Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi a Case Study

Wiqas Ghai, Navdeep Singh

Abstract*— Punjabi, Hindi, Marathi, Gujarati, Sindhi, Bengali, Nepali, Sinhala, Oriya, Assamese, Urdu are prominent members of the family of Indo-Aryan languages. These languages are mainly spoken in India, Pakistan, Bangladesh, Nepal, Sri Lanka and Maldive Islands. All these languages contain huge diversity of phonetic content. In the last two decades, few researchers have worked for the development of Automatic Speech Recognition Systems for most of these languages in such a way that development of this technology can reach at par with the research work which has been done and is being done for the different languages in the rest of the world. Punjabi is the 10th most widely spoken language in the world for which no considerable work has been done in this area of automatic speech recognition. Being a member of Indo-Aryan languages family and a language rich in literature, Punjabi language deserves attention in this highly growing field of Automatic speech recognition. In this paper, the efforts made by various researchers to develop automatic speech recognition systems for most of the Indo-Aryan languages, have been analysed and then their applicability to Punjabi language has been discussed so that a concrete work can be initiated for Punjabi language.

***Index Terms* — Maximum likelihood linear regression, Learning vector quantization, Multi layer perceptron, Cooperative heterogeneous artificial neural network.**

I. INTRODUCTION

Automatic Speech Recognition provides a vehicle for natural communication between man and machine. Automatic speech recognition has become an interesting and challenging area of research and development for the researchers across the world. It helps to provide a capability to a machine for responding properly to spoken language. Punjabi, Hindi, Marathi, Oriya, Gujarati, Sindhi, Bengali, Nepali, Sinhala, Assamese and Urdu are Indo-Aryan languages. Indo-Aryan languages are mostly phonetic in nature. For a phonetic language, there always exits a one to one mapping between their pronunciation and orthography. In addition to this, unlike English and other European languages, these languages possess a large number of phonemes such as retroflex and aspirated stops. Indo-Aryan languages, for which research on automatic speech recognition is being carried out, are Hindi, Marathi, Bengali, Sinhala, Urdu, Oriya, Assamese and Punjabi. Punjabi being a language with large number of speakers in the world and

Revised Manuscript Received on March 2012.

Wiqas Ghai, Assistant Professor, Department of Computer Science, Khalsa College (ASR) of Technology & Business Studies, Mohali. #01762-228457, (ghaialpha@gmail.com).

Mr. Navdeep Singh, Senior Lecturer, Post Graduate Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, (navdeep_jaggi@yahoo.com).

II. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the process of mapping an acoustic waveform into a text/the set of words which should be equivalent to the information being conveyed by the spoken words. This challenging field of research has almost made it possible to provide a PC which can perform as a stenographer, teach the students in their mother language and read the newspaper of reader's choice. The advent and development of ASR in the last 6 decades has resolved the issues of the requirements of certain level of literacy, typing skill, some level of proficiency in English, reading the monitor by blind or partially blind people, use of computer by physically challenged people and good hand-eye co-ordination for using mouse. In addition to this support, ASR application areas are increasing in number day by day. Research in Automatic Speech Recognition has various open issues such as Small/ Medium/ Large vocabulary, Isolated/ Connected/Continuous speech, Speaker Dependent/ Independent and Environmental robustness.

A. Modules of ASR

Automatic speech recognition system is comprised of modules as shown in the figure 1.

1.Speech Signal acquisition: At this stage, Analog speech signal is acquired through a high quality, noiseless, unidirectional microphone in .wav format and converted to digital speech signal.

