A Review on Different Approaches for Speech Recognition System

Suman K. Saksamudre
M.Tech Student
Dept. Of CS and IT
Dr. B. A. M. University

P.P. Shrishrimal
Research Student
Dept. of CS and IT
Dr. B. A. M. University

R.R. Deshmukh
Professor
Dept. of CS and IT
Dr. B. A. M. University

ABSTRACT

This paper presents the basic idea of speech recognition, proposed types of speech recognition, issues in speech recognition, different useful approaches for feature extraction of the speech signal with its advantage and disadvantage and various pattern matching approaches for recognizing the speech of the different speaker. 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.

General Terms

Pattern Recognition, Automatic Speech Recognition (ASR), Acoustic Modeling, Language Modeling.

Keywords

Feature extraction, pattern matching, ANN, HMM, DTW, MFCC.

1. INTRODUCTION

Automatic recognition of speech by machine has been a goal of research for more than four decades. In the world of science, computer has always understood human mimics. The idea which generated for making speech recognition system is because it is convenient for humans to interact with a computer, robot or any machine through speech or vocalization rather than difficult instructions [1]. Human beings have long been inspired to create computer that can understand and talk like human. Since, 1960s computer scientists have been researching various ways and means to make computer record, interpret and understand human speech [2].

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 [3] 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 are used to make better quality speech recognition systems. Feature extraction technique and pattern matching techniques plays important role in speech recognition system to maximize the rate of speech recognition of various persons.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

2.1 Types of speech recognition system based on utterances

2.1.1 Isolated Words

Isolated word recognition system which recognizes single utterances i.e. 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 [4].

2.1.2 Connected Words

A connected words system is similar to isolated words, but it allows separate utterances to be "run-together" with a minimal pause between them. Utterance is the vocalization of a word or words that represent a single meaning to the computer.

2.1.3 Continuous Speech

Continuous speech recognition system allows users to speak almost naturally, while the computer determines its content.

Basically, it is computer dictation. In this closest words run together without pause or any other division between words. Continuous speech recognition system is difficult to develop.

2.1.4 Spontaneous Speech

Spontaneous speech recognition system recognizes the natural speech. 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 speech recognition based on Speaker Model

Each speaker has special voice, due to his unique physical body and personality. Speech recognition system is classified into three main categories as follows:

2.2.1 Speaker Dependent Models

Speaker dependent systems are developed for a particular type of speaker. They are generally more accurate for the particular speaker, but could be less accurate for other type of speakers. These systems are usually cheaper, easier to develop and more accurate .But these systems are not flexible as speaker independent systems.

2.2.2 Speaker Independent Models

Speaker Independent system can recognize a 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.

2.2.3 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 [6]. They adapt operation according to characteristics of speakers.

2.3 Types of speech recognition based on Vocabulary

The size of vocabulary of a speech recognition system can affect the 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. FUNCTIONING OF SPEECH RECOGNITION SYSTEM

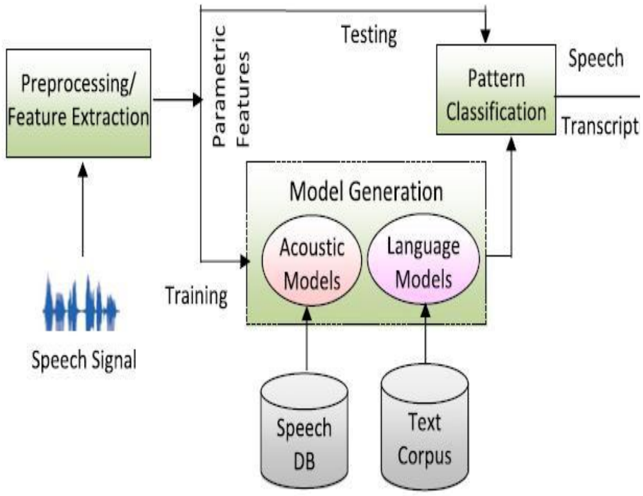

Fig 1: System Architecture of for Automatic Speech Recognition System [7]

3.1 Pre-processing/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 they can be processed. These digital signals are move to the first order filters to spectrally flatten the signals. This procedure increases the energy of signal at higher frequency. This is the preprocessing step.

