Speech Recognition: A Review of Literature

Kirandeep Singh

Department of Computer Science
Punjabi University Patiala, Patiala, India

Abstract— *Speech recognition is a process of identifying what a person speaks into a mike or any other similar hardware and reflects its meaning in any required form such as text, image or any event. This thesis provides a description of implementation of Speaker Independent Isolated Punjabi and English Digits Recognition system. The system is developed by using two different techniques, first is pattern based technique (DTW (Dynamic Time Wrapping)) and second is statistical based technique (HMM (Hidden Markov Model)). The system uses the Mel Frequency Cepstral Coefficients (MFCCs) technique for the purpose of features extraction. The developed system works for Punjabi as well as English digits recognition.*

Index Terms— *Speech Recognition, Acoustic Vector, Mel-Frequency Cepstrum Coefficients, Hidden Markov Model, Fast FourierTransform.*

I. INTRODUCTION

Speech is a basic unit of communication for humans. In this decade computer technology is drastically evolving, this era is an era of computer revolution. So to make communication easier between human and computer, speech recognition provides a great role. Speech recognition is a technique in which a human speaks to computer in his/her comfortable language and computer is such an intelligent to understand his/her spoken words and respond accordingly. Speech recognition technology helps to convert recognized and spoken language into text, image or any event stirring by computers and other computerized devices such as Smart Technologies and robotics.

At present, a lot of researches are going on for the development of much robust speech recognition systems. There are many exciting tools like HTK, KALDI, CMU SPHINX and others, used to develop speech recognition systems. Mobile phones, computers, ATM machines, household appliances and many others use speech recognition for the purposes of Speech to Text Conversion, specific event stirring, command and control, call center automation, voice calling, voice navigation on desktop, voice browser for internet, voice dialer, voiced controlled wheel chair and many other applications.

Typically, Speech recognition system has two phases first is training phase and second is testing phase. First step in speech recognition is to convert incoming human sound that is in analog signal, into digital signal or digital form. Digitized sampled signal is complex for direct processing by a system, so it needs to extract speech features from it. Many options available for feature extraction, such as, Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC), and others.

After feature extraction from a speech signal, next step in recognition is to compare these computed features with trained patterns in database to find a spoken word, this phase is known as testing phase. Different techniques exist for comparison step such as DTW, HMM, Neural Network, vector quantization and others. In final step, if computer correctly recognizes a spoken word then it would be used for any event stirring or print recognized text on screen or this output can work as input to further language processing as per user requirements.

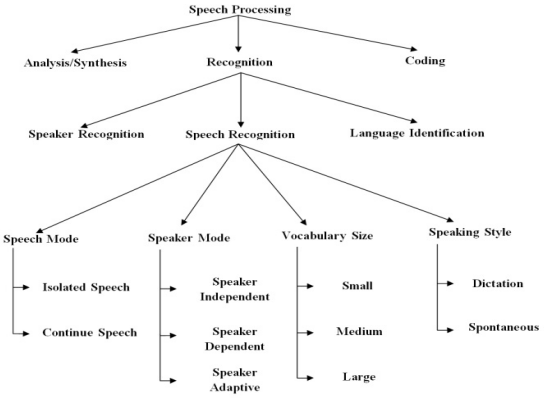

Fig. 1 Speech Recognition Systems Classification

II. AUTOMATIC SPEECH RECOGNITION SYSTEM

A. Speech recognition system can be developed for the grammatical structure and some statistical model can be used to improve word predication, but still there is a problem that how much world knowledge of speaking and encyclopedia can be modeled? Of course, we cannot model the world knowledge. So we cannot measure computer system up to human comprehensive.

B. Only speech does not participate in human communication, even some body signals are also used such as hand waving, eye moment and others. Consequently in any ASR system such information is completely missed.

C. Any unwanted information in any sound signal is a noise. While speaking in any environment, a radio playing somewhere downs the corridor, a clock ticking, another human speaker in the background are all examples of noise. ASR should be intelligent enough to detect such noise and filter it out from the speech signal.

