Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 2, Number 2; January-March, 2015, pp. 149-152

http://www.krishisanskriti.org/jbaer.html

Analysis of Algorithms in the Speech Recognition Process

Vijeta[1] and Dharam Veer Sharma[2]

[1]Research Student, M.Tech(CSE) Department of Computer Science, Punjabi University, Patiala
[2]Department of Computer Science, Punjabi University, Patiala
E-mail: [1] jindalvijeta5@gmail.com, [2]dveer72@hotmail.com

Abstract—*Speech is one of the most prominent and natural form of communication among humans. Human voice is a unique characteristic of each human. Nowadays, it has become a very useful biometric. A number of valuable biometric tools have been designed to recognize a particular set of words. Such tools have enormous commercial applications as voice dialing, call routing and many more. Also, it has evolved as a novel approach of security. This paper analyzes the various algorithms involved in the Feature Extraction and PatternMatching phase of Speech Recognition Process. A brief overview of different algorithms which include Linear Predictive Coding (LPC), Mel-frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction(PLP) and RASTA-PLP (Relative Spectra Filtering of Log Domain Coefficients) in Feature Extraction and Dynamic Time Warping (DTW), Hidden Markov Model(HMM) and Neural Networks(NN) in the Recognition phase, is given based on their advantages and limitations.*

Keywords: *Hidden Markov Model(HMM), Dynamic Time Warping(DTW),Mel frequency cepstral co-efficient (MFCC), Linear Predictive Coefficients (LPC), Perceptual Linear Prediction(PLP).*

1. INTRODUCTION

Speech recognition is a process of automatically extracting and determining the various linguistic features or information present in the speech signal. It is a research area in which human spoken words are translated into a digitized form or a computer recognizable form. To understand the operation of speech recognition some background information is needed on how speech sounds are produced. A person's vocal cords cause air to vibrate and generate sound waves. These waves travel through the air to the ear where the brain interprets the sounds. Words consist of speech sounds, which are known as phonemes. They have characteristics which allow humans to identify them.There are two types of speech recognition:

Independent speech recognition: It can be defined as the recognition of vocabulary items without regard to who is speaking. Independent speech recognition is working at 95% accuracy or better for populations of limited size.

Dependent speech recognition: It is the recognition of vocabulary items spoken by a particular speaker. It requires that users "train" the system to recognize vocabulary items of a particular voice. These systems create templates that will be used for subsequent comparisons to real time speech. Dependent speech recognition systems are performing at 98% or better unless the user who created the templates has some dramatic change in voice characteristics.

Speech recognition basically involves extracting the patterns from digitized speech patterns and representing them by an appropriate data model. Automatic speech recognition methods have been investigated for many years and the first technical paper on speech recognition was published in 1952. It described Bell Labs spoken digit recognizer. The system relied on measuring spectral resonances during the vowel region of each digit.

Some of the challenges being faced during the development of a speech recognizer-

- Difference in the speaker's accent and style for speaking the same word.
- Gender, age and voice patterns of different speakers.
- Background noises.
- Varying signal properties over time.
- Regional and social dialects of different speakers.

2. SPEECH RECOGNITION PROCESS

The speech recognition process begins by the creation of a speech database for a given vocabulary size. We need different speakers for uttering the words in various utterances of each word which are recorded using a microphone or a similar recording medium. The process can be shown diagrammatically as**:**

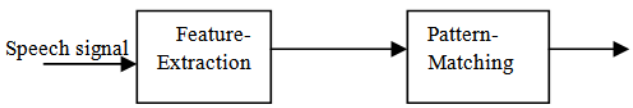

Fig. 1: Speech Recognition Process

Feature Extraction

Feature extraction is a process of retaining the necessary information of the signal while rejecting the redundant data. Sometimes one may lose useful information also while removing unwanted data. It involves extracting the discriminatory features from the speech data. Hence we can say that feature extraction transforms the signal into a form appropriate for models for classification. Here are some desirable properties of features being extracted[6]:

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

Phoneme is the basic unit of speech. The Feature Extraction process extracts the characteristic features out of the spoken utterance. The spoken utterance is then partitioned into frames of size approx. 16-32 msec and updated every 8-16 msec and performs analysis based on the technique being used.

Almost all the speech recognition systems use a parametric representation rather than the waveform itself as the basis for pattern recognition. The features can be extracted directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Few techniques generate patterns from the features and use them in classification by the degree of correlation, while other techniques use the numeral values of the features coupled to statistical methods. Broadly speaking there are two types of techniques used for feature extraction:

Temporal analysis: Here the speech waveform itself is used for analysis.

