A Detailed Survey on Large Vocabulary Continuous Speech Recognition Techniques

P Vanajakshi, [1]Assistant Professor,
Department of Computer Science and Engineering
Vivekananda Institute of Technology
Bangalore, India
vanaja_gowda@rediffmail.com

M. Mathivanan , Associate Professor
Department of Electronics and Communication Engineering
ACS College of Engineering
Bangalore, India
tmmathiv@yahoo.co.in

Abstract—**Speech recognition is a procedure of perceiving human speech by the PC and creating string yield in composed shape. A model is found out from an arrangement of sound recordings whose comparing transcripts are made by taking recordings of speech as sound and their content interpretations, and utilizing programming to make measurable meaning of the sounds that identify every word. Speech based applications are getting colossal prominence by joining Natural Language Processing (NLP) methods. Contribution to such applications is in common dialect and yield is acquired in regular dialect. If there should arise an occurrence of speech recognition, look into supporters are for the most part utilizing distinctive methodologies approach. ASR today finds across the board application in assignments that require human machine interfaces. This paper displays an audit on different existing procedures utilized in building ASR models and used to compare various speech recognition system methodologies and to find challenging issues in new research areas.**

Keywords-component; Modeling Approach, Acoustic-Phonetic approach, Pattern Recognition Techniques, Artificial Intelligence Approach, Language Modeling, ASR Tools, Automatic Speech Recognition (ASR), ASR grouping, Speech Analysis, Feature Extraction

I. INTRODUCTION

Programmed speech recognition framework can be characterized as free, PC driven translation of talked dialect into lucid content in genuine time [1]. Important applications of NLP (Natural Language Processing) are machine translation [2] and automatic speech recognition. ASR is innovation that permits a PC to distinguish the words that a man talks into a receiver or phone and change over it to composed content. Having a machine to see smoothly talked speech has driven speech look into for over 60 years. In spite of the fact that ASR innovation is not yet at the point where machines see all speech , in any acoustic environment, or by any individual, it is utilized on an everyday premise in various applications and administrations. A definitive objective of ASR research is to permit a PC to perceive progressively, with cent percent precision, every word which is understandably pronounced by any individual, free of terminology amount, commotion, speechifier qualities or complement. Today, if the framework is prepared to take in an individual speaker's voice, then much bigger vocabularies are conceivable and precision can be more noteworthy than 90%. Industrially accessible ASR frameworks ordinarily require just a brief time of speaker preparing and may effectively catch consistent speech with a huge vocabulary at ordinary pace with a high precision. Most business organizations assert that recognition programming can accomplish very near to 100% precision if worked under ideal situations. Ideal conditions' generally expect that clients have speech attributes which coordinate the preparation information, can accomplish legitimate speaker adjustment, and perform in a loud environment.

The primary objective of speech recognition domain to create methods and frameworks for speech contribution to machine. For reasons stretching out from inventive enthusiasm about the instruments for mechanical acknowledgment of human discourse abilities to desiring to motorize essential endeavors which requires human machine joint efforts and research in customized discourse acknowledgment by machines has pulled in a ton of thought for quite a while [3-4]. In light of significant advances in measurable demonstrating of speech, programmed speech verification frameworks nowadays find across the board application in errands that need man-machine interface, for example, programmed call handling in phone systems, question based data frameworks that give upgraded travel data, Stock value citations, Climate reports, Data section, Voice correspondence, Managing an account, Commands, Automobile entrance, Speech interpretation, Handicapped individuals (dazzle individuals) general store, Railroad reservations and so forth. Speech recognition innovation was progressively utilized inside phone systems to computerize and in addition to upgrade the administrator administrations. This report surveys significant highlights amid the most recent few years in the innovative work of programmed speech recognition, to give a mechanical point of view. Albeit numerous mechanical advances have been made, still there stay many research issues that inputs. It is relatively less complex and less demanding to actualize on the grounds that word limits are accessible and the words have a tendency to be obviously claimed which is the real favorable position of this sort. The detriment for this situation is that picking diverse limits influences the outcomes should be handled.