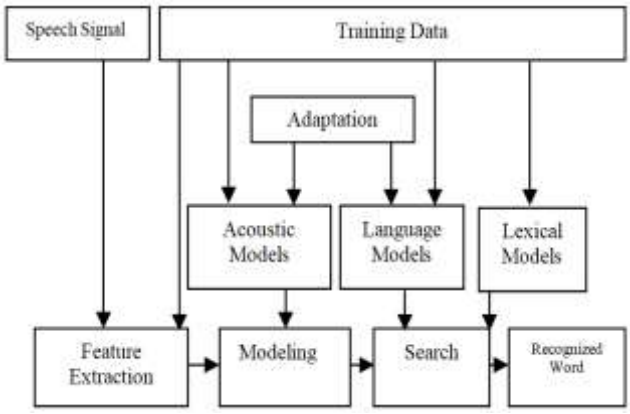

Figure 1: Block Diagram of ASR System

2. *Feature Extraction*: Feature extraction is a very important phase of ASR development during which a parsimonious sequence of feature vectors is computed so as to provide a compact representation of the given input signal. Speech analysis of the speech signal acts as first stage of Feature extraction process where raw features describing the envelope of power spectrum are generated. An extended feature vector composed of static and dynamic features is compiled in the second stage. Finally this feature vector is transformed into more compact and robust vector. Feature extraction, using MFCC, is the famous technique used for feature extraction (Figure 2).

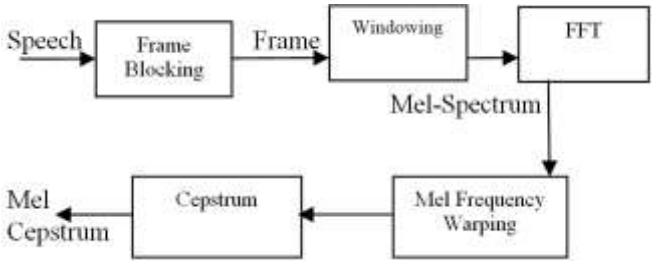

Figure 2: Block Diagram of Feature Extraction

3. *Acoustic Modelling*: Acoustic models are developed to link the observed features of the speech signals with the expected phonetics of the hypothesis word/sentence. For generating mapping between the basic speech units such as phones, tri-phones & syllables, a rigorous training is carried. During training, a pattern representative for the features of a class using one or more patterns corresponding to speech sounds of the same class.

4. *Language & Lexical Modelling*: Word ambiguity is an aspect which has to be handled carefully and acoustic model alone can't handle it. For continuous speech, word boundaries are major issue. Language model is used to resolve both these issues. Generally ASR systems use the stochastic language models. These probabilities are to be trained from a corpus. Language accepts the various competitive hypotheses of words from the acoustic models and thereby generates a probability for each sequence of words. Lexical model provides the pronunciation of the words in the specified language and contains the mapping between words and phones. Generally a canonical pronunciation available in ordinary dictionaries is used. To handle the issue of variability, multiple pronunciation variants for each word are covered in the lexicon but with care. A G2P system- Grapheme to Phoneme system is applied to better the performance the ASR system by predicting the pronunciation of words which are not found in the training data.

5. *Model Adaptation*: The purpose of performing adaptation is to minimise the system's performance dependence on speaker's voice, microphones, transmission channel and acoustic environment so that the generalization capability of the system can be enhanced. Language model adaptation is focussed at how to select the model for specific domain. Adaptation process identifies the nature of domain and, thereby, selects the specified model.

6. *Recognition*: Recognition is a process where an unknown test pattern is compared with each sound class reference pattern and, thereby, a measure of similarity is computed. Two approaches are being used to match the patterns: First one is the Dynamic Time Warping based on the distance between the acoustic units and that of recognition. Second one is HMM based on the maximisation of the occurrence probability between training and recognition units. To train the HMM and thereby to achieve good performance, a large, phonetically rich and balanced database is needed.

B. Data Preparation

1. Building Text Corpus: Text corpus means optimal set of textual words/sentences which will be recorded by the native speakers of a particular language. According to the domain specified for the ASR, the corresponding text is collected. Different context in which that text can be used, are also taken care of. Building a text corpus involves three steps: Text corpus collection, Grapheme to Phoneme Conversion, Optimal Text Selection.

