Advances in Voice Enabled Human Machine Communication

Shweta Sinha
Research Scholar
Dept of Information Technology
Birla Institute of Technology,
Mesra, India

S S Agrawal
Director General
KIIT World, Gurgaon
India

Aruna Jain
Associate Professor
Dept of Information Technology
Birla Institute of Technology,
Mesra, India

ABSTRACT
The inherent advantage of speech communication due to its variability, convenience and speed along with our increasing requirements to communicate with machines has driven the attention of researchers towards mechanical recognition of speech. Technological advancements and improvements in the fundamental approaches have shown a successful transition from small vocabulary isolated word recognition to large vocabulary continuous speech recognition. Even after years of research and development the accuracy of automatic speech recognition remains one of the major challenges. Design of speech recognition system requires careful selection of feature extraction technique and modeling approach to cover the challenges faced due to variability of speech-speaker characteristic, storage space and processing speed requirements. In this paper an effort has been made to highlight the progress made so far for mechanizing the recognition of speech along with the major challenges in this field. Authors have also presented a brief description of voice enabled service for common people. The objective of this paper is to summarize some of the well known methods used at various stages of speech recognition system along with their benefits and limitations.

General Terms
Speech recognition, Artificial Neural networks

Keywords
Dynamic Time Warping, Feature Extraction, Hidden Markov Model, Neural Network, Speech Recognition.

1. INTRODUCTION
Speech is the primary means of communication among human kind and an obvious choice of interaction with machine. Designing machine that can mimic human behavior, specially the capability of seeking and understanding has attracted a great deal of interest for several decades.

A device to understand speech needs an intelligent machine capable of making complex decisions and that too in a speed that matches human brain. This can be achieved by using the power of computers to develop Automatic Speech Recognition (ASR) system, which captures a relevant input signal and transforms it into written text. Textual translation of speech signal is the most common objective of ASR, but these systems can also support spoken queries, dictation systems, command and control medical applications and speech translation from one language to the other.

Most of the systems developed till date are based on English and other foreign language speech, which restricts the usage of these machine oriented speech based interface only among educated Hindi speaking population of India. It is a fact that effective communication always takes place in speaker's own language; hence Speech interface for database queries supporting Hindi is a special need. Very limited work has been done in support of Hindi and other Indian languages.

This paper initiates with the evolution of ASR systems, in the next section it covers signal processing techniques extending it from feature extraction methods to approaches to speech recognition. Language modeling, speech enabled service; challenges in ASR and conclusion have been covered in subsequent sections.

2. EVOLUTION OF ASR SYSTEMS
For the development of ASR it is important to have the basic understanding of speech as it will help us to better decide what aspects may be relevant for development of such a system.

The basic component of speech is a sequence of phonemes which combine to create words and thereafter sentences. Knowledge of human speech production is important in the context of seeking acoustic aspects that are important for speech perception as each phoneme corresponds to vocal tract shape [1].

Attempts to mimic speech by machines started as early as 18^{th} century, where readily available acoustic resonance tubes were used to approximate human vocal tract. Since each individual has a different vocal tract controlled by his brain, speakers following languages based on same linguistic rules have different speaking style, rate and accent. Even same word repeated by same speaker has variations, which can be readily observed in its digital representation. Technological advancements have tried to deal with such type of variability and also to reduce the complexity of speech systems.

Development of speaker dependent, isolated utterances of 10 syllables was built in RCA laboratories in 1950[2]. Then in 1970, speaker independent voice dialing system was developed at Bell Laboratories. In 1990s dialogue management systems were developed which mimicked the communicating capabilities of human [3-4]. Advancements in the research area of spontaneous speech processing took place in late 1990s [5]. With major technological advancements in dealing with temporal and spectral variability in speech signals, ASRs with capability to recognize continuous speech independent of speakers have been developed [6].