Spectral analysis: Here the spectral representation of the speech signal is used for analysis.

Comparison: Temporal analysis techniques involve less computation, ease of implementation but they are limited to determination simple speech parameters like power, energy and periodicity of speech,whereas for finding vocal tract parameters we need spectral analysis techniques [1].

Various algorithms used in Feature Extraction

Various algorithmic blocks used are Fast Fourier Transformation(FFT), calculation of logarithm(LOG), Discrete Cosine Transform(DCT) and sometimes Linear Discriminative Analysis(LDA). Cepstral coefficients obtained through Linear Predictive Coding(LPC) are widely used for modeling. Another well-known and widely used speech extraction is based on Mel-frequency Cepstral Coefficients(MFCC). Methods based on Perceptual Prediction which is good under noisy conditions are PLP and RASTA-PLP(Relative Spectra Filtering of Log Domain Coefficients). There are certain other methods available to extract features from speech[4].

Basic Idea of Pattern Matching and Various Algorithms Used

Dynamic Time Warping(DTW) and Hidden Markov Model(HMM),based on the concept of Dynamic Programming, are two well-suited and widely used pattern matching algorithms for recognition of spoken words.They provide us an optimal mapping from test signal to the template signal. As we know that brute force algorithm has exponential complexity which is reduced to the order of O(nm) by dynamic programming(where n and m are the length of the test and template signals). The research trend has transited from DTW to HMM in around 1990's because DTW is deterministic and lack of the power of DTW to model stochastic signals.Since DTW uses the concept of deterministic DTW and is not able to handle various speech or video signals that are stochastic in nature, a new algorithm called "stochastic DTW" is proposed [2]. This method uses conditional probabilities instead of local distances and transition probabilities instead of path costs as in DTW. This stochastic DTW is actually related to HMM. The result of stochastic DTW improved the recognition rate from 89.3% to 92.9% in word recognition experiments[5].

A more recent technique to independent speech recognition is to use neural networks. As HMM works by making certain assumptions about the structure of speech recognition, and then estimates system parameters as though the structures were correct. This technique may fail if the assumptions are incorrect. The neural network approach does not require any such assumptions to be made. This approach uses a distributed representation of simple nodes, whose connections are trained to recognize speech. Unlike HMMs where knowledge or constraints are not encoded in individual units, rules or procedures, but distributed across many simple computing units. Uncertainty is modeled not as unlikelihood of a single unit, but by the pattern of activity in many units. These computing units are simple in nature, and knowledge is not programmed into any individual unit's function; rather it lies in the connections and interactions between linked processing elements.

Optimization for large vocabulary

Generally isolated word recognition processes work well for a small vocabulary. But the number of computations is large and consumes a lot of time. So we need some optimization criteria for solving the problem with large vocabulary size and real time ASR systems. An optimization approach, called three step recognition method, diminishes the processing time for isolated word recognition presented in [3]. This method enabled to achieve the real time recognition goal. Also, as it is

considered that triphones always give better results than phones but it shows that this assertion is not always true.

3. LITERATURE SURVEY

Isolated Word Recognition system based on DTW and LPC was developed[8]. It used VQ to create reference templates. 12 words of Lithuanian language were pronounced ten times by ten speakers to evaluate the performance. The main source of errors recognized was that only one word of each speaker was used as a referencetemplate for recognition. The recognition error rates in speaker dependent and speaker independent mode 0.83% and 1.94% resp.

Mel-frequency cepstral coefficient (MFCC) and vectorquantization are used for the recognition system for a Marathi database[7]. It also compared the recognition systems based on the MFCC and LPC as the feature extraction methods. It concluded that the recognition accuracy is more with MFCC as compared to LPC.

For different feature extraction techniques, MFCC has better success rate than LPCC and LPC[10].It[16] concluded that PLP sometimes performs better than MFCC and LPC. It showed that RASTA-PLP outperforms any other feature extraction method. Also by increasing the number of gaussian mixtures per HMM state to a value of 10, system give a better recognition rate and an optimal HMM is used.

Speech signal modeling technique forrecognition of isolated word computed spectral and temporal features forrecognition of alphabet based on ISOLET database andHMM toolkit was used to implement it[17].Patternrecognition and audio processing were important aspectsin visual and audio stimuli in order to reflect intelligentbehavior. The use of neural network and linear predictivecoding techniques[18] together improved the performance ofspeech and text dependency recognition.

