FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES OF AUTOMATIC SPEECH RECOGNITION SYSTEM: A REVIEW

[1]Bhuvnesh Ratan Pippal, [2]Harsh Mohan Deshbandhu, [3]Hardik Kumar Prajapati, [4]Lucknesh Kumar

[1]Student, [2]Student, [3]Student, [4]Professor
[1]CSE Department,
[1]Galgotias College of Engineering & Technology,
Greater Noida, India

Abstract : Automatic speech recognition system are most prominent and important technique for human and computer interaction, controlling devices, extracting information. System user experience completely relies on accuracy, quality of speech notes and on the response time of the system. A system can be monolingual or multilingual, each works differently with different efficiency in accordance with the backend technique used. We have described about various ASR algorithms and feature extraction techniques and their comparison with each other in this paper. Feature extraction techniques like LPC, MFCCs, RASTA filtering, PLDA, LDA, PCA and algorithms like DTW, VQ, SVM, GMM, HMM.

IndexTerms **- SVM, VQ, DTW, GMM, ANN, MFCCs, ASR, RASTA filtering, LPC, PLDA.**

I. INTRODUCTION

Automatic speech recognition is the mechanism of converting speech signals to the series of words by the help of algorithm used on computer system. The objective of automatic speech recognition is to develop system for speech input signal to static modeling of speech. Recognition of speech is a special kind of pattern recognition.

Supervised pattern recognition consists of two phases, Training and Testing. The process of extracting the details and features related to classification is common in both phases. In the learning phase of the system, the parameters of the classification model are determine using a large number of data sets (Training Data). During the recognition or testing phase, the attribute of test pattern is matched with the learned model of each and every classification, If test data matches perfectly to model then it is declared that its belongs to that model[1]. Some of the applications of speech recognition include virtual reality, social media searches, multimedia searches, travel information and reservation, language translators and converters, natural language understanding and many more [4][5] . Now days speech recognition systems are based on HMM (Hidden Markov Models). The major advantage of using HMM is that HMM has parameters which can be learned automatically or trained and the procedures that are used for learning the system are less complex and are computationally reasonable to use [6]. Lot of research and development have been made time to time in speech recognition by machine but still not a single machine has been developed that understands the human speech by number of speakers in various environment and conditions. For multilingual users, one of the problem to system computer interaction is the common monolingual character and sequence of automatic speech recognition systems, users can interact in only a single preset language. According to some sources [1]–[3], multilingual speakers are much higher than monolingual speakers, and data analysis point to a higher number of multilingual speakers in the future. The ability to recognize multiple spoken languages is therefore a useful feature of ASR systems. In this paper we discuss about the major previous work in speech recognition over the years and describes about various feature extraction techniques and algorithms for automatic speech recognition.

II. SPEECH RECOGNITION TECHNIQUES

The main objective of automatic speech recognition system is to have ability to train from input data set, understand and then act on the spoken information. A speech recognition system mainly consist of four stages which are analysis, feature extraction, modeling, matching.

2.1 Feature Extraction Techniques

Feature extraction in speech recognition system is core part of the processing and also refers to as heart of the system. The work of this is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction compresses the magnitude of the input signal dimensionally without affecting the power of speech signal. There are many feature extraction techniques.

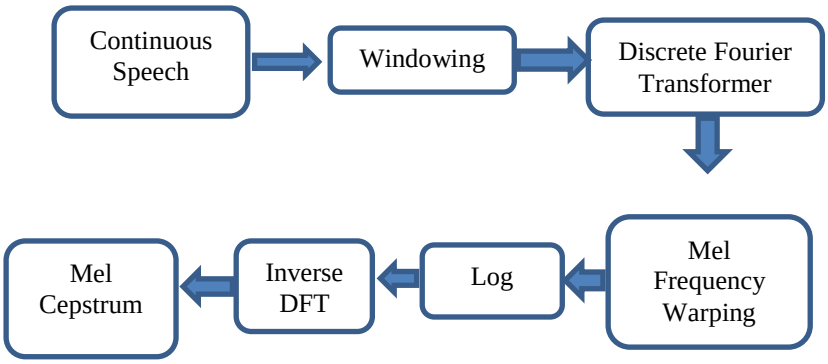

Fig 1: feature extraction diagram [11]