2. Building Speech Corpus: With the help of Text Corpus, the recordings of selected words/sentences are done with the help of high quality microphones. During the development of speech corpus, information, which is generally noted down, is Personal Profile of Speakers, Technical details of microphone, Date and Time of Recording, Environmental conditions of recording. During the recording session, the parameters of the wave file to be set are: Sampling rate, Bit rate, Channel. Building Speech Corpus involves three steps: Selecting a speaker, Data Statistics, Transcription Correction.

3. Transcription File: A transcript file is required to represent what the speakers are saying in the audio file. It contains the dialogue of the speaker noted exactly in the same precise way as it has been recorded. There are two transcription files: one is meant for training the system and second one is meant for testing the system.

4. Pronunciation Dictionary: It is a language dictionary which contains mapping of each word to a sequence of sound units. The purpose of this file is to derive the sequence of sound units associated with each signal. The important point, which is to be taken care of while preparing this dictionary, the sound units must be contained in this dictionary, must be in ASCII.

5. Language Model: Language model is meant for providing the behaviour of the language. The language model describes the likelihood or the probability taken when a sequence or collection of words is seen. A language model is a probability distribution over the entire sentences/texts. The purpose of creating a language model is to narrow down the search space, constrain search and thereby to significantly improve recognition accuracy. Language model becomes very important when Continuous speech is considered. Speech recognizers seek the word sequence W_s which is most likely to be produced from acoustic evidence A as per the following formula:

$$P(W_s \mid A) = \max_w P(W|A) = \max_w P(A|W)\,P(W)/P(A)$$

Where P (A): Probability of acoustic evidence. Language Model assigns a probability estimate P (W) to word sequences W= {W_1, W_2,....., W_n}. These probabilities can be trained from a corpus. Perplexity is a parameter to evaluate language model. Suppose sentences in a test sample contains 2000 words and can be coded using 10000 bits then the perplexity of language model = $2^{(10000/2000)}$=32 per word. Language model with low perplexity helps LM to perform well for the speech recognition system thereby compressing the test sample.

6. Filler Dictionary: It refers to a dictionary which contains the mapping of non-speech sounds to non-speech sound units. e.g. <sil> SIL

C. Performance Parameters

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which are described below:

1. Accuracy Parameters

Word Error Rate (WER): The WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set.

Formula: $WER = \dfrac{S + D + I}{N}$

Word Recognition Rate (WRR): It is another parameter for determining accuracy.

Formula: WRR=1-WER

e. g. REF: Misunderstandings | usually | develop.

S1: Misunderstandings | using | develop.

The substitution error of the word *using* for the word *usually* would be scored as one substitution error, as opposed to one error of deletion (*usually*) and one error of insertion (*using*). *Single Word Error Rate* (SWER) and *Command Success Rate* (CSR) are two more parameters to determine accuracy of a speech recognition system.

2. Speed Parameter

Real Time Factor is parameter to evaluate speed of automatic speech recognition.

Formula: $RTF = \dfrac{P}{I}$

where P: Time taken to process an input

Duration of input I

e. g. RTF= 3 when it takes 6 hours of computation time to process a recording of duration 2 hours.

RTF ≤ 1 implies real time processing.

D. Performance Degradation

Automatic speech recognition suffers degradation in recognition performance due to following inevitable factors:

i. Prosodic and phonetic context
ii. Speaking behaviour
iii. Accent & Dialect
iv. Transducer variability and distortions
v. Adverse speaking conditions
vi. Pronunciation
vii. Transmission channel variability and distortions
viii. Noisy acoustic environment
ix. Vocabulary Size and domain