D. Written language and spoken language are essentially different in nature. Written language is one way communication while spoken language is dialog oriented. In spoken language we give feed back to the sound that we understand. So in last few years it has been observed that spoken language is grammatically less complex whereas in written language, grammatical possibilities should always be kept in mind. As normally speech contains repetitions, slips of the tongue, changes of subject in the middle of phrase, hesitations etc. such disfluencies are commonly ignored by human listener. In ASR, such kind of behavior should be represented by the machine and these differences should be identified and addressed carefully.

E. Communication does not have natural pause between words of a spoken sentence, usually pauses comes at the beginning and end of a speech. ASR should be capable to convert a sound wave into a sequence of spoken words.

F. All persons in this world have their special voices; because of their distinctive physical body. There are some variations, even within a one specific speaker, listed below.

(i) If a person speaks same word again and again then there will definite be a small variation in same spoken word. The realization of sound changes over time.

(ii) All humans speak according to place and their emotions. For example a person speaks differently with parents, with friends, with teachers, in banks, in market, same as speaking style also varies on expressions. We speak differently when we are happy, sad, stressed, frustrated, disappointed etc.

(iii) Different aged persons have different speaking style and different speaking sound. Some persons have different speaking sound at different ages.

III. DIFFERENT SPEECH RECOGNITION APPROACHES

A. ***Template Matching:*** In TM all information contained in the templates is kept and used to recognize the pronounced word, no priori assumptions are made and a word can be identified by only a few samples. When a pronunciation dictionary is not available and there are only a few samples per word, template matching (TM) seems to be the most suitable approach. In template based approach [1], a collection of prototypical speech patterns is stored as a reference pattern. Whenever an unknown spoken utterance comes, it is matched with the each stored reference pattern and the pattern having best match is selected. The unknown speech pattern is compared against each reference pattern and measure of similarity (distance) between test pattern and reference pattern is computed. This approach has the advantage in its simplicity and uses perfectly accurate word models. Normally templates for entire words are constructed so errors due to segmentation and classifications can be avoided [2]. The system is insensitive to sound class so the basic techniques developed for one sound class can be easily applied to another sound class with little or no modification. The disadvantage is that the prerecorded templates are fixed so for speech variations, we have to store many templates per word. The speaking environment and characteristics of the transmission medium can affect the efficiency of reference patterns. However, its generalization capabilities are weak

and its performances are not as competitive as HMM-based approaches.

B. ***Stochastic Approach*:** Stochastic modeling is based on probabilistic modeling that deals with uncertain and incomplete information. In speech recognition, uncertainty and incompleteness take place from many sources; for example, variability in speakers, confusable sounds, homophones words, and contextual effects. The most popular stochastic approach is hidden Markov modeling. Thus, stochastic approach is predominantly appropriate approach to speech recognition. A Hidden Markov Model is described as a finite state Markov model and a set of output distributions. Hidden Markov modeling has more solid mathematical foundation as compared to template based approach. HMMs facilitate easy integration of knowledge sources into a compiled architecture. A negative side effect of this is that it does not afford much insight on the recognition process. Therefore, it is often complicated to analyze the errors of an HMM system in an attempt to improve its performance.

C. ***Vector Quantization (VQ):*** ASR frequently uses Vector Quantization (VQ) [3]. It is valuable for speech coders, i.e. efficient data reduction. While transmission rate is not a major issue for ASR, so usefulness of VQ lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. In isolated word recognition, each different vocabulary word has its own codebook that is trained by multiple repetitions of that specific word. During the testing of an incoming word, all codebooks are evaluated and ASR system chooses the codebook that raises the lowest distance measure. Basically VQ does not have any time related information (e.g., the temporal order of phonetic segments in each word and their relative durations are ignored), as codebook entries are not ordered and can come from any part of the training words. The average distance, across all training frames that are corresponding to longer acoustic segments, is minimized by selecting codebook entries.