978-1-4673-8855-9/17/$31.00 ©2017 IEEE

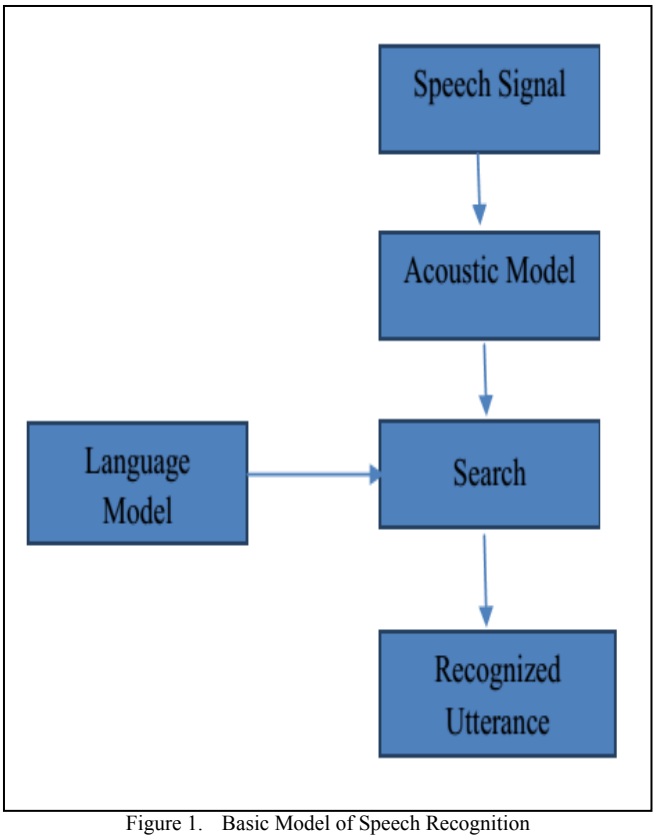

Figure 1. Basic Model of Speech Recognition

II. TYPES OF SPEECH RECOGNITION

Speech recognition frameworks can be characterized in a few unique classes [5] in view of the kind of speech articulation, sort of speaker model, kind of channel and the sort of terminology that they can perceive.

A. Based on Speech Utterance

An articulation is a solitary word or gathering of words that gives extraordinary intending to the framework. The words can take for verification into single or set of words for processing based on the requirement.

1) Isolated Words

Confined word recognizers as a rule require every articulation to have calm on both sides of the example window. It more often than not doesn't imply that it acknowledges single word, yet requires a solitary expression at once. This is useful for circumstances where the client gives just a single word reactions or charges, however is exceptionally unnatural if there should be an occurrence of numerous word.

2) Connected Words

Associated word frameworks (all the more effectively connected utterances) are like segregated words, however permit isolate expressions to go together with a negligible interruption among them.

3) Continuous Words

Ceaseless speech recognizers permit clients to talk actually, while the PC essentially decides the substance. It fuses a considerable measure of "co-clarification", where close-by words go together without stops or whatever different clear division between words. Consistent ASR frameworks are most hard to make since they should use uncommon systems to decide articulation limits. As vocabulary size develops bigger, disarray between various words successions likewise develop.

4) Spontaneous Words

Spontaneous speech is not practiced but rather regular. An ASR framework with unconstrained speech ought to have the capacity to handle a wide assortment of common speech elements, for example, Unconstrained speech may likewise incorporate errors, false-begins, and not-words.

B. Based on Speaker Model

All persons have their individual kind of voice input, because of their identical physical body nature. Speech recognition framework is comprehensively ordered into two principle bunches in light of speaker models to be specific speaker ward and speechifier free.

1) Speaker dependent models

Speaker subordinate frameworks are intended for a particular Speaker. In this model mostly more exact for the specific speaker, yet a great deal less exact for different speakers. These frameworks are typically less demanding to create, less expensive and have more exactness, yet are less adaptable than speaker versatile or Speaker free frameworks.

2) Speaker independent models

Speaker autonomous frameworks are produced for assortment of Speaker. It perceives the speech examples of a vast gathering of individuals. This sort of framework is most hard to grow, most costly and provides less exactness than Speaker subordinate frameworks. Be that as it may, these frameworks are more adaptable and tolerant.