3.2 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 extractor is to keep the relevant information and discard irrelevant one. To act upon this operation, feature extractor divides the acoustic signal into 10-25 ms. Data acquired in these frames is multiplied by window function. There are many types 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) [8], Linear Predictive Cepstral Coefficient (LPCC), Perceptual Linear Prediction (PLP), wavelet and RASTA-PLP (Relative Spectral Transform) [9]Processing etc.

3.3 Acoustic Modeling

Acoustic modeling is the fundamental part of ASR system [10]. 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 [11]. 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 [12] as it is efficient algorithm for training and recognition. Many models or techniques are there for training the system as explained in section 5.

3.4 Language Modeling

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 [13]. Each language has its own constraints. Generally Speech recognition systems uses 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. For example, in American English, the phrases like “recognize speech" and "wreck a nice beach" have same pronunciation but mean very different things. These ambiguities are easier to resolve when evidence from the language model is incorporated with the pronunciation model and the acoustic model.

3.5 Pattern Classification

Pattern Classification (or recognition) is the process of 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.

4. DIFFERENT FEATURE EXTRACTION TECHNIQUE USED IN SPEECH RECOGNITION

Table 1. Different Feature Extraction Methods Used In Speech Recognition System

Sr. No.	Method	Property	Advantage	Disadvantage
1	Principal component Analysis (PCA)	Nonlinear feature extraction method, Linear map; rapid; Eigen vector-based,	Good result for Gaussian data.	The directions maximizing variance do not always maximize information.
2	Linear Discriminate Analysis(LDA)	Supervised linear map, Depend on Eigen vector, Nonlinear feature extraction method.	Better than PCA for classification, Handles the case where the within-class frequencies are unequal and their performance has been examined on randomly generated test data.	If the distribution is significantly non-Gaussian the LDA projection will not be able to preserve any complex structure of the data, which may be needed for classification.
3	Independent component Analysis(ICA)	Nonlinear feature extraction method, Linear map, iterative non-Gaussian.	Blind than PCA for classification	Extracted components are not ordered.
4	Linear Predictive coding	10 to 16 lower sequence coefficient, Static feature extraction method	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel frequency scale	Frequencies are weighted equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic.
5	Filter Bank analysis	Filter tuned required frequencies	It provide a spectral analysis with any degree of frequency resolution (wide or narrow), even with non-linear filter spacing and bandwidths.	always take more calculation and processing time than discrete Fourier analysis using the FFT
6	Mel-frequency Cepstrum Coefficients (MFCC)	Power spectrum is computed by implementing Fourier Analysis.	This method is used for find our features.	MFCC values are not very robust in the presence of additive noises it is common to normalize their values in speech recognition system to reduce the influence of noise.
7	Kernel based feature extraction method	Nonlinear transformations	Dimensionality reduction leads to better classification and it is used to remove noisy and redundant features and improvement in classification error.	Slow similarity calculation speed [14].
8	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allows better time resolution at high frequencies than Fourier transform	It requires longer compression time.
9	Cepstral Mean Subtraction	Robust Feature Extraction	It is same as MFCC but working on Mean statically parameter.	
10	RASTA Filtering	For Noisy speech	It find out feature in noisy data	It increases the dependence of the data on its previous context.

5. APPROACHES FOR PATTERN MATCHING IN SPEECH RECOGNITION

5.1 Template- Based Approach

Template based approach has a collection of prototypical speech patterns. These patterns are stored as reference patterns representing the dictionary of words. Speech is recognized by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Normally Templates for entire words are constructed. Errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided [15].