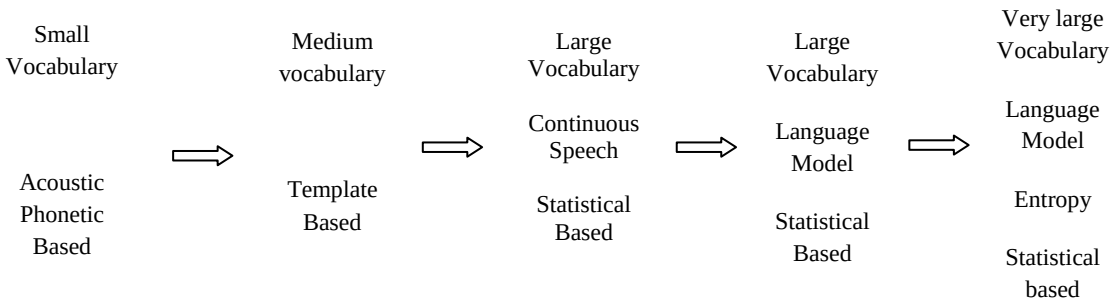

Figure 1: Milestones in speech research

3. SIGNAL PROCESSING FOR ASR

One of the major goals of ASR is to accept an input waveform, act upon it and transform it to written text. This requires the acoustic knowledge as well as the linguistic knowledge to process the spoken utterances.

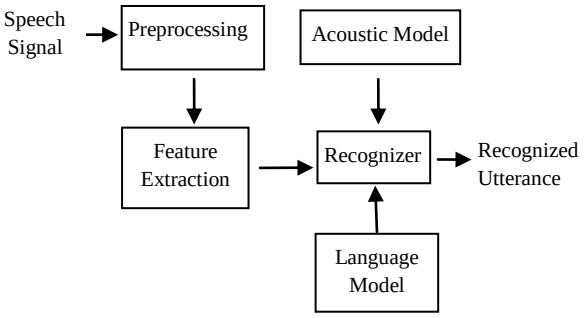

Figure 2: Automatic speech recognition System

The overall ASR is executed in two phases: Pre-processing and post processing phase. During pre-processing, noise and unwanted speech segment is removed and further features are extracted from the clean speech. In post processing phase, speech recognition engine is developed which is based upon the knowledge of acoustic and language model. The model so obtained is used by the recognition engine to correctly identify the most likely match for the test input.

The most important task of ASR system development is the adoption of suitable feature extraction mechanism and the recognition approach in order to decode the spoken utterances. Feature extraction helps us to obtain important characteristics which can be used for comparison during recognition.

4. SPEECH FEATURE EXTRACTION TECHNIQUES

Every speech has different individual characteristics embedded in each utterance. These characteristics can be separated from one another by a suitable feature extraction method. A major challenge in the selection of feature extraction technique is the choice of how to reduce or compress the acquired data, while minimizing the loss of relevant information. The extracted feature occurring frequently and naturally in speech should be stable over time. Two signal processing techniques are generally used for feature extraction: Spectral based and Cepstral based methods. Cepstral features are based on auditory perspective and research [7] shows that for classification purpose these feature extraction technique produces less error as compared to spectral based. Several methods falling under these two categories have been defined in literature but the most widely used are Linear Predictive Coding (LPC) and its variations, MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Predictive), RASTA (Relative Spectra) and its variants.

4.1 Linear Predictive Coding

No sooner did the concept of linear prediction [8] came into existence for speech coding, the fundamental concept of Linear predictive coding [LPC] for speech recognition was formulated [9] [10]. The goal of LPC is to determine the coefficients of linear predictive filters. For this, the speech signal is divided into consecutive fixed length frames and coefficients are computed so as to minimize the mean square prediction of any given signal frame. The estimation of source-vocal tract (filter) response separation is simplified by sampling the continuous scalar signal with a fixed sample rate. Owing to its simplicity many works [11-12] in the area of speech recognition have been done. Several variations of LPC as, LPCC [13-14], LPCMCC [15] also exists. Figure 3 represents the block diagram of LPC.

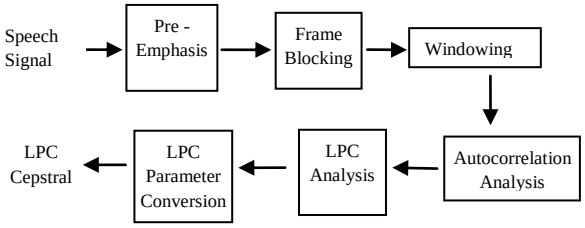

Figure 3: LPC feature extraction process

To minimize analysis complexity, the speech signal is usually assumed to come from an all-pole source, i.e., its spectrum has no zeros. Since actual speech has zeros due to the usual glottal source excitation and due to multiple acoustic paths in nasal and unvoiced sounds this assumption was not acceptable [23]. LPC is still in use for cell phone speech transmission, but has been replaced by MFCC for ASR.