C. Based on Vocabulary

1) Vocabulary Types

The span of vocabulary of a speech recognition framework impacts the unpredictability, handling prerequisites and the exactness of the framework. A few applications just need a couple words (e.g. numbers just), others require huge word references. In ASR frameworks the sorts of vocabularies are classified as 1.small vocabulary set - tens of words, 2. Medium vocabulary set – hundreds of words, 3. Large vocabulary set - thousands of words, 4. Very Large vocabulary set – tens of thousands of words.

III. RELATED WORK

D. Huggins-Daines[8] proposed a technique contains TIMIT database and MFCC or PLP feature extraction technique following event-based speech recognition system for English language as well as the performance is measured to HMM systems.

A. Rathinavelu [13] contains Small vocabulary Speaker independent phoneme as dataset and first five format values as feature extraction technique following neural network recognition technique for Tamil language with 81% accuracy.

G. Muhammad [9] having small vocabulary speaker dependent isolated digits as dataset and MFCC as feature extraction technique following HMM recognition technique for

Bangla language with 95% for digits (0-5) less than 90% for digits (6-9).

Y.-H. B. Chiu and R. M. Stern [10] will have DARPA resource management and MFCC feature extraction technique following CMU SPHINX-III as recognition technique for English language with improved speech recognition output accuracy is compared with MFCC modelin different background.

B. A. Al-Qatab [1] uses speaker dependent dataset, MFCC feature extraction technique and HMM recognition technique for Arabic language with 97.99% accuracy.

J. Ashraf [13] consists of minimum amount of vocabulary speaker independent isolated word and MFCC feature extraction technique following HMM recognition technique for Urdu language with minimum change in WER for new speakers.

M.A.Anusuya [13] contains a speaker subordinate preparing information, speaker autonomous preparing information, test and valuation information and PCA,MFFC,LDA are feature extraction technique for Kannada language with accuracy higher than 95%.

Hemakumar G [14] contains LM Domain like test, train data and MFCC as feature extraction technique following HMM as recognition technique for Indian languages with accuracy of 80% and 65%

E. Zavarehei [9] contains some speaker independent continuous data and Mel-Frequency CepstralCoefficients (MFCC) along with the feature extraction methods following Hidden Markov Model (HMM) recognition technique for Tamil language offers high performance.

H. Behravan [14] contains National foreign language certificate (FSD) corpus furthermore the English NIST 2008 SRE corpus and MFCC extraction technique following HMM recognition technique for English dialect with up to a 15% of the relative blunder reduction than the extremely solid vector arranged acknowledgment framework.

Hemakumar G [14] allows training or adaptation data and ICA, MFFC, LDA are feature extraction technique following HMM as recognition techniques for Kannada language with 75% to 95% accuracy.

M.A.Anusuya [13] having speech database,PRAAT software is used and MFCC for feature extraction techniques following HMM as recognition technique with error recognition accuracy has been decreased from 0.83 to 0.14for the speaker dependent application.

Hemakumar G [14] contains LM Domain like test, train data and MFCC as feature extraction technique following HMM as recognition technique for Indian languages with accuracy of 80% and 65% when the language models are used for Tamil speech recognizer.

IV. SPEECH RECOGNITION TECHNIQUES

The main aim of the speech recognition is to understand, identify and response or act based on the spoken information.

A. Modeling Technique

Speech recognizers are essentially decent concentrated measurable example recognizers. If there should arise an occurrence of factual displaying, a great choice of model based components is significant for the execution of the framework. For acoustic elements to be helpful in speech acknowledgment, they ought to incorporate much number of properties: they ought to be as spellbinding and distinctive as would be prudent, yet at the comparable time they ought to exclude intemperate excess. In addition, the speech recognizers are not for the most part moved in all the data the discourse flag contains. On the off chance that the objective or target is to separate the component words behind the speech, a wide range of helpful data about the speaker saw from his or her voice is not required, and the tone or accentuation of the speech. Truth be told, the less data about these futile qualities the acoustic elements contain, the less demanding it is to show the varieties of discourse we are really intrigued in and the better recognizers we can build. The Following are the exhibiting which can be used as a grammatical form acknowledgment handle.

