of abstract meaning of the speech. This process is highly difficult since sound has to be matched with stored sound bites on which further analysis has to be done because sound bites do not match with pre-existing sound pieces. Various feature extraction methods and pattern matching techniques plays important role in speech recognition system to maximize the rate of speech recognition of various persons.

This paper is organized as follows: section II describes the classification of speech recognition system. Section III gives architecture of speech recognition system. Section IV describes different types of Feature Extraction Techniques in speech recognition and section V describes about the Pattern Classification or Pattern Recognition Techniques in Speech Recognition.

2. Classification of Speech Recognition System

2.1 Types of speech based on utterances

1) *Isolated Words Speech:* Isolated word recognition system which recognizes single utterances (single word). Isolated word recognition is suitable for situations where the user is required to give only one word response or commands, but it is very unnatural for multiple word inputs. It is simple and easiest for implementation because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The drawback of this type is choosing different boundaries affects the results [2].
2) *Connected Words Speech:* Connected words are similar to isolated words, the only difference is that is allows separate utterances to be connected with less halts betwwen them. Utterance is the vocalization of a word or words that represent a single meaning to the computer. The disadvantage of this type is choosing different boundaries affects the results.
3) *Continuous Speech:* Continuous speech system recognizer is the one which requires the user to speak in a natural way. The continuous voice may or may not contain pauses. The computer determines the content that is contained in the input signal. Among all other systems, the most complex system to create and recognize is the continuous speech recognition systems as it requires special techniques to determine utterance boundaries. The larger is the dataset more is the complexity and lesser becomes the accuracy and performance. Complexity also increases due to the presence of noise in the signal [3].
4) *Spontaneous Speech: Spontaneous voice is unplanned or unprepared voice in which rehearsal is not done and*

which is naturally said by the user. Spontaneous speech is natural that comes suddenly through mouth. An ASR system with spontaneous speech is able to handle a variety of natural speech features such as words being run together. Spontaneous speech may include mispronunciation, false-starts and non-words.

2.2. Types of speaker Mode

Every individual who is the speaker for recognition systems is considered to have unique voice due to their distinguishing voice characteristics such as pitch, timbre, vibrato, glottal shape, vocal tract, etc. Speech recognition system is classified into three main categories as follows:

1) *Speaker Dependent Models:* Speaker dependent systems are developed for a particular type of speaker and depends on the speakers voice characteristics. Some of the features specified above can be used to distinguish and individually identify the particular speaker. During training phase specific user's voice is used to train the system. In the testing phase, pattern matching method searches for the best match between the test input and the voice that is already stored in the database. This model works very well with great accuracy for a particular singer or speaker whose voice is already stored in database, but much less accurate for rest of the singers.
2) *Speaker Independent Models:* Speaker independent systems are generally designed for live data or for variety of speakers without any prior training. A speaker independent system is developed to operate for any particular type of speaker. It is used in Interactive Voice Response System (IVRS) that must accept input from a large number of different users. But drawback is that it limits the number of words in a vocabulary. Implementation of Speaker Independent system is the most difficult. Also it is expensive and its accuracy is lower than speaker dependent systems.
3) *Speaker Adaptive Models:* Speaker adaptive speech recognition system uses the speaker dependent data and adapt to the best suited speaker to recognize the speech and decreases error rate by adaption [4].

2.3. Types of Vocabulary

The size of vocabulary of a speech recognition system can affect eh complexity, processing and the rate of recognition of ASR system. So that ASR system are classified based on the vocabulary as following:

- Small Vocabulary : 1 to 100 words or sentences
- Medium Vocabulary : 101 to 1000 words or sentences
- Large Vocabulary : 1001 to 10,000 words or sentences
- Very large Vocabulary : More than 10,000 words or sentences

3. Architecture of Speech Recognition System

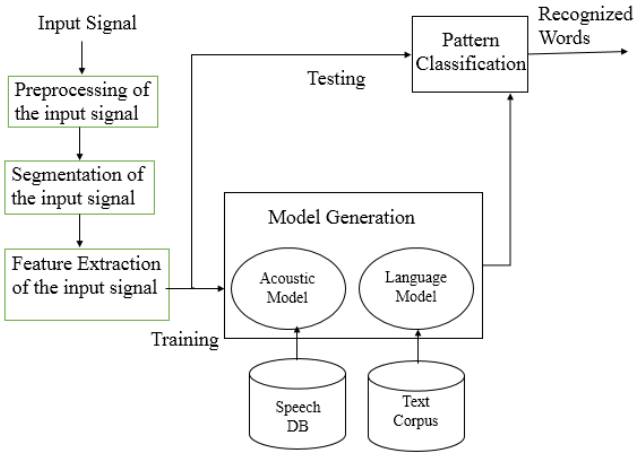

Figure 1: System Architecture for Continuous Speech Recognition System

3.1 Preprocessing / Digital Processing

The recorded acoustic signal is an analog signal. An analog signal cannot directly transfer to the ASR systems. So these speech signals need to transform in the form of digital signals and then only the can be processed. These digital signals are moved to the first order filters to spectrally flatten the signals. This procedure increases the energy of signal at higher frequency. Moreover speech enhancement can be made in this step of speech recognition system to improve speech quality and intelligibility of degraded speech signal. This is the preprocessing step.