III. INDO_ARYAN ASRs

A. Assamese Speech Recognition System

Assamese, mainly spoken in Assam & Arunachal Pradesh, has official language of Assam state. It has derived its phonetic character set and behaviour from Sanskrit. It has 11 vowels and 41 consonants in its script. Assam is a semi-literate state and as a result, research in the area of Automatic speech recognition system and thereby the development of ASRs for public domain will surely help the people of this state. Total absence of any retroflex sounds, Extensive use of velar nasal /ŋ/ and Presence of voiceless velar fricative are the unique features of Assamese language. Sarma & Sarma [1] worked for the development of numeral speech recognition system for Assamese language. Gender and mood variations were given consideration during the recording speech signals of 10 numeral digits at 8 KHz in mono channel mode. For feature extraction, a hybrid feature set of Linear Predictive Code (LPC) and Principal Component Analysis (PCA) was extracted so as to identify and differentiate speech sounds, insensitive noise and other irrelevant factors. Learning Vector Quantization (LVQ) blocks in cooperative architecture, comprised of SOM – Self Organised Map and MLP – Multi Layer Perceptron, were used as classifiers. The purpose of this architecture was to minimise classification error, maximize prediction rate and tackle Assamese numeral speech input with mood, gender and recording condition variation. Success rate, for simple LVQ system and as well as for cooperative LVQ system, has been found to more than 95%. Training time with cooperative LVQ system has been found to be 1.5 times more than that with unitary LVQ system.

Sarma et al. [2] have proposed the design of an optimal feature extraction block and ANN based cooperative architecture CHANN. MLP acted as class mapper network. LPC coefficient was transformed to robust set of parameters i.e. Cepstral coefficient 20 because actual LPC show high variance. Since phoneme discriminating power of MLP deteriorates for consonants, so Recurrent Neural Network with one hidden layer and tan-sigmoid activation function was taken as classifier. Training of RNN was done with BPM – back propagation with momentum. To improve the training and testing time performance of RNN based ASR, cooperative heterogeneous ANN architecture was adopted. Several sets of noise corrupted test signals were created to determine the performance of filter structures and thereby making the ASR compatible with noisy environment. Taking noise free, noise mixed, stress free and stressed samples with gender specific recognition, performance of CHANN block was better than singular RNN block.

B. Bengali Speech Recognition System

Bengali is native to the region of eastern South Asia known as Bengal, comprising Bangladesh, the Indian state of West Bengal and parts of the Indian states of Tripura and Assam. In India, Bengali is the second most commonly spoken language. The important point about this language is that National Anthem of India has been written and composed in Bengali Language and the National animal of India is also named as The Royal Bengal Tiger. For Bangladesh, Bengali/Bangla is national and official language. Phonemic inventory of Bengali contains 29 consonants and 14 vowels including 7 nasalised vowels. Chowdhury [3] has conducted his research centred on the methodology to develop Domain based Continuous Speech Recognition system for Bangla language by using CMU-Sphinx tool. The system was tested by varying speaker, environment and microphone. Audio inputs from two speakers were used for testing. Decoder Pocket Sphinx gave the recognition accuracy 90.65%. Decoder Sphinx 4 gave average recognition accuracy of 71.38% by varying speakers where as 86.79% by varying the microphone and environment. It was also observed that due to considerable period of silence in the middle of speech, the performance of decoder declines.

Hasnat et al. [4] worked for the development of Isolated speech recognizer as well as Continuous speech Recognizer by using HTK tool. Adaptive filer was used to eliminate noise from the recorded speech signal. End point detection algorithm was used for start and end point detection for continuous speech where as for isolated speech, the additional purpose was to eliminate noisy signal and unwanted signal within the speech. For feature extraction, MFCCs were used. For training the system, word based HMM model was used for isolated speech recognition where as phoneme based HMM model for continuous speech recognition. Speaker dependent isolated word recognition system's accuracy was found to be 90% where as it was 80% for SD continuous speech recognition system. There was 20% decline in performance for speaker independent systems.