On the other hand if there ought to be an event of speaker acknowledgment machine should amass speaker qualities in the acoustic banner. The essential purpose of speaker recognizing evidence is taking a gander at a discourse movement from a dark speaker to parameters of understood speaker. The system can see the speaker, which has been set up with different speakers where acknowledgment can similarly be apportioned into two methodologies, content ward and substance free procedures. In substance ward method the speaker say catchphrases or sentences having a comparative substance for both planning and acknowledgment trials, however message free does not rely on upon a specific works being talked. The Following are the exhibiting which can be used as a speech recognition handle:

1) The Acoustic-Phonetic Approach

This procedure is in reality sensible and has been analyzed in exceptional significance for over forty years. The approach is unending supply of acoustic phonetics and proposes. The soonest approaches to manage discourse acknowledgment relied on upon finding talk sounds and giving legitimate imprints to these sounds. It is the commence of the acoustic-phonetic approach (Hemdal and Hughes 1967) which recommends that there exist restricted, unmistakable phonetic units is talked lingo and that these units are widely depicted by a game plan of acoustics properties that are appeared in the talk movement after some time. Notwithstanding the way that, the acoustic components of phonetic units are extraordinarily consider, both with speakers and with neighboring sounds that it is normal in the acoustic-phonetic approach that the standards controlling the variability are immediate and can be expeditiously learned by a machine [7]. Formal evaluations drove by the National Institute of Science and Technology (NIST) in 1996 demonstrated that the most ideal approach to manage modified lingo recognizing confirmation (LID) uses the phonotactic substance of a talk banner to isolate among a plan of lingos.

2) Pattern Recognition Approach

The example coordinating methodology (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) includes two basic strides in particular, design preparing and design correlation. The basic component of this approach is that it utilizes an all-around planned numerical structure and sets up steady speech design representations, for solid example correlation, from an arrangement of named preparing tests by means of a formal preparing calculation. For example recognition has been created more than two decade got much consideration and connected broadly an excessive number of pragmatic example recognition issue. A discourse outline representation can be as a discourse format or a quantifiable model (e.g., Hidden Markov Model or HMM) and can be associated with a sound (more diminutive than a word), a word, or an expression. In the case relationship period of the approach, a prompt examination is made between the dark converses with each possible case learned in the planning arrange remembering the end indicate choose the identity of the dark as showed by the tolerability of match of the cases. The example coordinating methodology has turned into the prevalent strategy for speech recognition in the most recent six decades.

3) Knowledge Based Approaches

A specialist learning about varieties in speech is extracting features into a framework. This has the benefit of unequivocal displaying varieties in speech; yet lamentably such master information is hard to acquire and utilize effectively. Along these lines this approach was confirmed to be unfeasible and customized learning procedure was searched for the Vector Quantization (VQ) and is every now and again associated with ASR. This is accommodating for discourse coders, i.e., gainful data diminishment. Since transfer speed is not an important issue for ASR, but rather the utilization of VQ here in the efficiency of using insignificant codebooks for unique models and codebook searcher set up of all the more costly evaluation strategies. For Isolated word recognition, each vocabulary word gets its own VQ codebook, in perspective of get ready progression of a couple of redundancies of the word. Here the test discourse is evaluated by greatest number of codebooks and ASR picks the components whose codebook yields the most negligible partition measure.

4) The Artificial Intelligence Approach

a) The fake cognizance approach attempts to designed the acknowledgment method According to the way a man applies its knowledge in envisioning, analyzing, in conclusion settling on a decision on the consider acoustic parts. Ace structure is used by and large as a piece of this approach. In this Artificial Intelligence approach which contains just 50% of the acoustic phonetic approach and illustration acknowledgment approach. In this, enterprises the musings and thoughts of Acoustic phonetic and illustration acknowledgment techniques. Learning based approach uses the information as for semantic, phonetic and spectrogram. Some discourse researchers made acknowledgment system that used acoustic phonetic figuring out how to make arrange rules. While layout based methodologies have been exceptionally compelling in the plan of an assortment of speech recognition frameworks; they gave little understanding about human speech handling, along these lines making mistake investigation and information based framework improvement troublesome. In its immaculate shape, learning building configuration includes the immediate and unequivocal fuse of master speech information into a recognition framework. This information is normally gotten from watchful investigation of spectrograms and is joined utilizing tenets or methods.