As from the diagram, we input the speech signals in continuous form for the process of windowing. In the windowing process the disruptions present at the start and end of the data frame are minimized. Then the continuous data of speech signal is transformed into windowed frames. After that these minimized windowed frames are loaded into the discrete Fourier transformer which transforms the minimized windowed frames into magnitude spectrum. After that, spectral analysis is performed with a fixed finite resolution along a subjective frequency scale i.e the Mel-frequency scale which generates a Mel-spectrum. The spectrum is then loaded to Log for processing and after that to inverse of discrete Fourier transform which produces Mel-Cepstrum. The Mel-Cepstrum consists of the attributes that are useful for speaker identification. Feature extraction techniques are:

2.1.1 Linear Predictive coding: LPC is a technique which is used for speech processing. LPC is based on an supposition: In a sequence of speech data samples, by which we can make the prediction about the nth sample then which can be represented by adding up the target signal's previously used data samples. Inverse filter should be done produced so that is analogous to the formant regions of the speech data samples. Application of filters into these data samples is the LPC process [7]. Some characteristics of LPC are, it provides auto-regression based features [8], it is a formant estimation and static technique [6] and in the process of LPC the residual sound is much close to the vocal tract loaded signal [7]. Major advantages of using LPC are, it is a reliable, robust and accurate technique which varies linearly according to the time which represent the vocal tract [9]. Computational speed is good and provides with accurate parameters of speech. It is highly for encoding speech at low bit rate. But this has some flaws like, is not able to distinguish the words with similar vowel sounds [10], unable to identify speech because of the supposition that signals are stationary that's why it is not able to examine the local events accurately, LPC produces residual error as output which means certain amount of useful speech gets left in the remaining part resulting in distorted speech quality.

2.1.2 Mel-frequency cepstrum (MFFCs): It is based on the variation of the human ear's critical bandwidth having frequency less than 1Khz. The most important purpose of the MFCC processor is to copy the behaviour of human ear, the steps of MFCC procedure the given in following below figure:-

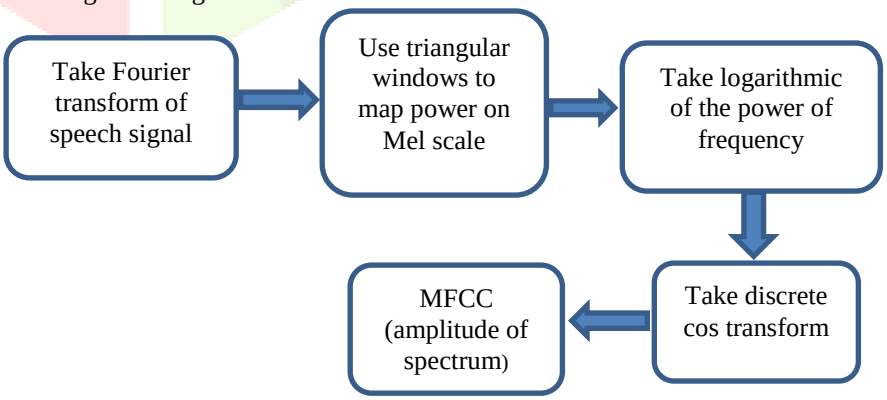

Fig 2: MFCC's functioning [18]

The main characteristics of MFCC is use for speech processing task. It is also capable to mimic the human auditory system. MFCC linear frequency spacing below 1 Khz and a log spacing above. The main advantages of MFCC is that is produce the high recognition accuracy that means performance rate of MFCC is high. It have low complexity algorithm. MFCC captures main attributes of phones in speech. The main disadvantage is that if there is any background noise then MFCC might not give accurate result. Sometimes performance get affected by the use of number of filter.

2.1.3 RASTA filtering: RASTA is short for relative spectral. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the notion of the data of speech signals in RASTA are band pass filtered. Initially it was just used to reduce the effect of noise in speech signal but now it is also used to directly improve the signal [12]. Figure 3 shows the functioning of RASTA The main idea here is to overcome the constant factors [14] .Is a band pass filtering technique. Designed to decrease the effect of noise and to enhance speech. That is, it is a technique which is widely used for the speech signals that have background noise or simply noisy speech. Eliminates the slow changing environmental variations as well as the fast variations in artefacts [13]. This mechanism does not depend on the alternative of microphone or the place of the microphone to the mouth, hence it is robust [14]. Captures frequencies with low modulations that correlate to speech [15]. This mechanism causes a minor decrement in performance for the clean information or a data but it also slashes the error in half for the filtered case [14]. RASTA combined with PLP gives a superior performance ratio [15].