4.2 Mel Frequency Cepstral Coefficients

The usual objective in selecting a parametric representation of speech is to compress it by eliminating information that are not crucial for phonetic analysis of speech data while enhancing those aspects of the signal that contribute most to the detection of phonetic differences[16].Motivated by human perceptual factors MFCC approach came into existence.

The MFCC are based on the known variation of the human ear's critical bandwidth frequencies. Filters are spaced linearly at low frequencies and logarithmically at high frequencies to

capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linearly spaced at frequency below 1000 Hz and logarithmically spaced at frequency above 1000 Hz [17]. Figure 4 represents the block diagram of MFCC feature extraction. Speech signal is divided into overlapping frames of 20-25 ms and windowed, typically using Hamming window to remove discontinuity at the frame boundary. FFT (Conversion from time domain to frequency domain) is obtained for each frame and is passed through set of triangular filters spaced according to the perceptual Mel scale (Mel-Frequency warping). Logarithm of spectral amplitude is then taken and the output so obtained is finally converted back to time domain by an inverse FFT (Cepstrum) process to give MFCCs.

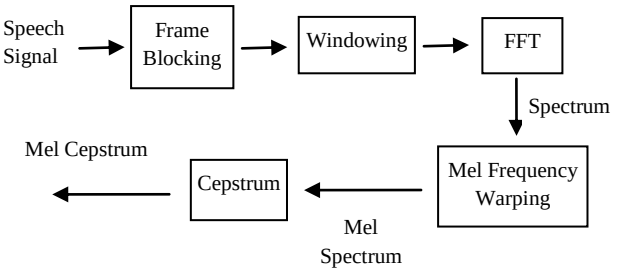

Figure 4: MFCC feature extraction process

Several work based on this approach have been done [18-21] and is being continuously used till now. MFCC allows better suppression of insignificant spectral variations in the higher frequency bands. Initially it was observed that six coefficients capture most of the relevant information but further research showed that to capture more and more speaker guided aspects of data during transition from one sound to other delta and delta-delta variations need to be extracted[16].Variations of MFCC have also been practiced to obtain much better results [22].

Even though MFCCs have been widely used, they too are suboptimal. The first MFCC (C0) is the energy and C1 is interpreted as indicating global energy balance between low and high frequencies, but other MFCC coefficients are difficult to relate them to any clear aspect of speech production or perception. Also MFCCs give equal weight to both low and high amplitudes in the log spectrum despite the fact that high energy dominates perception, leading to deterioration of MFCCs in case of speech corrupted with noise [23].

4.3 Perceptual Linear Predictive (PLP) Analysis of Speech

Lack of interpretability of the MFCCs forces to use simple merging of distributions to handle different speakers. This type of merging leads to larger variances and hence lowered ability to discriminate against other phoneme models. To deal with this a related approach based on nonlinearly compressed power spectrum made its place in development of some ASR. This approach was named as perceptual linear prediction [24].

PLP is a LP-based analysis method that successfully incorporates a non-linear frequency scale and other known properties from the psychophysics of hearing. In PLP analysis, a Fourier transform is first applied to compute the short-term power spectrum, and the perceptual properties are applied while the signal is represented in filter-bank form. The spectrum is transformed to a Bark scale, and this spectrum is pre-emphasized by a function that approximates the sensitivity of human hearing at different frequencies. The output is compressed to approximate the non-linear relationship between the intensity of a sound and its perceived loudness. The all-pole model of LPC is then used to give a smooth, compact approximation to the simulated auditory spectrum, and finally the LP parameters are transformed to obtain cepstral coefficients for use as recognition features. Figure 5 presents the block diagram of PLP computation. Several research based on PLP [25-28] have been carried out for speech and speaker recognition giving very promising result.