D. ***Neural Network:*** The artificial intelligence approach [4] is the way to automate the recognition procedure likewise a person applies his intelligence in visualizing, analyzing, and characterizing speech based on a set of measured acoustic features. This approach [5] has not been extensively used in commercial systems. The spotlight in this approach has been mostly in the representation of knowledge and integration of knowledge sources. Connectionist models significantly depend on the good learning or training strategies. In connectionist models, knowledge or constraints are scattered across many simple computing units as an alternative to encode them in individual units. Nature of computing units is simple, and knowledge is not programmed into any individual unit function; rather, it lies in the connections and interactions among linked processing elements.

E. ***Support Vector Machine (SVM):*** SVM [6] is a tool for pattern recognition that uses a discriminative approach. For data classification purpose, it uses non-linear and linear separating hyper-planes. This approach cannot be willingly applied to task that involves variable length data classification as it can only classify fixed length data vectors. Before SVMs can be used, variable length data should be transformed into fixed length vectors.

F. ***Artificial Intelligence Approach (Knowledge Based Approach):*** The Artificial Intelligence approach is a hybrid approach that incorporates the thoughts and ideas of acoustic phonetic approach and pattern recognition approach. Knowledge based approach utilizes the information on the subject of phonetic, spectrogram and linguistic. Knowledge engineering design involves the direct and explicit incorporation of expert's speech knowledge into a recognition system [7]. This knowledge usually comes from suspicious study of spectrograms and is incorporated by using rules or procedures. Speech recognition system uses classification rules for speech sounds by using acoustic phonetic knowledge. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand. The problem in limited success of this approach is due to the difficulty in quantifying expert knowledge and incorporation of many levels of human knowledge such as syntax, pragmatics, lexical access, semantics, phonetics and phonotactics. In more indirect forms, knowledge has also been used to design

the models and algorithms of template matching and stochastic modeling techniques. Artificial intelligence approach provides a great role in different phases of speech recognition, such as designing of recognition algorithm, exhibition of speech and depiction of suitable and appropriate inputs units. Artificial intelligence is mainly dedicated to develop such kind of machines that are capable to simulate the behaviors of human beings. Artificial machine collects information from their respective environments and respond in an intelligent manner, calculating appropriate and adequate steps, formulating answers, and present desired results [8]. Other applications of artificial intelligence in different areas are robotic equipment, smart phones, video games, signals and traffic lights, credit card transactions and home security etc.

IV. LITERATURE SURVEY

The earliest attempts in speech recognition were made during 1950 and 1960s.

In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built an isolated digit recognition system [9] for a single speaker using the formant frequencies measured/estimated during vowel regions of each digit.

In 1956 at RCA Laboratories, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words [10].

In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer [10] to recognize four vowels and nine consonants. They used spectrum analyzer and pattern matcher for the recognition. By incorporating statistical information, they increased the overall phoneme recognition accuracy for words consisting of two or more phonemes. Their work marked the first use of statistical syntax in automatic speech recognition.

In 1960s, Martin and his colleagues at RCA Laboratories developed a set of elementary time-normalization methods [11] to detect speech starts and ends that significantly reduced the variability of the recognition scores. At the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods generally called dynamic time warping for time aligning a pair of speech utterances including algorithms for connected word recognition.

In 1970, the area of isolated word or discrete utterances became a viable and usable technology based on the studies in Russia and Japan. The Itakura of Bell laboratories [10] introduced that through the use of an appropriate distance measure based on LPC spectral parameters, linear predictive coding (LPC) could be used in speech recognition. Also researchers here, started experiments aiming at making speaker independent systems. A wide range of clustering algorithms was used to achieve this goal. In 1973, Hearsay I system by CMU was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. CMU's Harpy system was able to recognize speech using vocabulary of 1011 words with reasonable accuracy. These projects were funded by DARPA (Defense Advanced Research Projects Agency).