Template based approach to speech recognition has provided a family of technique that has advanced the field considerably during the last two decades. This approach is simple. It is the process of matching unknown speech against a set of pre-recorded words or template in order to find the best match [16]. This approach has the benefit of using perfectly accurate word models; but this has the drawback that the pre-recorded templates are fixed. So variations in speech signals can only be modeled by using many templates per word, which certainly becomes impractical. Template training and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. This method is rather inefficient in terms of both required storage and processing power needed to perform the matching. Template matching is also tediously speaker dependent. Continuous speech recognition is not possible using this approach.

5.2 Knowledge-Based Approach

The use of knowledge/rule based approach to speech recognition has been proposed by several researchers and applied to speech recognition. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram [17]. The expert knowledge about variation in speech is hand-coded into a system. It takes set of features from the speech and then train the system to generate set of production rules automatically from the samples. These rules are resulted from the parameters that provide useful information about a classification. The effort of recognition is performed at the frame level, using an inference engine to implement the decision tree and classify the firing of the rules. This approach has the benefit of explicitly modeling variation in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach is considered as impractical and automatic learning procedures were sought instead.

5.3 Neural Network-Based Approach

Another approach for pattern matching in the speech recognition system is the use of neural networks. Neural Networks are capable of solving more complicated recognition tasks, but could not perform as excellent as Hidden Markov Model (HMM) when it comes to large vocabularies [18]. They can grip low quality, noisy data and speaker independency. This type of systems can achieve more accuracy than HMM based systems when there is training data and the vocabulary size is limited. A more familiar approach using neural networks is phoneme recognition. This is dynamic area of research, but generally its results are better than HMMs. There are also an NN-HMM hybrid system that uses the neural network as part of phoneme recognition and the HMM as part of language modeling.

Artificial neural network technology in speech recognition due to the following reason [19]:

- It reduces the modeling unit, generally in the phoneme modeling to advance the recognition rate of the entire system by improving the recognition rate of phonemes.
- Depth learning of the acoustic model, the brain operation structure, the introduction of context information, to reduce the impact of changes in voice more than the speech signal.
- Various feature are extracted from speech signal, a hybrid network model (HMM + NN) and to apply variety of knowledge sources i.e. characteristics, vocabulary and meaning of the word for speech recognition to understand the research, to advance system properties.

The application of artificial neural network in the field of speech recognition has been significantly developed in recent years. Artificial neural networks in speech recognition process can be divided into the following areas: **Firstly** improve the performance of artificial neural networks. **Secondly,** can be used to develop combine a hybrid system. **Thirdly**, mathematical methods represent the unique nature of neural network and applied to the field of speech recognition process. Artificial neural network in speech recognition has become a new emerging trend. Use of Artificial neural network technology has been successfully applied to solve pattern classification problems [20] and has shown to have enormous energy, speech recognition systems using artificial neural network will appear in the market and people will adjust their own way of speaking to accommodate a variety of recognition system.

5.4 Dynamic Time Warping (DTW) Based Approach

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. It is used in ASR, to cope with different vocalization speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences with certain restriction i.e. the sequences are “warped” non-linearly to match each other. This sequence ordering method is often used in the situation of HMM. In general, DTW is an approach that allows a computer to find an optimal match between two given sequences with certain limitations. This technique is useful for isolated word recognition and can be modified to recognize connected words also [21].

5.5 Statistical- Based Approach

In this approach, variations in speech are modeled statistically (e.g. HMM) using training methods. This approach represents the current state of the art [22]. Present general purpose speech recognition systems are based on statistical acoustic and language models. Acoustic model and language models for ASR in unlimited domain require large amount of acoustic and linguistic data for parameter estimation. Processing of extreme amounts of training data is a key element in the development of an effective ASR technology. The main drawback of statistical models is that they must make a priori modeling presumption, which is liable to be inaccurate, restrict the system’s performance.