Immaculate learning designing was additionally propelled by the premium and research in master frameworks. Nonetheless, this approach had just restricted achievement, generally because of the trouble in evaluating master learning. Another troublesome issue is the coordination of many levels of human learning phonetics, phonotactics, and lexical get to, sentence structure, semantics and pragmatics. Then again, joining free and non-agreeing data sources in a perfect world remains an unsolved issue. In more deviant structures, learning has in like manner been used to deal with the arrangement of the models and computations of various strategies, for instance, format planning and stochastic showing. This type of learning application makes an imperative refinement amongst information and calculations. Calculations empower us to take care of issues. Information empowers the calculations to work better. This type of information based framework improvement has contributed extensively to the outline of all effective procedures reported. It assumes a vital part in the choice of an appropriate info representation, the meaning of units of speech, or the outline of the recognition calculation itself.

B. Matching Techniques

The Speech recognition engine is outlined like contrasting a recognized word with deference with a known word by any of the accompanying systems.

1) Whole-word matching

The engine breaks down the moving toward cutting edge sound banner against a prerecorded configuration of the word. This framework takes considerably less get ready than sub-word organizing, yet it requires that the customer (or some person) prerecord every word that will be seen as a less than dependable rule a couple of hundred thousand words. Entire word layouts additionally require a lot of capacity (somewhere around 50 and 512 bytes for each word) and they give very effective results if the identified vocabulary is known.

2) Sub-word Matching

The engine scans for sub-words which is typically phonemes and a while later performs furthermore outline affirmation on those. This technique takes more get ready than whole word organizing, nonetheless it needs significantly less limit (some place around 5 and 20 bytes for each word). What's more, the articulation of the speech can be speculated from English content without the need of the client to talk the word in advance.

V. ASR ARCHITECTURE.

A. Preprocessing

Speech signal is continuous time varying analog signal, which cannot be understand by the digital devices, therefore this analog signal should be converted into digital format that

can be recognized by the system. In the preprocessing step, the digitized speech signal is filtered to remove the unnecessary high frequency noise signals and also the dc offset. Further the continuous speech signal is segmented at regular pause intervals by obtaining the Spectral Centroid and the spectral energy. The obtained parameters are compared with the threshold to detect the pause in signal. These segmented signals are further used for feature extraction.

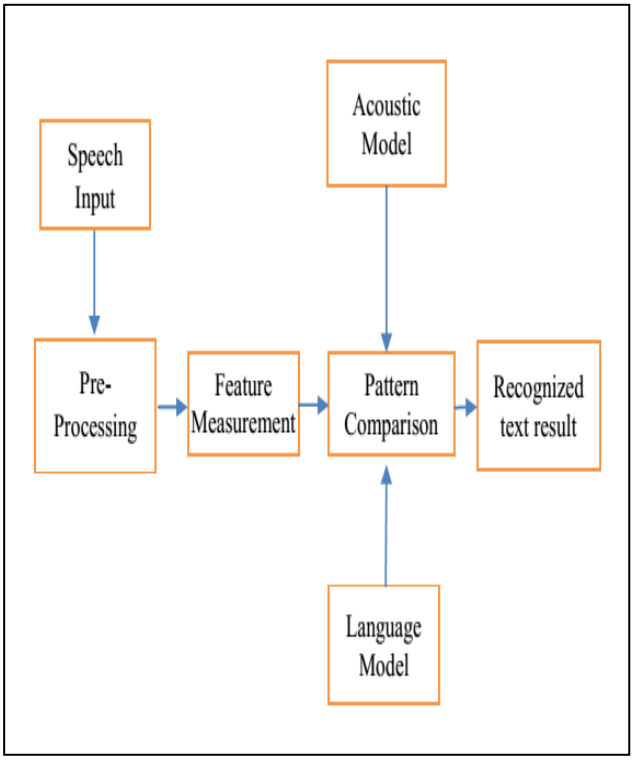

Figure 2. Structure of proposed ASR

B. Feature Extraction

Feature extraction is also called feature measurement; an arrangement of estimations is connected on the info flag to call as the "test pattern". The feature measurements of the speech gotten from the yield of otherworldly examination procedure, such as a channel bank analyzer, a straight prescient coding investigation, Mel-frequency Cepstral Coefficients or a Discrete Fourier Transform analysis. The different feature extraction techniques are as shown in below Figure3.