3.2 Segmentation of Speech Signal

Speech segmentation problem has been studied over the span of several decades. In recent years, various approaches have been proposed for the design of an automatic speech segmentation system, such as the detection of variations or similarities in spectral and prosodic parameters extracted from the input speech, the discriminative learning segmentation, and the template matching using dynamic programming. However, the most frequently used approach is based on HMM phone models [5]. Segmentation of speech signal is very important part of the continuous speech recognition system. Because the continuous speech must be segmented to get the smaller speech units for good performance system.

3.3 Feature Extraction

Feature extraction step finds the set of parameters of utterances that have acoustic correlation with speech signals and these parameters are computed through processing of the acoustic waveform. These parameters are known as features. The main focus of feature extraction is to keep the relevant information and discard irrelevant one. To act upon this operation, feature extractor divides the acoustic signal into 10-25ms. Data acquired in these frames is multiplied by window function. There are many type of window functions that can be used such as Hamming Rectangular, Blackman, Welch or Gaussian etc. In this way features have been extracted from every frame. There are several methods for feature extraction such as Mel-Frequency Cepstral

Coefficient (MFCC), Linear Predictive Cepstral Coefficient (LPCC), and Perceptual Linear Prediction (PLP) etc. [6].

3.4 Acoustic Modeling

Acoustic modeling is the fundamental part of ASR system. In acoustic modeling, the connection between the acoustic information and phonetics is established. Acoustic model plays important role in performance of the system and responsible for computational load. Training establishes co-relation between the basic speech units and the acoustic observations. Training of the system requires creating a pattern representative for the features of class using one or more patterns that correspond to speech sounds of the same class. Many models are available for acoustic modeling out of them. Hidden Markov Model (HMM) is widely used and accepted as it is efficient algorithm for training and recognition. Many models or techniques are there for training the system [7].

3.5 Language Model

A Language Model contains the structural constraints available in the language to generate the probabilities of occurrence. It induces the probability of a word occurrence after a word sequence [8]. Each language has its own constraints. Generally Speech Recognition Systems use bi-gram, tri-gram, n-gram language models for finding correct word sequence by predicting the likelihood of the nth word, using the n-1 earlier words. In speech recognition, the computer system matches sounds with word sequence. The language model distinguishes word and phrase that has similar sound.

3.6 Pattern Classification

Pattern Classification (or Recognition) is the process of comparing the unknown test pattern with each sound class comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. After completing training of the system at the time of testing patterns are classified to recognize the speech. There are many pattern recognition method in the speech recognition system that are described in the section V.

4. Different Types of Feature Extraction Techniques used in Speech Recognition

Feature extraction is a technique used to find the features of speech that is spoken by different speaker. The feature extraction is a data reduction process that captures speaker specific properties [9]. Extracted features represent the whole speech. Many feature extraction techniques are used in speaker recognition. Many feature extraction techniques are described in this paper.

4.1 Linear Predictive Coefficient (LPC)

LPC is one of the powerful techniques for feature extraction. LPC gives vocal tract information of a speaker. Speech sample can be approximated as a linear combination of past few samples. The pre-emphasis of speech signal is the first step of flattening the spectrum of speech signal. Pre-emphasis boosts the higher frequencies in the signal. The next step is to frame the signal and multiply it by window function in order to reduce spectrum leakage in speech frame. In the last step, cepstrum is calculated by means of cepstral analysis. One disadvantage of LPC is that it does not capture spectral valleys. LPC is not so good features for identification of speakers. However, it is good for speech recognition [10]. LPC parameter is not so acceptable because of its linear computation nature. As human voice is nonlinear in nature, Linear Predictive Codes are not a good choice for speech estimation.