5.5.1 Hidden Markov Model (HMM)-Based Speech Recognition

Hidden Markov Model based speech recognition system has become popular. Because HMM can be trained automatically and computationally feasible to use [23]. HMMs are simple networks that can generate speech using a number of states for each model and modeling the short-term spectra associated with each state. The parameters of the model are the state transition probabilities the means, variances and mixture weights that represent the state output distributions. Each word or phoneme, will have a different output distribution [24]; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes. Modern HMM based large vocabulary speech recognition systems are often trained on hundreds of hours of acoustic data. The word sequence, pronunciation dictionary and HMM training process can automatically determine word. This means that it is relatively straightforward to use large training corpora. It is the main advantage of HMM which will extremely reduce the time and complexity of recognition process for training large vocabulary.

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 improved and better quality 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.

7. ACKNOWLEDMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. This work is supported by University Grants Commission.

8. REFERENCES

[1] Malay Kumar, R K Aggarwal, Gaurav Leekha and Yogesh Kumar “Ensemble Feature Extraction Modules for Improved Hindi Speech Recognition System”, International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012. Pukhraj P. Shrishrimal, Vishal B. Waghmare, Ratnadeep Deshmukh, “Indian Language Speech Database: A Review”, I international Journal of Computer Application, Vol 47– No.5, June 2012.

[2] Hemakumar, Punitha, “Speech Recognition Technology: A Survey on Indian languages”, International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013.

[4] 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.

[5] M. A. Anusuya, S.K. Katti, “Speech Recognition by Machine: A Review”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[6] 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.

[7] R.K. Aggarwal, M. Dave “Integration of Multiple acoustic and language models for improved Hindi speech Recognition system”, Springer Science Business Media, LLC 2012.

[8] Bishnu Prasad Das1, Ranjan Parekh, “Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers”, International Journal of Modern Engineering Research ,Vol.2, Issue.3, May-June 2012.

[9] S.B. Magre, R.R. Deshmukh, “A Review on Feature Extraction and Noise Reduction Technique”, International Journal of Advanced Research in Computer Science and Software Engineering Vol 4, Issue 2, February 2014.

[10] Rajesh Kumar Aggarwal, Ph.D_Thesis, “Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement”, December 2012.

[11] Ankit Kumar, Mohit Dua, “Continuous Hindi Speech Recognition using Monophone based Acoustic Modeling”, International Journal of Computer Applications® (0975 – 8887), 2014.

[12] Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, Mahua Bhattacharya, “Development of Application Specific Continuous Speech Recognition System in Hindi”, Journal of Signal and Information Processing, 2012, 3, 394-401

[13] Preeti Saini, Parneet Kaur, “Automatic Speech Recognition: A Review”, International Journal of Engineering Trends and Technology- Volume4Issue2-2013.

[14] Hanmin Jung, Sung-Pil Choi, Seungwoo Lee, “Survey on Kernel-Based Relation Extraction”.Chapter 1.

[15] 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.

[16] Vimala.C, Dr.V.Radha, “A Review on Speech Recognition Challenges and Approaches”, World of Computer Science and Information Technology Journal, Vol. 2, No. 1, 1- 7, 2012.

[17] Ranu Dixit, Navdeep Kaur, “Speech Recognition Using Stochastic Approach: A Review”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 2, February 2013.

[18] Joe Tebelskis, “Speech Recognition using Neural Networks”, PHD thesis, School of Computer Science Carnegie Mellon University Pittsburgh, Pennsylvania May 1995.

[19] Abhishek Thakur, Naveen Kumar, “Automatic Speech Recognition System for Hindi Utterance with Regional Indian Accents: A Review”, International Journal of Electronics & Communication Technology, Vol. 4, April – June 2013.

[20] Simone Marinai, Marco Gori, “Artificial Neural Networks for Document Analysis and Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 1, January 2000.

[21] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, “A Review on Speech Recognition Technique” International Journal of Computer Applications, Vol 10– No.3, November 2010.

[22] M. Chandrasekar, M. Ponnavaikko, “Tamil speech Recognition: a complete model”, Electronic Journal «Technical Acoustics» 2008.

[23] Rashmi C R, “Review of Algorithms and Applications in Spe ech Recognition System”, International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014.

[24] R. Vinay Chand, M.Veda Chary, “Wireless Home Automation System with Acoustic Controlling”, International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 9- Sep 2013.