In 1980, there was a shift in methodology from template based to more rigorous statistical modeling framework. One of the key technologies was Hidden Markov Model (HMM) although this technique became widely applied in mid-1980s. Furui proposed the use of cepstral coefficients as spectral features in speech recognition. The n-gram model defining the probability of occurrence of an ordered sequence of n words was introduced by IBM for large vocabulary speech recognition systems. The primary focus was the development of a language model which describes how likely a sequence of language symbols appear in a speech signal.

In 1990's DARPA program was continued. The emphasis was laid on the different speech understanding application areas such as transcriptions of broadcast news and conversational speech. The BN transcription technology was integrated with information extraction and retrieval technology, and many application systems, such as automatic voice document indexing and retrieval systems, were developed [12]. Various other techniques were developed viz. the maximum likelihood linear regression (MLLR) , the model decomposition, parallel model composition (PMC) , and the structural maximum a posteriori (SMAP) method to reduce the mismatch caused by background noise , microphones , voice individuality etc.

Rebiner and Sambur have proposed "A Statistical Decision Approach to the Recognition of Connected Digits" [13]. Each utterance which was a string of three digits, was first analyzed to find end points and a voiced-unvoiced-silence part of the utterance was obtained. The digit string was then segmented into individual digit based on the voiced-unvoiced-silence information. The voicing region in each segmented digit is analyzed using linear predictive coefficients (LPC). The LPC coefficients are converted to parcor or reflection coefficients and linearly warped to compute average digit length. The recognition of each digit within the string is done by using a distance measure based on minimal residual error. This measure also takes into account the effect of co articulation and multiple repetitions. This system can be used for both speaker independent and speaker-dependent situations. The recognition system has been tested on six speakers in the speaker-dependent mode. The accuracy achieved is 99 percent. In speaker-independent mode, the system was tested with 10 new speakers and reported accuracy was 95 percent.

Rebiner and Wilpon proposed "simplified, robust, training procedure for speaker trained, Isolated word recognition systems" [14]. The method has been proposed in order to overcome the extensive burden of training required in statistical analysis. The method gives a training procedure which has advantages of both averaging and clustering techniques. The proposed method is more reliable and robust than casual training. The word spoken by the user is measured and saved for the first time. When user speaks the word second time, DTW distance is computed between new pattern and previously stored pattern. If the distance comes out to be below a threshold, a reference pattern is created and training for that word is completed; otherwise third or subsequent passes are executed to save word reference pattern again. This procedure continues until all words are completed or until a maximum word repetition count is met. For testing the effectiveness of the training procedure, an experiment was performed taking nine talkers (five males, four females). Word reference template was created for 39-word vocabulary consisting of alphabets, the digits 0-9 and three command words. The experiment showed that for 95.2% of all words, a single reference pattern is obtained from the first four replications of that word by a given talker.

Lee and Hon presented a "Large Vocabulary Speaker Independent Speech Recognition System Using HMM" [15]. In their paper, they described about SPHINX which is a HMM based speaker independent large vocabulary recognizer. The system uses two types of HMM models: context-independent phone models and function-word-dependent phone models. Each word in SPHINX is represented by pronunciation network of phones and set of sentences accepted by grammar is represented by network of words. In order to add knowledge to HMM, three set of parameters viz. instantaneous LPC cepstrum coefficients, differenced LPC cepstrum coefficients and power and differenced power are computed. The speech is sampled at 16-khz and 12 LPC cepstral coefficients are computed which are then transformed to Mel-scale using bilinear transform and vector quantized into three codebooks which improves recognition accuracy and reduces VQ distortion. SPHINX is a phone-based HMM recognizer. A total of 153 HMM are created using a set of 105 HMM to model phones in 42 selected function words. The 153 HMM are trained through the use of a forward-backward algorithm which runs on 4160 sentence database. For recognition of speech, a time-synchronous Viterbi beam search technique is used. A threshold is determined and at a particular time, all states which are worse than the best state by more than the threshold are pruned. The system can recognize speech for no language model, a word pair language model and a bigram language model. The system has been tested for the 997 words and accuracy for bigram, word pair and no language model comes out to be 93%, 87.9% and 53.4%.