4.2 Mel Frequency Cepstrum Coefficient (MFCC)

Most of today's automatic speech recognition (ASR) systems are based on some type of Mel-frequency cepstral coefficients (MFCC), which have proven to be effective and robust under various conditions. To enhance the accuracy and efficiency of the extraction process, speech signals are normally pre-processed before feature are extracted. The MFCC technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. MFCC can be computed by using the formula [11].

$$\text{Mel}(g) = 2595 * \log 10 (1+g / 700) \quad (1)$$

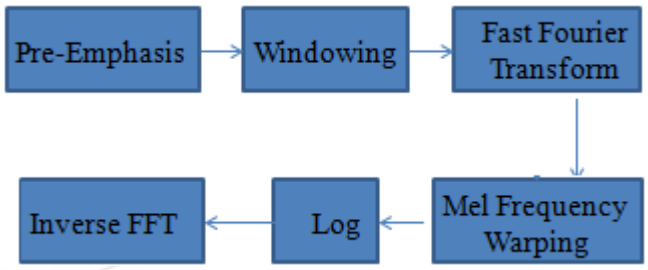

Figure 2: Block Diagram of MFCC

Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages.

In 2014, M.Kalamani, Dr.S. Valarmathy and C.poonkuzhali described about the MFCC advantages as following: Even though many speech recognition systems have obtained satisfactory performance in clean environments, recognition accuracy significantly degrades if the test environment is different from the training environment. These environmental differences might be due to additive noise, channel distortion, acoustical differences between different speakers, and so on. Hence, MFCC method has been developed to enhance the Accuracy and reduce the computational time for environmental robustness of speech recognition systems [12].

4.3 Perceptual Linear Predication Coefficient (PLPC)

PLP is used to calculate power spectrum of the speech signal. It modifies the spectrum of speech signal by several transformations. The power spectrum is integrated using a Bark-scale filter bank, which models the critical band frequency selectivity inside the human cochlea. The relationships between Bark scal and linear frequency is given by,

$$f_{bark} = 6 * \ln\left[\frac{f}{600} + \left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right] \quad (2)$$

The bark scale filters are trapezoidal in shape. The pre-emphasis of the signal is done using equal loudness curve after frequency integration. After this step, inverse Fourier transform is applied to the filter outputs to obtain autocorrelation sequence and Linear Predictive analysis is performed to smooth the spectrum. The final features are also obtained using cepstral recursion from the LP coefficients [13]. PLP model is identical to LPC model except that in PLP spectral characteristics are transformed to match the characteristics of human auditory system. DFT and LP techniques are merged in PLP scheme. LPC, MFCC and PLP are the most frequently used features extraction techniques in the fields of speech recognition and speaker verification applications. PLP and MFCC are derived on the concept of logarithmically spaced filter bank, clubbed with concept of human auditory system and hence had the better response compare to LPC parameters.

5. Pattern Classification/Pattern Recognition Techniques in Speech Recognition

The last step for speaker recognition technique is classification. It is used to classify different speaker from one over other. Many classification techniques are available for speech recognition. In this paper discusses about pattern matching methods as HMM, GMM and ANN.

5.1 Hidden Markov Model (HMM)

A model commonly used for speech recognition is the HMM, which is a statistical model used for modeling an unknown system using an observed output sequence [14]. HMM is a mathematical approach to recognize speech. It is a doubly embedded stochastic process with an underlying stochastic process that is not directly observable (it is hidden) but can be observed only through another set of stochastic processes that produce the sequence of observations. 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. Hidden Markov Model is a collection of states connected by transitions. Each transition carries two sets of probabilities: transition probability and output probability. Transition probability which provides the probability for taking this transition and output probability which defines the conditional probability of emitting each output symbol from a finite alphabet given that a transition is taken.

In [15] the minimum classification error (MCE) is used along with extended Baum Welch algorithm to optimize the HMM parameters is proposed. The experiments were performed on the speaker independent connected digits. The MEC has faster convergence rate and is stable than Generalized Probabilistic Descent (GPD).

5.2 Gaussian Mixture Model (GMM)

A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMM are commonly used as a parametric model of the probability distribution of continuous measurement or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum a Posterior estimation from a well-trained prior model. In GMM each speaker has the independent GMM model. Text-independent (TI) speech recognition can be done using Gaussian mixture models (GMM) [16].

The use of a GMM for representing feature distributions in a biometric system may also be motivate by the intuitive notion that the individual component densities may model some underlying set of hidden classes. In [17] GMM is used to separate the usable speech segments are from the interface which occurs when more than one individual speak together. GMM can identify 84 % of the usable speech while the other techniques can identify only 75.5% male 5 female speech utterances where taken from TIMIT (Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) speech corpus to perform the experiments.

5.3 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a network inspired by biological neural network which are used to estimate or approximate functions that can depend on a large number of inputs that are generally unknown. To learn couples functions, generalize effectively, tolerate noise, and support parallelism are some of the characteristic for neural networks [18]. Artificial neural network contain large number of processing elements, called as neuron which influence each other's behavior via a network of excitatory or inhibitory weights. Each unit simply computes a nonlinear weighted sum of its inputs, and broadcasts the result over its outgoing connections to other units. A training set consists of feature vector for different speaker that are assigned to designated input and /or output units.