4. CONCLUSIONS.

A lot of research work has been done on English language and digits[2,9,12]. Research has been carried on some regional languages of India[7,15] and some foreign languages also[2,8].

As it is known that human voice is non-linear in nature, so Linear Predictive Codes are not considered to be a good choice. PLP and MFCC have a better response comparable to LPC parameters[4]. Among MFCC and LPC, MFCC gives a better recognition accuracy[7]. Words classified using a combination of features based on LPC, MFCC, Zero Crossing Rate(ZCR) and Short Time Energy(STE) gives a betterrecognition accuracy than achieved using these features individually[9]. It also concludes that MFCC has better recognition rates than LPC. Also, ANN gives a better accuracy compared to other classifiers as Euclidean Distance.RASTA-PLP method when used with wavelets for noisy speech yields better results[15]. But its accuracy decreases if it deals with clean speech. In contrast, other feature extraction methods like MFCC and LPC have a higher accuracy rate with clean speech as compared to their accuracy for noisy speech.

In Pattern Matching process, DTW and HMM are two well known algorithms being used. But the trend has transited from DTW to HMM during 1990's[5]. This occurred because DTW is deterministic in nature and is not able to model the stochastic signals. So HMM is considered to be the best choice for pattern matching process and is extensively used with varying success rates[13,14]. DTW is used only for application related to time series[11].

REFERENCES

[1] Manish P. Kesarkar, "Feature Extraction for Speech Recognition", M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Submitted November2003.

[2] Seiichi Nakagawa, etc, "Speaker-Independent English consonant and Japanese word recognition by a Stochastic Dynamic Time Warping method", Journal of Institution of Electronics and Telecommunication Engineers, 1988.

[3] Horia CUCU, Andi BUZO,Corneliu BURILEANU, "Optimization Methods For Large Vocabulary,Isolated Word Recognition In Romanian Language", U.P.B. Sci. Bull., Series C, Vol. 73, Iss. 2, 2011, ISSN 1454-234x, 179-192.

[4] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, Volume 1, Issue VI, July 2013.

[5] Chunsheng Fang, "From Dynamic Time Warping (DTW) to Hidden Markov Model (HMM)", University of Cincinnati, 2009/3/19.

[6] Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur, "The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 2, February 2012 ISSN: 2277 128X.

[7] Leena R Mehta, S.P.Mahajan, Amol S Dabhade, "Comparative StudyofMFCC andLPC for Marathi Isolated WordRecognition System, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 6, June 2013,ISSN (Print) : 2320 – 3765, ISSN (Online): 2278 – 8875 .

[8] Antanas LIPEIKA, Joana LIPEIKIEN E, Laimutis TELKSNYS, "Development of Isolated Word Speech RecognitionSystem", INFORMATICA, 2002, Vol. 13, No. 1, 37–46.

[9] Bishnu Prasad Das, 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 (IJMER) Vol.2, Issue.3, May-June 2012 pp-854-858 ISSN: 2249-6645.

[10] Bansod N.S, Seema Kawathekar and Dabhade S.B, "Review ofdifferent techniques for speaker recognition system", Volume 4, Issue1, 2012, pp.-57-60.

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

[12] R. Low, R. Togneri, Speech recognition using the probabilistic neural network, *Proc. 5th Int. Conf. on Spoken Language Processing*, Australia, 1998.

[13] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE, 77(2), 1989, 257-286.

[14] M. S. Rafiee, A. A. Khazaei, A novel model characteristics for noise-robust automatic speech recognition based on HMM, *Proc. IEEE Int. Conf. on Wireless Communications, Networking and Information Security (WCNIS)*, 2010, 215-218.

[15] M.A.Anusuya, S.K.Katti, "Comparison of Different Speech Feature ExtractionTechniques with and without Wavelet Transform to Kannada Speech Recognition", International Journal of Computer Applications (0975 – 8887) Volume 26– No.4,July 2011, 19-24.

[16] T.Schiirer, "Comparing Different Feature Extraction Methods for Telephone Speech Recognition based on HMM's", Institute Fur Fernmeledtechnik Techiscihe Universitat Berlin, 234-237.

[17] Karnjanasecha Montri and Zahorian Srephen A., "Signalmodeling for high performance robust isolated word recognition" , IEEE transactions on speech and audioprocessing , vol.9, no.6, September 2001.

[18] Panagiotakis Costas and Tzititas George, "IEEE Transaction on Multimedia", **7**, No. 1, February 2005.