Kita et. al. proposed "HMM Continuous Speech Recognition Using Stochastic Language Models" [16]. Their system uses HMM-LR method which is an integration of Hidden Markov Models and LR parsing. First, the LR parser predicts the phoneme candidates and then these candidates are verified using HMM phoneme models. During the process of verification, all possible partial parses are constructed and the HMM verifier updates an array containing end point candidates and their probabilities. This partial parse is pruned whenever the highest probability in the array is lower than a threshold value. To improve recognition accuracy, word bigram/trigram model has been applied to Japanese syllables as they have special stochastic structure. The LR parser in the system is a stochastic shift reduces parser as it is closely related to stochastic context free grammar.

Suzuki et. al. proposed a speech recognition system based on acoustic models by considering variations in voice characteristics [17]. This system works by constructing voice-characteristic dependent acoustic model by using tree based clustering technique. The phonetic context is judged from linguistic phonetic knowledge using triphone models. To construct the voice-characteristic-dependent acoustic models, each speaker's voice is labeled according to the result of listening test. Since the context-dependent triphones can be very large, so these are grouped into number of clusters. So a tree based clustering technique is applied to speaker's voice characteristics. The simultaneous clustering of voice characteristics along with phonetic context allows the construction of voice-characteristic-dependent acoustic models. For recognition of speech, each leaf node having same phonetic context but different voice characteristics is integrated as a mixture distribution. Either the Yes or No node regarding phonetic context and both Yes and No nodes regarding voice characteristics are chosen and process is repeated for root to leaf nodes. At the end we get set of leaf that differs only in voice characteristics. The system has been trained using 20000 sentences spoken by 130 speakers of each gender and tested using total of 100 sentences spoken by 23 speakers of each gender. For the evaluation, the speech data is down-sampled to 16 kHz and parameterized to 12 Mel-cepstral coefficients. Three states left to right HMMs were used to model 43 Japanese phonemes, 146 phonological context questions and 20 voice characteristic questions. Embedded training has been applied before and after integrating voice-characteristic-dependent acoustic model. The result shows that proposed method performs better than conventional 4-mixture model in case of males and in case of females the proposed method performed well than conventional 8-mixture models.

Revathi and Venkatramani developed Speaker Independent Continuous Speech and Isolated Digit Recognition using VQ and HMM [18] which is based on perceptual features of speech. The system uses combination of Vector Quantization and HMM for speech recognition. The perceptual features are extracted by first computing the power spectrum of windowed speech and the grouping is done to 21 critical bands in bark scale. In order to simulate power law of hearing, loudness equalization and cube root compression is performed. After performing IFFT and LP analysis, the LP coefficients are converted into cepstral coefficients. Speech recognition using VQ consists of extracting features from training and testing data and building VQ codebooks for all 0-9 digit and continuous speeches. The codebooks are generated from training data using K-means clustering algorithm. Further HMM models are developed with state transition probability, observation symbol probability distribution and initial probability distribution to optimize the likelihood of the training set observation vectors. For discrete HMM, models are initialized with 256 observation sequences and 8 states. Code books indices are used as input to train the models. Observation sequences from feature vectors of all test speeches are given to HMM models and probability density values are calculated. After that all probabilities are compared and speech is selected whose likelihood is the maximum. Average accuracy of the system for speaker independent isolated digit using VQ+HMM is 93% and for speaker independent continuous speech is 100%.

Dua et.al. has developed Punjabi Automatic Speech Recognition System using HTK based on Hidden Markov Model [19]. The GUI of the system has been developed using JAVA platform in Linux environment. The system architecture consists of four phase's viz. Training data preparation, Acoustic Analysis, Acoustic model generation and GUI based decoder. The first phase deals with the recording and labeling of speech signal. The system is trained using 115 distinct Punjabi words which are recorded using a unidirectional microphone. The data is sampled at 16 kHz. 8 speakers recorded the data and each word is spoken 3 times by each speaker. The 2nd phase is the feature extraction phase in which original recorded waveform is converted into series of acoustical vectors. The features are extracted using MFCC (Mel frequency cepstral coefficient) technique. For this signal is segmented in a series of frames, each having length between 20 to 40 ms. Each frame is multiplied by a windowing function and after that a vector of acoustical coefficients is extracted from each windowed frame. In acoustic model generation phase, comparisons are made to recognize unknown utterances. First HMM is initialized by generating some prototype for each word. For generating prototype, some topology is used which consists of 4 observation functions and two non emitting states. After that optimal values for HMM parameters are estimated using HRest tool. In order to recognize speech, the test signal is converted to series of acoustic vectors. This data along with HMM definition, Punjabi word dictionary, task network and