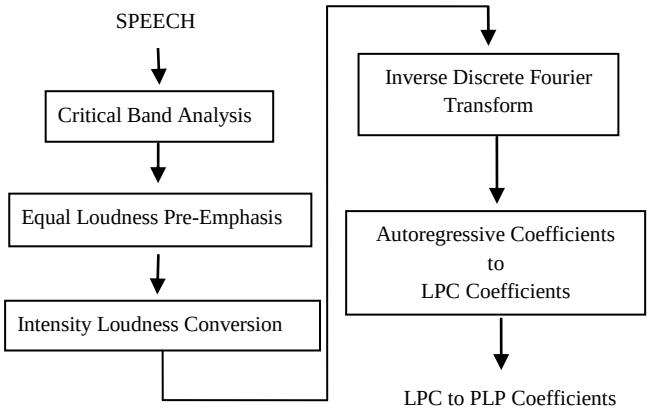

Figure 5: Perceptual linear predictive process

4.4 Relative Spectra (RASTA) – PLP Speech Analysis

Human recognition of speech is less sensitive to its slowly changing or steady state factors. To get the same effect in ASR, spectral estimate is applied on speech samples in which each frequency channel is band pass filtered by a filter with sharp spectral zero at zero frequency to suppress constant and slowly varying component in each channel and get spectral estimate less sensitive to slow variations in short term spectrum. This process is applied as initial step during PLP coefficients calculation and is known as RASTA [29].

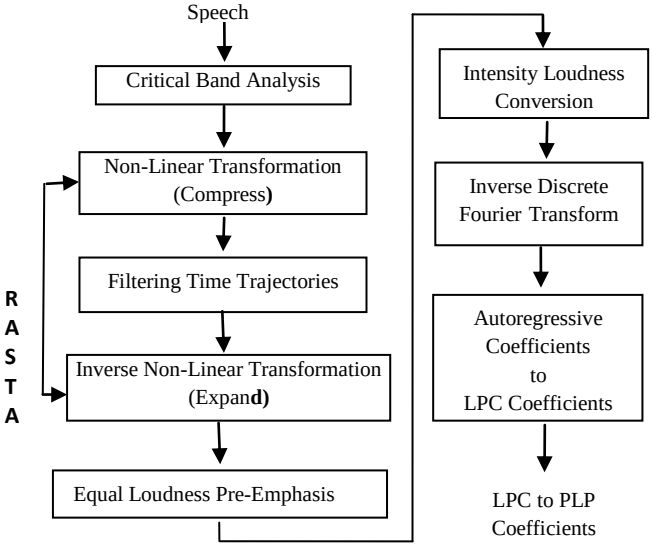

Figure 6: Feature Extraction using RASTA-PLP

RASTA-PLP makes recognizer much more robust to factors like microphone types and its position. Experiments show that RASTA-PLP does not show much better performance up gradation for filtered speech, but for unfiltered speech error rate is reduced to half. This method increases context dependency; hence performance in case of sub word recognizers degrades. Variations of RASTA processing have been produced to handle both additive and convolution noise [29] and has made Rasta and its variants a very common choice [30-33] for speech recognition.

Due to reduced complexity and noise robustness MFCCs are till date the most widely used feature extraction method. Many researches [34] show that MFCC gives better performance than any other method.

5. APPROACHES TO SPEECH RECOGNITION

Once the speech signal is modeled some mechanism has to be devised for recognizing the spoken utterance. There are mainly three approaches to speech recognition. The first one is acoustic phonetic approach, where the machine attempts to decode the speech signal in a sequential manner based on the observed acoustic features of signal and their known relations with phonetic symbols [35]. Due to difficulty in decoding phonetic units into word strings this method did not get much success [37-38].

The second approach is the artificial intelligence (AI) approach. Two key concepts of artificial intelligence are automatic knowledge acquisition and adaptation. These concepts are generally applied using artificial neural network (ANN). Neural network (NN) is a connectionist model of neurons distributed for parallel processing. It attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing and making decisions regarding measured acoustic features [39]. Several works in the area of ASR have been published based on the use of ANN. ANN models using multilayer perceptrons, recurrent networks, Kohonen's self organizing maps and learning vector quantizers have been widely used in the development of many systems [10] [41-42]. Even though NN do not scale well with large vocabulary their capability to solve complicated recognition tasks and its robustness and fault tolerance makes it a preferable choice for ASR till today. Combination of ANN with DTW, HMM and other matching techniques are being used in many of the ongoing researches.