This research mainly focuses on obtaining the features form the input speech signal using Mel-frequency Cepstral Coefficients (MFCC). Mel-frequency scale or Mel scale is very nearly equal to human auditory system consider to be one of the advantage of this technique.

The Mel scale is a logarithmic scale pulls the features from the input speech signal and also for increasing the recognition rate.

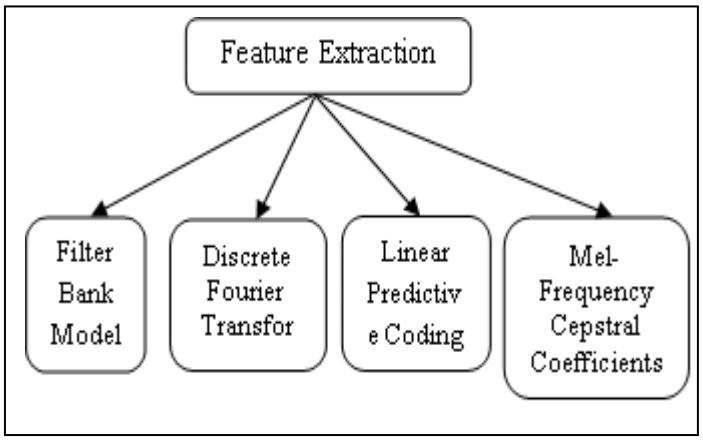

Figure 3. Feature Extraction Techniques

C. Pattern Classification

The obtained train features are compared with the test features using pattern classifier to recognize the test signal. To compare speech patterns the local distance measure (LDM) can be utilized. The local distance is calculated with the support of the spectral "distance" between the space of spectral vectors.

The rms log spectral distance (final single value) is given by [8],

$$dtf_{rms} = \sum_{n=-\infty}^{\infty} (c_n - c_n')^2$$

Where Cn Train signal features

Cn' Test Signal features.

D. Acoustic model

An An acoustic model is represented into a group of sound data that forms up a word. From these statistical representations phoneme is obtained. An acoustic model is created by forming the statistical representation of data into called a speech corpus by using training algorithms for language. These interpreted representations are called HMM Model. Each phoneme in a language has its own HMM.

E. Language Model

Language Model illustrated with a set of continuous words, such language models can include some important limitations of the languages and the recognition task.

The probabilistic model of discourse acknowledgment where W creates an acoustic watched grouping O, with likelihoo P(W, O).The Objective is then to change the relating word, in view of the acoustic perception succession, which will frame the string has the most extreme a back MAP probability[6]

$$P(W/O) = \arg\max_w WP(W/O)$$

Utilizing Bayes rule [6]

$$\frac{P(O/W) * P(W)}{P(O)}$$

P (W/O) =

Where P (O/W) is called Acoustic model

P (W) is called Language Model.

VI. TOOLS FOR ASR

The summarization of apparatuses tools utilized for performing the ASR [8].

HTK [11]: The fundamental utilization of open source Hidden Markov Toolkit (HTK), composed totally in ANSI C, is to manufacture and control shrouded Markov models.
SPHINX [12]: Sphinx 4 is a most recent variant of Sphinx arrangement of speech recognizer instruments, composed totally in Java programming dialect. It gives a more adaptable structure to investigate in speech recognition.
KALDI [15]:The principle commitments are the incorporation of Kaldi to a comprehensive assessment of open-source ASR frameworks and the use of standard corpora making our outcomes comparable to different distributions in the field. A free and open-source discourse recognition is a toolbox. The toolbox right now bolsters demonstrating of setting ward telephones of subjective setting lengths and all ordinarily utilized procedures that can be assessed utilizing most extreme probability.

VII. PEFORMANCE ANALYSIS

The Performance of speech recognition is measures as far as acknowledgment precision and execution speed. The acknowledgment exactness is otherwise called word error rate (WER).The execution investigation should likewise be possible by utilizing phoneme error rate (PER).

A. Word Error Rate

Most of the common metric the word error rate is used in speech recognition performance. Due to different length of word sequence, it is difficult to measuring the performance [9].