generated HMM list is given to HTK tool HVite which compares it against recognizer's markov models and recognized word is displayed in text form. The performance of the system is tested in different environments using total of 6 distinct speakers each uttering 35-50 words. The average performance comes out to lie in the range of 94 to 96%.

Kumar et.al. Proposed a system named Continuous Hindi Speech Recognition using Gaussian Mixture HMM [20]. In this, the performance of the system is compared against different number of Gaussian mixture. The aim is to find the optimal number of Gaussian mixture that exhibits maximum accuracy. System uses the database of 51 words which are recorded at the sampling rate of 16 kHz. Features are extracted through MFCC technique. 39 MFCC are used in the experiment. In HMM training of continuous Hindi speech recognition system 5 states left right with no skips is used as a prototype model and 40 prototypes HMM model for all Hindi mono-phones are created. The mono-phone model further extends to triphone model to increase recognition accuracy. Different types of experiments have been conducted in order to test the performance of the system. Experiments with different vocabulary size showed that system has higher performance with small vocabulary. Experiments were performed five times with different number of Gaussian mixture. Tri-phone based continuous speech recognition system reported high accuracy with 4 mixtures GMM. Another experiment showed that tri-phone based system which is a context dependent system has better performance than mono-phone based system which is context independent. The authors are able to achieve 97.04% accuracy with 51 word vocabulary size.

Baby et.al. have proposed the enhancement of automatic speech recognition system based on deep neural network using exemplar based technique [21]. The system used coupled dictionaries as a pre-processing stage. The noisy speech is first decomposed as a weighted sum of atoms in an input dictionary having exemplars sampled from a domain of choice. In order to directly obtain the estimations of speech and noise, the resulted weights are applied to a coupled output dictionary having exemplar sampled in short time Fourier transform (STFT). The system has been evaluated using three different input exemplar spaces namely Mel, magnitude STFT and MS spaces. Three types of settings have been used as DFT-DFT setting, Mel-Mel and Mel-DFT settings and MS-DFT settings. In DFT-DFT setting DFT exemplar space is chosen as the input exemplar. In order to create the input dictionary using DFT exemplars, a random segment of acoustic data spanning T frames is taken and its full resolution magnitude STFT of size F*T is considered. In Mel-Mel and Mel-DFT, NMF based decomposition is done using Mel dictionary having Mel exemplars. In order to obtain Mel exemplar, magnitude STFT of size F*T is pre-multiplied with STFT-to-Mel matrix. MS-DFT setting makes use of MS exemplars to obtain compositional model using NMF. MS-exemplars are obtained by considering T frames of acoustic data and filtered using a filter bank having B channels. The resulting B band-limited signals are half-wave rectified to model non negative nerve firings and low pass filtered at a 3 db cut-off frequency. The system has been trained and tested using AURORA-4 database with both clean and multi-condition training. Average word error rates are used to evaluate and compare the performance of various settings. The system yielded average overall WERs of 26.8% and 11.9% with clean and retrained DNN respectively.