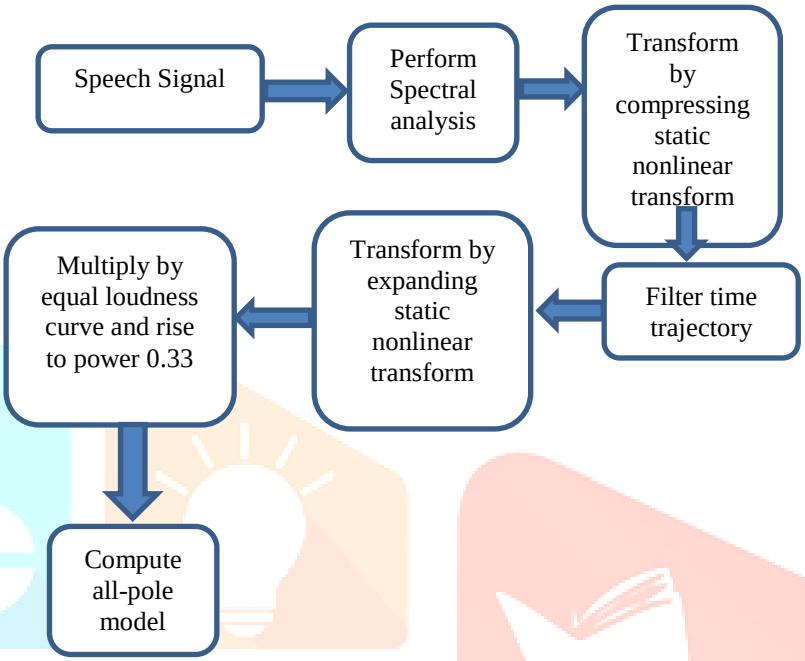

Fig 3: functioning of RASTA technique [14]

2.1.4 Probabilistic Linear Discriminate Analysis (PLDA): This technique is an extension for linear probabilistic analysis (LDA). Initially this technique was used for face recognition but now it is used for speech recognition. Based on i-vector extraction. The i-vector is one which is full of information and is a low dimensional vector having fixed length. This technique uses the state dependent variables of HMM. PLDA is composed of generative model. Is a flexible acoustic model which makes use of variable number of interrelated input frames without any need of covariance modelling [16]. High recognition accuracy. The Gaussian assumption which are on the class conditional distributions. This is just an assumption and is not true actually. The generative model is also a disadvantage. The objective was to fit the date which takes class discrimination into account [17].

2.1.5 Principal Component analysis (PCA): Principal component analysis (PCA) is mainly used as procedure or a technique for data compression without any loss of data or information. It is used to transform one set of variable into other smaller set, where the newly created variable is not easy to understand. In many applications, PCA is often used only to give information on the true dimensionality of a data set. If the data set includes X variables, all X variables may not represent all particular information. PCA changes a data set of correlated variables into a new data set of noncorrelated variables that are known as principal components. But if the information is already noncorrelated the PCA is of no use. Principal component analysis will be useful and can be applied to any data sets having any number of variables [19][20].

2.1.6 Linear discriminant analysis (LDA): Linear discriminant analysis is a widely known as method for evaluating a linear subspace with certain discriminative properties. The main objective is to find a projection of the data where the variance between the classes is highly compared to the variance within the classes.

LDA is supervised learning method, it means LDA required class definitions for the estimation process. One option is to use phonemes as classes [23], that is a popular option with MLPs (e.g. in [22]). With LDA, using hidden Markov model (HMM) states as the classes have been proven to give improved speech recognition performance [22]. Beulen et al. [24] showed that classes consistent with the HMM models are the best choice despite this large number of classes. Additional issue with LDA is the kind of the spectral data which is applied as the input for the method. The spectral information can be given out with logarithmic filter bank energies [23], some changed features like PLP [25] or spectral energies augmented with time derivatives [26]. LDA is unalterable to linear transformations, so there is no need of spectral transformations. Beulen et al [23] searched that

with Gaussian mixture densities the time derivatives do not change LDA over using the spectral energies. With the help of several research, we use in our research logarithmic mel-spaced filter bank energies as short-time spectral features.

2.2 ASR feature classification

The most common techniques used for speech classification are discussed in short. These system consist of complex mathematical functions and they extracts hidden details from the input processed signal.