The third and the most widely used approach is the pattern recognition approach. Due to its simplicity it has gained acceptance in very large scale and till today it is being widely used. The execution is carried out in two phases, in the first phase system is trained with speech utterance and a template is created representing the utterance. In the second phase the test utterance is compared against the template. For isolated word recognition Dynamic Time Warping (DTW) is the most widely used template matching technique. In recent years statistical pattern recognition has replaced traditional methods of template creation and matching. They create simple probabilistic model with large number of training sets easily. Hidden Markov Models are the most widely used statistical model in recent researches.

5.1 Dynamic Time Warping

In early years, dynamic programming techniques have been developed to solve the pattern-recognition problem. To handle the temporal and spectral variability in speech, templates are stretched nonlinearly (warp) and are compared trying to synchronize similar acoustic segments in test and reference patterns. This procedure combines alignment and distance computation in one dynamic programming procedure to find the optimal path. The aim of time warping here is to minimize the total distance measures, summing the distance measures of successive frame to frame match [43]. This technique is quite efficient for isolated word recognition and can be modified to recognize connected word also [44-46].

5.2 Hidden Markov Model

HMM is a network representation of acoustic events based on their statistical information in speech. First order Markov chain is considered as the base for this model, where the likelihood of being in a given state depends only on the immediate prior state. These networks are called HMM because the models are inferred through observation of speech output, not from any internal representation of speech production [36].

HMM works as a finite state machine which is assumed to be built up from a finite set of possible states having some probability density function (PDF) associated with them. Fundamental problems of HMMs are probability evaluation, determination of the best sequence, and parameter estimation. Several search algorithms and methods are available for probability evaluation and for finding the best sequence. For parameter estimation in ASR HMM uses maximum likelihood estimation (MLE) using a forward-backward procedure [47]. Several discriminative training methods have been proposed in recent years to boost ASR system accuracy like maximum mutual information estimation (MMIE); minimum classification error (MCE); and minimum word error/minimum phone error (MWE/MPE) [48-52].

At present, much of the recent researches on speech recognition involve recognizing continuous speech from a large vocabulary using HMMs or a hybrid of HMMs with other techniques like ANN.HTK toolkit is available for research in ASR based on HMM [53].

6. LANGUAGE MODEL

Initially ASR only used acoustic knowledge during evaluation process of text hypothesis. Soon it was realized that speech usually follow linguistic rules, so incorporating knowledge about text being spoken would enhance ASR performance. Also, knowledge about prior words in an utterance guides in finding out the most likely next word and language models are statistical description of likelihood of text in a sequence.

N-gram models estimates the likelihood of each word given the preceding n-1 words and are the generalization of Markov model which assumes that we can predict the probability of some future unit without knowledge of much back history[35].unigram, bigram, trigram etc are the most common n-gram model [54] which helps in estimating the next word. As the vocabulary size increases the language model grows exponentially leading to increased demand for storage and high computational capability, so the value of n is kept not more than 3.

7. VOICE ENABLED INFORMATION ACCESS

One of the most demanded voice based service today is information access through speech interface. This type of system requires two components, one for speech recognition

and the other for query execution. Figure 8 represents the architecture of such a system.