Nguyen et. al. have improved the English ASR system using two approaches of Deep Neural Network Hybrid and bottleneck [22] features based on de-noising en-coders. Deep Neural Network architecture for Hybrid HMM/GMM consists of large number of fully connected hidden layers followed by final classification layer. Architecture for bottleneck feature extraction is similar to hybrid HMM/GMM but it has a small bottleneck layer. For training the acoustic model, the authors have used TED talk lectures consisting of 22 hours of audio distributed among 920 talks. The non spoken sounds have been filtered out using segmentation and the remaining audio used for training was around 175 hours of speech. One eighth of the Giga corpus filtered according to the Moore-Lewis approach has been used for language modeling. During supervised training, the neural network predicts context dependent HMM states. The auto-encoders are pre-trained using gradient descent method with learning rate of 0.01%. The input vectors are corrupted by masking noise. Bottleneck consisting 39 units is then added to the remaining layers. The authors have evaluated the system using 2012 development set and 2013 test set. The word error rate in baseline system is 30% on dev2012 and 36.1% on test2013. The hybrid DNN/HMM combination outperforms the baseline setup showing the error rates of 18.7% and 22.7%.

Lee et. al. proposed a multistage enhancement technique for Automatic Speech Recognition [23] In first stage the multi-channel speech enhancement method works on spatial information of speech signal for improvements. The second stage enhances the performance of the system at server side by employing a data driven approach based single channel speech enhancement method. The single channel speech enhancement method uses a priori and posteriori SNR to train the noise reduction gain function. The performance of the proposed method is evaluated by recording the 1200 spoken sentences by 12 Korean persons (5 females, 7 males). The speech samples are recorded in various noisy environments such as car, street, café etc. The word recognition rate in proposed method is 77.9% which is higher than conventional method (65.7%).

Mohan and Babu have implemented a speech recognition system in MATLAB environment using MFCC (Mel-Frequency Cepstral Coefficients) and DTW (Dynamic Time Warping) [24]. The system employs two phases: Feature Extraction and Feature Matching. Before extracting features using MFCC, the voice signal is converted from Analog to Digital by following Pre-Emphasis and filtering. For Pre-emphasis FIR filter is used which increase the higher frequency magnitude with respect to lower frequencies. Voice sample is framed within ranges of 20 to 30ms. Each frame is then multiplied by a hamming window. After that Fast Fourier Transform is taken for each frame to transform the signal into frequency domain. Each resultant frame is then multiplied by Triangular MEL filter bank. The resulting values are called MFCC. After feature extraction, DTW algorithm is used for feature matching by calculating least distance between features of spoken word and reference templates. Among the calculated scores, the reference template with least value is selected as detected word.

Kumar Ravinder has purposed a speech recognition system for isolated word recognizer for Punjabi language [25]. Purposed system is a speaker dependent system and it works in real time. System is developed using the Hidden Markov Model (HMM) and Dynamic Time Warp (DTW) techniques for small vocabulary of isolated spoken words in Indian regional language (Punjabi). After development of systems, research is further extended to comparison of those developed systems. The purposed work has focused on template-based recognizer approach using linear predictive coding (LPC) for feature extraction with dynamic programming computation and vector quantization with Hidden Markov Model based recognizers. In purposed systems, end point detection is done by finding energy in speech signal wave. End point detection is done to extract background silence before and after the input speech. The systems are trained for Punjabi numerals five times. The performance of system for Punjabi language numerals with DTW is 92.3% and with HMM is 87.5%. The performance of the systems is showing that DTW approach is more appropriate for Punjabi numerals and isolated spoken words.

V. CONCLUSION

In this paper work, we have presented a survey on challenges, different approaches and evaluation metrics used for different speech recognition systems. We have also listed some of the existing speech recognition systems. From the survey we have found that almost all existing language speech recognition systems are based on HMM and pattern based approaches. We have tried to list down the works of few different scholars and institutions but there might exist some more groups and organizations that are involved in the development of speech recognition systems.

REFERENCES

[1] De Wachter, M., et al., *Template-based continuous speech recognition.* IEEE Transactions on Audio, Speech, and Language Processing, 2007. **15**(4): p. 1377-1390.

[2] Resch, B., *Automatic Speech Recognition with HTK.* Signal Processing and Speech Communication Laboratory. Inffeldgase. Austria. Disponible en Internet: http://www. igi. tugraz. at/lehre/CI, 2003.