2.2.1 Hidden Markov Model (HMM): It is the highly successful pattern used currently which recognition techniques in speech recognition. It is a mathematical model signalized on the Markov Model and a set of output distribution. This technique is more general and has a secure mathematical foundation as compared to knowledge based approach and template based approach. In this techniques the voice is divided into small audible entities so that it shows the state in hidden markov model. According to the probabilities of transition, there exists a transition from one state to another [6]. Hidden Markov model is widely used in speech recognition due to its ability to represent the time-varying aspect of speech signal. The origin of HMM is the famous Markov chain of probability theory which can be used for sequential modelling. HMM is a finite state automaton it has a limited number of states through which the machine makes transition from one state to another state. It also has a limited set of input and output symbols. When a machine is at states at time instant t, depending upon the input symbol provided, it will transit to another state emitting a certain output with a certain possibilities for the another time instant.

2.2.2 Dynamic Time Warping (DTW): Technique compares words with reference words. It is an algorithm to compute the resemblance between the two sequences that can vary in time or speed [6]. In this technique, the time dimensions of the unknown words are changed until they match with that of the reference word.

2.2.3 Vector Quantization (VQ): It is a procedure in which the mapping of vector is calculated from a large vector space to a limited number of region in that particular space. This procedure relies on block coding principle. Each region is known as cluster and that can be shown by its center known by code-word. Code book is the collection of all code-words [27].

2.2.4 Gaussian mixture model (GMM): A Gaussian mixture model is known as a probabilistic model that suppose all the data points that give rise to form a mixture of a limited number of Gaussian distributions with parameters unknown. The GMM is basically a density checker. Expectation maximization algorithm is mainly used to find out the mean, covariance parameters. During recognition, a series of features is extracted from the given input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The model that provides the highest likelihood score is verified as the identity of the speaker [28].

2.2.5 Support Vector Machine (SVM): It is known as a supervised learning algorithm which discriminative classifier mainly defined by a dividing each hyperplane given any labeled training data that is commonly known as supervised learning, the algorithm outputs an perfect hyperplane which distributed to form a new examples.

In this algorithm training tool is required before classification procedure get starts. The basic SVM requires the input set of data and analyze it, for each individual input. The hyper plane is formed by defining the weights Y, data point Z and offset o this implies $Y.Z + o = 0$ [29]. Since Y.Z is dot product of the data and the normal vector with respect to the hyper plane. The parameter o determines the offset value of the hyper plane at the origin along the normal vector [32].

III. CONCLUSION AND FUTURE SCOPE

Lot of research has been happened in the area of speech recognition but still the speech recognition systems are not hundred percent accurate till date. Systems developed till now have some limitations: Definite number of vocabularies in the current systems and work need to be done towards expanding this vocabulary, multilingual speech recognition problem system does not identifies speech from multiple user at the same time, the user should be in noise free environment for an accurate recognition, problem with the accent or tone and the pronunciation of the speaker. In near future the speech recognition systems need to be free from these flaws to give hundred percent accuracy and result. In this paper we give description of various speech recognition techniques. A speech recognition system should mainly consist of four stages: Analysis, Feature Extraction, Modelling and Matching techniques as described in the paper. Also, through this paper we show six techniques used in feature extraction: Linear Predictive Coding, Mel-frequency cepstrum, Relative Spectral, Probabilistic Linear Discriminate Analysis, Linear Discriminate Analysis and Principal component analysis. After studying each of these techniques we can say that each technique has its own advantages and disadvantages according to various use in different situation. Through research we come to the conclusion that Mel frequency cepstrum feature extraction technique is used widely in many speech recognition systems since it is able to mimic the speaker auditory system and provides better performance rate. And then we described about feature classification techniques like GMM, HMM, SVM, DTW and VQ. After research we can conclude that GMM performs better since it requires lesser amounts of data to train the classifier hence the usage of memory also decreases for the system.

REFERENCES

[1] G. Tucker and A. Tucker, “A global perspective on bilingualism and bilingual education, ser. ERIC (Collection),” ERIC Clearinghouse on Languages and Linguistics, 1999 [Online].