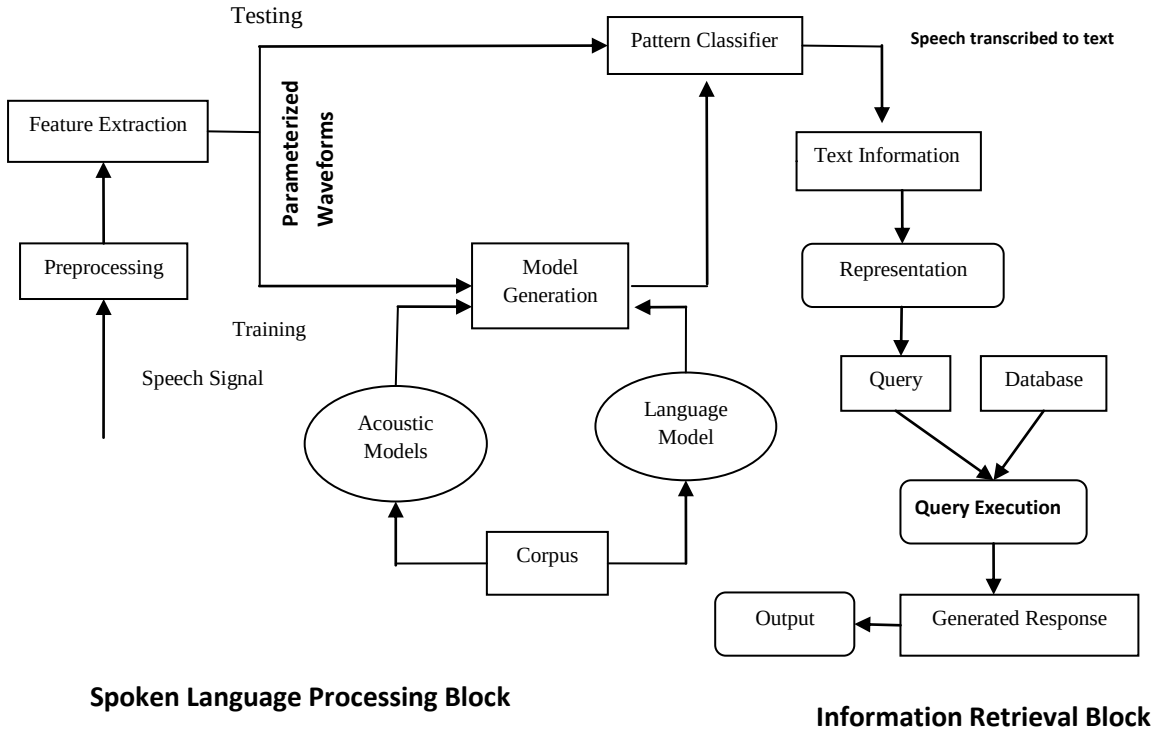

Figure 8: Voice enabled speech interface for data access

Speech interface in these systems is responsible for capturing spoken utterances. Inherent advantages of MFCC features can be taken into account for feature extraction. The test samples can be compared with the models created during the training phase for recognition by the ASR engine. These recognized utterances can be used for query generation obeying the syntactical structure of database used in backend. Their execution will produce the desired output that can be in terms of text or images.

Development of this type of system for Indian language in domains of general interest like agriculture, travel etc. would definitely help the Indian population who are not well versed with English in retrieving information from data store.

8. CHALLENGES IN ASR

Several types of signal (image, speech etc.) can be processed by human being and huge diversity lies between speech signals and other signals. This diversity is mainly due to variability persistent in speech, which is caused by uniqueness of vocal tract of each individual. The vast difference in anatomy and physiology between the speech production and perception systems in humans makes its analysis difficult. The dialectal variations of languages make its recognition challenging [36].

Another big challenge in the development of continuous speech recognition system is the word boundary detection. Determining the exact start and stops in continuous speech, i.e., classification as speech versus background noise or silence is very difficult. Noise robustness in ASR is a major challenge in current scenario. Noise may come from natural sources, machine as well as communication link distortion. All these noise tend to degrade system performance. Creating a system which can produce high recognition score in noisy environment is a major challenge in the area of speech research.

9. CONCLUSION

Speech recognition system has been in development for more than 50 years now. Number of practical limitations still exists which hinders the widespread deployment of such systems. Literature shows that the gap between human and machine recognition is still very large and current ASR give accurate performance for limited size vocabularies only. Need for large vocabulary speaker independent continuous speech has highly increased and several researches have been initiated towards filling this gap. Very few ASR system supporting Indian languages, especially Hindi exists [55] [56] and speech interface supporting Indian languages is very much in need. Based on the review it can be concluded that HMM approach along with MFCC features is more suitable for these requirements and offers good recognition result. These techniques will enable us to create increasingly powerful systems, deployable on a worldwide basis in future.

10. REFERENCES

[1] J. L. Flanagan, Speech Analysis, Synthesis and Perception, Second Edition, Springer-Verlag, 1972.

[2] B.H. Juang, Lawrence R. Rabiner, Automatic Speech Recognition – A Brief History of the Technology Development, Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005

[3] V. Zue et al, Jupiter: A Telephone-Based Conversational Interface for Weather Information, IEEE Trans. On

Speech and Audio Processing, Vol. X, pp. 100-112, Jan. 2000.

[4] J. Glass and E. Weinstein, Speech Builder: Facilitating Spoken Dialogue System Development, 7th European Conf. on Speech Communication and Technology, Aalborg Denmark, Sept.2001.