[3] Kekre, H., A.A. Athawale, and G. Sharma. *Speech recognition using vector quantization*. in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*. 2011. ACM.

[4] Moore, R.K., *Twenty things we still don't know about speech*, in *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology*. 1994.

[5] Tebelskis, J., *Speech recognition using neural networks*. 1995, Siemens AG.

[6] Pan, Y., P. Shen, and L. Shen, *Speech emotion recognition using support vector machine.* International Journal of Smart Home, 2012. **6**(2): p. 101-108.

[7] Anusuya, M. and S.K. Katti, *Speech recognition by machine, a review.* arXiv preprint arXiv:1001.2267, 2010.

[8] Alhawiti, K.M., *Advances in Artificial Intelligence Using Speech Recognition.* World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2015. **9**(6): p. 1397-1400.

[9] Furui, S., *Theory and Applications.* 1 ed. Speech Technology, ed. K.J. Fang Chen. US: Springer. XXVII, 331.

[10] Furui, S., *50 years of progress in speech and speaker recognition.* SPECOM 2005, Patras, 2005: p. 1-9.

[11] Gulzar, T., et al., *A systematic analysis of automatic speech recognition: an overview.* Int. J. Curr. Eng. Technol, 2014. **4**(3): p. 1664-1675.

[12] Furui, S., *Cepstral analysis technique for automatic speaker verification.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981. **29**(2): p. 254-272.

[13] Sambur, M.R. and L.R. Rabiner, *Statistical decision approach to the recognition of connected digits.* The Journal of the Acoustical Society of America, 1976. **60**(S1): p. S12-S12.

[14] Rabiner, L. and J. Wilpon, *A simplified, robust training procedure for speaker trained, isolated word recognition systems.* The Journal of the Acoustical Society of America, 1980. **68**(5): p. 1271-1276.

[15] Lee, K.-F. and H.-W. Hon. *Large-vocabulary speaker-independent continuous speech recognition using HMM*. in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. 1988. IEEE.

[16] Kita, K., T. Kawabaa, and T. Hanazawa. *HMM continuous speech recognition using stochastic language models*. in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. 1990. IEEE.

[17] Suzuki, H., et al. *Speech recognition using voice-characteristic-dependent acoustic models.* in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. 2003. IEEE.

[18] Revathi, A. and Y. Venkataramani. *Speaker independent continuous speech and isolated digit recognition using VQ and HMM*. in *Communications and Signal Processing (ICCSP), 2011 International Conference on*. 2011. IEEE.

[19] Dua, M., et al., *Punjabi automatic speech recognition using HTK.* IJCSI International Journal of Computer Science Issues, 2012. **9**(4): p. 1694-0814.

[20] Kuamr, A., M. Dua, and T. Choudhary. *Continuous hindi speech recognition using gaussian mixture HMM*. in *Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on*. 2014. IEEE.

[21] Baby, D., J.F. Gemmeke, and T. Virtanen. *Exemplar-based speech enhancement for deep neural network based automatic speech recognition.* in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015. IEEE.

[22] Nguyen, Q.B., T.T. Vu, and C.M. Luong. *Improving acoustic model for English ASR System using deep neural network.* in *Computing & Communication Technologies-Research, Innovation, and Vision for the Future (RIVF), 2015 IEEE RIVF International Conference on*. 2015. IEEE.

[23] Lee, S., Y. Lee, and N. Cho. *Multi-stage speech enhancement for automatic speech recognition*. in *2016 IEEE International Conference on Consumer Electronics (ICCE)*. 2016. IEEE.

[24] Mohan and B. Jagan. *Speech recognition using MFCC and DTW.* in *Advances in Electrical Engineering (ICAEE), 2014 International Conference on*. 2014. IEEE.

[25] Ravinder, K. (2010, November). *Comparison of hmm and dtw for isolated word recognition system of punjabi language.* In *Iberoamerican Congress on Pattern Recognition (pp. 244-252).* Springer Berlin Heidelberg.