WER = $\frac{SUB+DEL+INS}{N}$

Where SUB is add up to number of substituted words

DEL is add up to number of erased words

INS is add up to number of embedded words

N is add up to number of words.

Performance of speech recognition can likewise be assessed utilizing word acknowledgment rate (WRR) is used instead of WER.

WRR = 1- WER

$$WRR = 1 - \frac{SUB+DEL+INS}{N}$$

B. Real Time Factor

The speed of recognition is measured in the form of real time factor. The processing time is Tp and an input of duration is Dt then, the real time factor is given by the equation [9],

$$RTF=\frac{T_p}{Dt}$$

C. Phoneme Error Rate (PER)

Phoneme error rate is computed by the proportion of misclassified phonemes to the aggregate number of testing phonemes utilized.

VIII. EXPERIMENTAL SETUP

The database contains speech signals of 92 speakers having age ranges between 18 to 30, out of which 40 signals belongs to male and remaining 50 signals belongs to female. These speech signals are captured by using sensitive microphone by using sound recorder software in a laboratory environment. The signals are acquired in both noisy and noiseless environment. The pause is removed from all the samples by calculating spectral centroid and spectral energy which is stored in the wave format files and the sampling rate is 16KHz.and signed 16 bits with monotype.
The proposed large continuous speech recognition system acquires the continuous Kannada speech signal. The acquired signal is pre-emphasized to energy in the signal at higher frequencies to remove the pause.
The below figure 4 shows the Input Signal

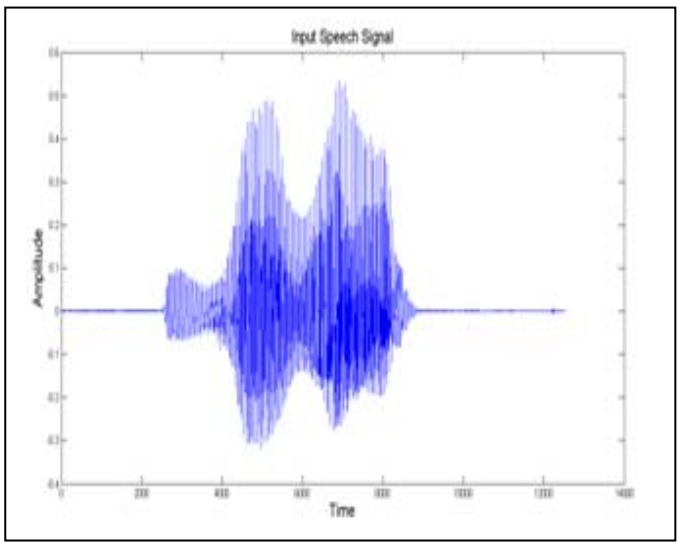

Figure 4. Input Signal

Hamming window is applied for the pre-emphasized signal to observe a signal in a finite time.