[5] Sadaoki Furui, Recent Advances in Spontaneous Speech Recognition and Understanding, ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, 2003

[6] K.-F. Lee, Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system, Ph.D. Thesis, Carnegie Mellon University, 1988.

[7] Hong C. Leung , Benjamin Chigier and James R. Glass, A comparative study of signal representations and classification techniques for speech recognition, Proc of IEEE International Conference(ICASSP-93) on Acoustics, Speech, and Signal Processing, 1993

[8] John Makhoul, Linear Prediction: A Tutorial Review ,Proc of IEEE, Vol 63, no 4 April 1975

[9] B S Atal and S L Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Formant Frequencies, Electronics and Communications in Japan, Vol. 53 A, pp. 36-43, 1970

[10] F. Itakura, Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans.Acoustics, Speech and Signal Proc., Vol. ASSP-23, pp. 57-72, Feb. 1975.

[11] Cristhian Manuel Durán Acevedo, Martín Gallo Nieves, Integrated System Approach for the Automatic Speech Recognition using Linear predict Coding and Neural Networks, Fourth Congress of Electronics, Robotics and Automotive Mechanics 2007.

[12] Antanas Lipeika, Joana Lipeikien ˙ E, Laimutis Telksnys,Development of Isolated Word Speech Recognition System , INFORMATICA, 2002, Vol. 13, No. 1, 37–46

[13] Cuntai GUAN, Yongbin CEBN and Boziu WU, Direct modification on LPC coefficients with application to Speech enhancement and improving the performance of Speech recognition in noise , Proc of IEEE international conference(ICASSP-93) on Acoustics, Speech, and Signal Processing, 1993

[14] Jialong He, Li Liu, and Gunther Palm, On the use of Residual Cepstrum in Speech Recognition, Proc of IEEE international conference(ICASSP-96) on Acoustics, Speech, and Signal Processing, 1996

[15] Xueying Zhang, Yueling Guo, Xuemei Hou, A Speech Recognition Method of Isolated Words Based on Modified LPC Cepstrum, 2007 IEEE International Conference on Granular Computing

[16] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. ASSP 28 (1980) 357--366.

[17] Deng, D. Yu, and A. Acero, Structured speech modeling, IEEE Transactions on Audio, Speech and Language Processing 2006, Vol. 14, No. 5, pp. 1492-1504

[18] Joseph Picone, Signal Modeling Techniques in Speech Recognition, Proc IEEE June199

[19] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617

[20] Akira Sasou, Futoshi Asano, Satoshi Nakamura, Kazuyo Tanaka, HMM-based noise-robust feature compensation, Speech Communication , Vol 48 (2006)

[21] Siva Prasad Nandyala , T. Kishore Kumar, Real Time Isolated Word Recognition using Adaptive Algorithm, Proc of International Conference on Industrial and Intelligent Information (ICIII 2012)

[22] Hubert Wassner and Gerard Chollet, New time frequency derived cepstral coefficients for automatic speech recognition, Proc of ICSLP 1996

[23] D. O'Shaughnessy, Speech Communications, IEEE Press, New York, 2000

[24] H Hermansky, Perceptual linear predictive(PLP) analysis of speech, J. Acoustic Society of America,Vol 87,No 4, 1990

[25] N Morgan, H Hermansky, H Bourlard, P Kohn, C Wooters, Continuous speech recognition using PLP analysis with multilayer perceptron, ICASSP'91, Proceedings of IEEE International Conference on Acoustic,Speech and Signal processing 1991.

[26] Corneliu Octavian DUMITRU , Inge GAVAT, A comparative study of feature extraction methods applied to continuous speech recognition in Romanian language, Proc of 48th International Symposium ELMAR-2006, 07-09 June 2006

[27] Cini Kurian, Kannan Balakrishnan, Malayalam isolated digit recognition using HMM and PLP cepstral coefficient, International Journal of Advanced Information Technology (IJAIT), Vol 1,No 5 2011

[28] A Revathi,R Ganapathy,Y Venkatramani, Text independent speaker recognition and Speaker independent speech recognition using iterative clustering approach. International Journal of Computer Science and Information Technology, Vol 1, No 2,2009.