6. Conclusion

In this review paper the basics of speech recognition system and different approaches available for feature extraction and pattern matching has been discussed. Using these various techniques rate of speech recognition can be developed. In future there will be focus on development of large vocabulary speech recognition system and speaker independent continuous speech recognition system. For

developing such systems in future Artificial Neural Network (ANN) and Hidden Markov Model (HMM) will be used at high level as in recent these techniques have become popular techniques in speech recognition process.

References

[1] A Pukhraj P. Shrishrimal, Vishal B. Waghmare, Ratnadeep Deshmukh, "Indian Language Speech Database: A Review", International Journal of Computer Application, Vol 47-No.5, June 2012.
[2] Sanjivani S. Bhabad, Gajanan K. Kharate, " An Overview of Technical Progress in Speech Recognition" , International Journal of Advanced Research in computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
[3] Swati Atame, Prof.Shanthi Theress S., Prof.Madhuri Gedam, "A Survey On: Continuous Voice Recognition Techniques", International Journal of Emerging Trends and Technology in Computer Science, Volume 4, Issue 3 , May-June 2015.
[4] Pratik K. Kurzekar, Ratndeep R. Deshmukh, Vishal B, Waghmare, Pukhraj P, Shrishrimal, " Continuous Speech Recognition System: A Review", Asian Journal of Computer Science and Information Technology, 2014.
[5] G. Aversano, A. Esposito, and M. Marinaro, " A New Text-Independent Method for Phoneme Segmentation", in Proc.of 44th IEEE Midwest Symposium on Circuits and System, v.2, pp.516-519, 2001.
[6] S.B Magre, R.R. Deshmukh, "A Review on Feature Extraction and Noise Recuction Technique", International Journal of Advanced Research in Computer Science and Software Engineering , Vol-4 , Issue 2, February 2014.
[7] Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, Muhua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing , 2012, 3, 394-401.
[8] Ankit Kumar, Mohit Dua, "Continuous Hindi Speech Recognition using Monophone based acoustic Modeling", International Journal of Computer Applications, (0975-8887), 2014.
[9] S. Sivan and C. Gopakumar, "An MFCC Based Speaker Recognition using ANN with Improved Recognition Rate," International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), Vol. 8, Issue 4, March-May 2014, pp. 365-369.
[10] S.B.Dhonde and S.M.Jagade, "Feature extraction Techniques in Speaker Recognition", ISSN:2349-947, Volume:2 Issue:5, IJRMEE, May 2015.
[11] Vimala, C., Radha, V., " A review on speech recognition challenges and approaches", World Computer. Sci. Inf. Technol., 2012, 2, (1), pp. 1-7.
[12] M.Kalamani, Dr.S. Valarmathy, C. Poonkuzhali and R. Karthiparksh, " Comparison of cepstral and mel frequency cepstral coefficient for various clean and noisy speech signals",Vol.2, Special Issue 1, ISSN, March2014.
[13] M. J.Alam, T.Kinnunen, P.Kenny, P. Ouellet and D. O'Shaughnessy, " Multitaper MFCC and PLP features for speaker verification using i-vectors", Speech Communication, ScienceDirect, Volume. 55, Issue 2, Pages 237-251 February 2013.
[14] M.T.Bala Murugan and M.Balaji , "SOPC-Based Speech to text Conversion", Nios || Embedded Processor Design Contest, Outstanding Designs 2006.
[15] Xiaodong He, Li Deng and Wu Chou, "A Novel Learning Method or Hidden Markov Models in Speech and Audio Processing", IEEE 8th Workshop on multimedia signal processing 2006.
[16] Chee-Ming Ting, Salleh S. H., Tian-Swee Tan and Ariff A.K., "Text independent Speaker Identification using Gaussian mixture model", International Conference on Intelligent and Advanced Systems, pp.194- 198, 25-28 Nov. 2007.
[17] Yantorno R.E., Smolenski B.Y., Iyer A. N. and Shah J. K., "Usable speech detection using a context dependent Gaussian mixture model classifier", *Proceedings of the International Symposium on Circuits and Systems*, vol. no. 5, pp. V-619- V-623, 2004.
[18] S. Rajasekaran and G.A. Vijayalakshmi Pai, "Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications", PHI, 2003.

Author Profile

Dr. Renu is working as an associate professor at department of information and communication technology in University of Technology (Yantanarpon Cyber City), Pyin Oo Lwin. She received master degree from University of Computer Studies, Yangon. She received doctoral degree at the same university.

Yin Win Chit is at present Ph.D candidate student from University of Technology (Yantanarpon Cyber City), Pyin Oo Lwin. She received Master of Computer Science (M.C.Sc) from University of Computer Studies, Magway. Her current research in Automatic Speech Recognition for Myanmar Language.