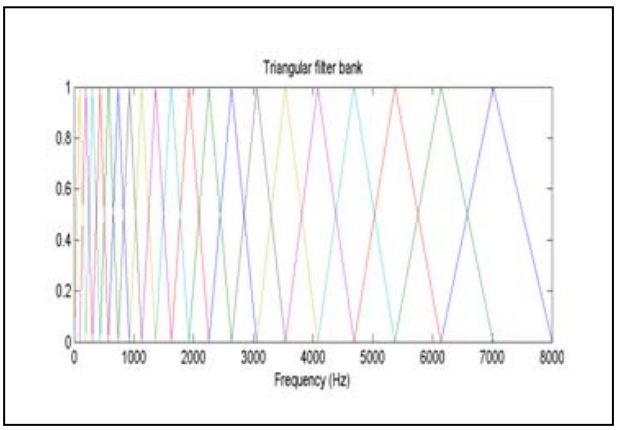

Figure 5. Mel–Frequency Filter Bank

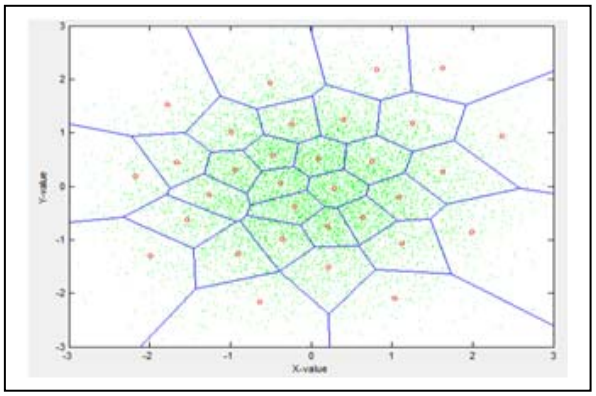

Figure 6. An example of a 2-Dimentional VQ

From the figure 5, the Mel-frequency filter bank output is displayed for various filter banks with respect to frequencies.
Figure 6, stars are called code vectors and the regions having borders called encoding regions. The set of code vectors is called codebooks and set of encoding regions are called the partitions of the space.
Vector Quantization is a fixed to fixed length algorithm and based on the principle of block coding, lossy data compression method have been developed. A region having Star shape shows every pair of numbers.

IX. CONCLUSION

Speech based applications are getting greatly influence as they turn out to be fundamentally instructed. Nowadays part of research is being done and additionally parcel of work should be done with regards to ASR for a specific type of an Indian dialect which is indistinguishable to a particular geographical range. In this paper explains different ASR systems and have advanced a portion of the fundamental data and essentials for the same by studying a little part of the mammoth work yet to be done in this field. We have quickly examined the Speech Recognition System and different methodologies utilized as a part of ASR created for different form of an Indian language which is particular to a specific area. Concealed Markov Model and Hidden Markov Model Toolkit (HTK) has been utilized generally.

REFERENCES

[1] B. A. Al-Qatab and R. N. Ainon.Arabic speech recognition using Hidden Markov Model toolkit (htk). in Information Technology (ITSim) 2010 International Symposium in, vol. 2, pp. 557–562, IEEE, 2010.

[2] J.H.Martin and D.Jurafsky. Speech and language processing. International Edition, 2000.

[3] Sadaoki Furui. 50 years of Progress in speech and Speaker Recognition Research, ECTI Transactions on Computer and Information Technology, Vol. 1, No. 2, November 2005.

[4] B.H. Juang, Lawrence R. Rabiner. Automatic Speech Recognition – A Brief History of the Technology Development. Georgia Institute of Technology, Atlanta and Rutgers University and the University of California, Santa Barbara.

[5] B. S. Atal and L. R. Rabiner. A pattern recognition approach to voiced unvoiced-silence classification with applications to speech recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 24, no. 3, pp. 201–212, 1976.

[6] Taabish Gulzar, Anand Singh, Dinesh Kumar Rajoriya and Najma Farooq. A Systematic Analysis of Automatic Speech Recognition: An Overview., IJCET, 2014.

[7] P.Satyanarayana.Short segment analysis of speech for enhancement. Institute of IIT Madras Feb 2009.

[8] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravi Shankar, A. Rudnicky, et al.Pocket Sphinx: A free, real-time continuous speech recognition system for hand-held devices. IEEE International Conference on in Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings, vol. 1, pp. I–I..

[9] G. Muhammad, Y. Alotaibi, M. N. Huda, et al. Automatic Speech Recognition for Bangla digits. in Computers and Information Technology, 2009. ICCIT-09. 12th International Conference on, pp. 379–383, IEEE, 2009.

[10] Y.-H. B. Chiu and R. M. Stern.Minimum variance modulation filter for robust speech recognition. in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pp. 3917–3920, IEEE, 2009.

[11] J. Ashraf, N. Iqbal, N. S. Khattak, and A. M. Zaidi. Speaker independent Urdu speech recognition using hmm. in Informatics and Systems (INFOS), 2010 The 7th International Conference on, pp. 1–5, IEEE, 2010.

[12] H Behravan, V.Hautamaki, S.Siniscalchi, T.Kinnunen and C-H, Lee,.i-vector modelling of speech attributes for automatic foreign recognition.

[13] 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

[14] Hemakumar G, Punitha P. Speech Recognition Technology: A Survey on Indian Languages., International Journal of Information Science and Intelligent System, Vol.2, No.4, 2013.

[15] Daniel Povey, Arnab Ghoshal. The Kaldi Speech RecognitionToolkit.,ICASSP2010