[29] Hynek Hermansky, RASTA processing of speech, IEEE Transaction of Speech And Audio Processing, Vol 2, No 4, 1994

[30] Katrin Kirchhoff, Gernot A. Fink , Gerhard Sagerer, Combining acoustic and articulatory feature information for robust speech recognition, Speech Communication , Vol 37, 2002

[31] Brian ED Kingsbury, N Morgan, Steven Greenberg, Robust speech recognition using the modulation spectrogram, Speech Communication, Vol 25, 1998

[32] T Schurer, An experimental comparison of different feature extraction and classification methods for telephone speech, Proc of IEEE Workshop on Interactive Voice Technology for communications Applications, 1994

[33] Brian ED Kingbury,N Morgan, Recognizing reverberant speech with RASTA-PLP, Proc of ICASSP-1997

[34] Z.Hachkar, B.Mounir, A. Farchi, J. El Abbadi, Comparison of MFCC and PLP Parameterization In pattern recognition of Arabic Alphabet Speech, Canadian Journal on Artificial Intelligence, Machine Learning & Pattern Recognition Vol. 2, No. 3, April 2011

[35] Daniel Jurafsky, James H Martin, Speech and Language processing, Pearson education Inc.

[36] Douglas O'Shaughnessy, Automatic speech recognition: History, methods and challenges, Pattern Recognition ,Vol 41 , ELSEVIER, 2008

[37] Spector, Simon Kinga and Joe Frankel, Recognition, Speech production knowledge in automatic speech recognition,Journal of Acoustic Society of America, 2006

[38] M.A Zissman, Predicting,diagonosing and improving automatic Language identification performance,Proc.Eurospeech97,Sept.1997 ol.1,pp.51-54 1989.

[39] R. P. Lippmann, Review of Neural Networks for Speech Recognition, Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, 1990

[40] W. S. McCullough and W. H. Pitts, A Logical Calculus of Ideas Immanent in Nervous Activity, Bull. Math Biophysics, Vol. 5,1943.

[41] DongSuk Yuktt and James Flanagan, Telephone Speech Recognition Using Neural Networks and Hidden Markov Models, , Proceedings of IEEE International Conference on Acoustic,Speech and Signal processing 1999

[42] Judith Justin, Ila Vennila, Performance of Speech Recognition using Artificial Neural Network and Fuzzy Logic, European Journal of Scientific Research,Vol.66 No.1 (2011)

[43] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. ASSP 26 (1978)

[44] H. Silverman, D. Morgan, The application of dynamic programming to connected speech segmentation, IEEE ASSP Mag. 7 (3) (1990)

[45] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, D. Van Compernolle, Template-based continuous speech recognition, IEEE Trans. ASLP, 15 (2007) 1377--1390.

[46] Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C.Mehrotra,Marathi Isolated Word Recognition System using MFCC and DTW Features, Proc. of Int. Conf. on Advances in Computer Science 2010

[47] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257--286.

[48] X. Huang, Y. Ariki, M. Jack, Hidden Markov Models for Speech Recognition,Edinburgh University Press, Edinburgh, 1990.

[49] R. K. Aggarwal and M. Dave,Using Gaussian Mixtures for Hindi Speech Recognition System, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011

[50] Akira Sasou , Futoshi Asano , Satoshi Nakamura , Kazuyo Tanaka, HMM-based noise-robust feature compensation, Speech Communication, Vol 48 (2006)

[51] Carsten Meyer , Hauke Schramm, Boosting HMM acoustic models in large vocabulary speech recognition, Speech Communication, Vol. 48 (2006)

[52] Jen-Tzung Chien, Chuang-Hua Chueh, Joint acoustic and language modeling for speech recognition, Speech Communication 52 (2010)

[53] S. Young, et. al., the HTKBook, http://htk.eng.cam.ac.uk/

[54] William B. Cavnar and John M. Trenkle, N-Gram-Based Text Categorization, In Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1994

[55] Rajat Mathur, Babita, Abhishek Kansal,Domain Specific Speaker Independent Continuous Speech Recognition Using Julius, Proceedings of ASCNT–2010, CDAC, Noida, India, pp.55- 60.

[56] F. Reena Sharma and S. Geetanjali Wasson, Speech Recognition and Synthesis Tool: Assistive Technology for Physically Disabled Persons, Proc of International Journal of Computer Science and Telecommunications, Vol 3, Issue 4, pp. 86-91,April 2012.