[2] F. Grosjean, Bilingual: Life and reality. Harvard, MA, USA: Harvard Univ. Press, 2010 [Online].
[3] D. Waggoner, "The growth of multilingualism and the need for bilingual education: What do we know so far?" Bilingual Res. J. vol. 17, no.1–2,pp.1–12,1993[Online].
[4] Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn "An Evaluation of AudioVisual person Recognition on the XM2VTS corpus using the Lausanne protocol" MIT Lincoln Laboratory, 244 Wood St., Lexington MA
[5] W. M. Campbell_, D. E. Sturim W. Shen D. A. Reynolds_, J. Navr´atily "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition" MIT Lincoln Laboratory IBM 2006.
[6] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yennawar, "A Review on Speech Recognition Techniques", IJCA Vol. 10, No. 3, pp. 16-24, November 2010.
http://dx.doi.org/10.5120/1462-1976
[7] Celso Auguiar, in CCRMA - Center for Computer Research in Music and Acoustics. Stanford University on Modelling the Excitation Function to Improve Quality in LPC's Resynthesis.
[8] Tomyslav Sledevic,˘ Artu¯ras Serackis, Gintautas Tamulevici˘ us, Dalius Navakauskas, International Journal of Electrical, Computer, Electronics and Communication on Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System Vol:7 No:12, 2013.
[9] Shanthi Therese Chelpa Lingam, International Journal of Scientific Engineering and Technology (ISSN: 2277-1581) a Review of Feature Extraction Techniques in Automatic Speech Recognition, Volume No.2, Issue No.6, pp: 479-484 1 June 2013.
[10] Navnath S Nehel and Raghunath S Holambe Journal on Audio, Speech, and Music Processing, on DWT and LPC based feature extraction methods for isolated word recognition, 2012.
[11] Wiqas Ghai and Navdeep Singh International Journal of Computer Applications (0975 – 8887) a Literature Review on Automatic Speech Recognition, Volume 41– No.8, March 2012.
[12] Hynek Hermansky, Eric A. Wan, and Carlos Avendano, Oregon Graduate Institute of Science & Technology Department of Electrical Engineering and Applied Physics, Speech enhancement based on temporal processing.
[13] Chia-Ping Chen Jeff Bilmes and Daniel P. W. Ellis, Department of Electrical Engineering University of Washington Seattle, WA on Speech Feature Smoothing for Robust ASR
[14] H. Hermansky and N. Morgan, Rasta processing of speech, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 578{589, 1994.
[15] Yuxuan Wang, Kun Han, and DeLiang Wang, Fellow, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, Exploring Monaural Features for Classification-Based Speech Segregation, 2012.
[16] Liang Lu, Member, IEEE and Steve Renals, Fellow, IEEE, IEEE SIGNAL PROCESSING LETTERS, Probabilistic Linear Discriminant Analysis for Acoustic Modelling, VOL. X, NO. X, 2014
[17] Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng and Frank Soong, at Center for Speaker and Language Technologies (CSLT), on I-vectors, a Discriminative Scoring for Speaker Recognition Based, 2014.
[18] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication 54 (4):543–565. doi:10.1016/j.specom.2011.11.004.
[19] Lawrence Rabiner, Biing Hwang Juang, B.Yegnanarayana, Fundamentals of Speech Recognition, First Impression 2009 by Dorling Kindersley (India) Pvt. Ltd.
[20] V Susheela Devi, M Narasimha Murty, Pattern Recognition-An Introduction, 2013, Universities Press(India) Private Limited.
[21] Qifeng Zhu, Barry Chen, Frantisek Grezl, and Nelson Morgan, "Improved MLP structures for data-driven feature extraction for ASR," in Proceedings of Interspeech, 2005, pp. 2129–2132.
[22] Sachin S. Kajarekar, B. Yegnanarayana, and Hynek Hermansky, "A study of two dimensional linear discriminants for ASR," in Proceedings of ICASSP, 2001, pp. 137–140.
[23] Klaus Beulen, Lutz Welling, and Hermann Ney, "Experiments with linear feature extraction in speech recognition," in Proceedings of Eurospeech, 1995, pp. 1415–1418.
[24] Panu Somervuo, Barry Chen, and Qifeng Zhu, "Feature transformations and combinations for improving ASR performance," in Proceedings of Eurospeech, 2003, pp. 477– 480.
[25] Reinhold Haeb-Umbach and Hermann Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in Proceedings of ICASSP, 1992, pp. 13–16.
[26] Kirandeep Kaur, Neelu Jain, "Feature Extraction and Classification for Automatic Speaker Recognition SystemA Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.
[27] Lindasalva Muda, "Voice Recognition Algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, Issue 3, March 2010.
[28] Kirandeep Kaur, Neelu Jain, "Feature Extraction and Classification for Automatic Speaker Recognition System A Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.
[29] A Survey on Speech Recognition Algorithms by Gaganpreet Kaur, Dr. Dheerendra Singh , Gagandeep Kaur, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-5)