C. Hindi Speech Recognition System

Hindi is used as the sole official state language in the states of UP, Bihar, Rajasthan, Himachal Pradesh, MP, Chattisgarh, Jharkhand etc. Hindi contains more vocabulary from Sanskrit. Hindi was first used in writing during the 4th century AD with Brahmi script. 11th century AD onwards, it has been written with the Devanāgarī script. Hindi language possesses 40 consonants, 10 vowels and 2 modifiers. Samudravijaya [5] came up with the development of a speaker independent, continuous speech recognition system for Hindi. Their system recognized spoken queries in Hindi in the context of railway reservation enquiry task. The sentence corpus contained 320 sentences and the size of the vocabulary was 161 words. Two microphones were applied for database collection. A high quality & directional SM 48 microphone was used for recording sentences. An ordinary microphone was mounted at desktop mount to contain the effect of room acoustics and background noise. A spoken sentence is represented as a sequence of 48 context independent acoustic-phonetic units, each modelled by a hidden Markov model. The performance of the system has been reported for test as well as training data collected from 10 speakers, both male and female.

Kumar and Aggarwal [6] have developed an Automatic speech recognition system for Hindi language for recognizing isolated words. HMM was used to train and recognize the speech. 39 MFCCs (12 Mel Cepstrum + Log energy + 1st and 2nd Order derivatives) features were extracted. HMM tool kit was used. A data set of only 30 words was developed and 960 speech files were created with the help of 5 male and 3 female speakers. Velthuis transliteration was used for creating transcription file. Word accuracy of 94.63% and word error rate of 5.37% were achieved.

It has been found that Gaussian evaluation of acoustic signal is a computationally expensive task. A range of 8 to 64 mixture components per state have been found to be useful depending upon the amount of training data. Aggarwal and Dave [7] have proposed an approach to speed up the statistical pattern classification by reducing the time consumed in the likelihood evaluation of feature vectors with the help of optimal number of Gaussian mixture components. They applied extended MFCC procedure by extracting 52 MFCC features (39 MFCC + 13 triple delta features) and then reducing them to 39 by using HLDA – Hetroscedastic linear discriminant analysis technique. They performed experiments to analyse the results with due regard to effect of change in the number of Gaussians, size of vocabulary, training method, modelling unit on the recognition accuracy. The observations made from the experiments carried out are:

i. It is not preferable to use whole word model beyond 200 vocabulary size.
ii. Extended MFCC gave improvement over standard MFCC.
iii. Improvement in recognition accuracy up to 4 mixture components
iv. Discriminative MPE is better than MLE training method
v. Tri-phone model performs better than other modelling units

Sivaraman and Samudravijaya [8] have made an attempt to compensate the mismatch between training and testing conditions with the help of unsupervised Speaker adaptation. MLLR - Maximum Likelihood Linear Regression, a speaker adaptation technique requiring small amount of data, has been used. A global MLLR transform was used due to scarcity of adaptation data. Tri-phones, modelled by left-to-right 5-state semi-continuous HMMs, have been used as acoustic units. Back-off trigram grammar has been used as Language model. A multi-speaker Hindi continuous speech database "rlwRes", related to railway reservation availability, was used for testing the efficacy of online speaker adaptation. The database used for training acoustic models was different from "rlwRes" database. It was observed that online speaker

adaptation led to mild deterioration in ASR accuracy when the accuracy of complex, general purpose SI system was less than 70%.

D. Marathi Speech Recognition System

Marathi language is mainly spoken in western and central India. Standard Marathi and Warhadi Marathi are the two dialects of this language. Standard Marathi is the official language of Maharashtra state. Its vocabulary possesses 36 consonants and 16 vowels. Gawali et al. [9] worked to develop a speech database and speech recognition system for isolated words. CSL: Computer Speech Laboratory has been used for collecting speech data i.e. vowels, isolated words starting from each vowel and simple Marathi sentences. One experiment for speech recognition was conducted using MFCC and another experiment was conducted using DTW. The results showed that recognition accuracy of 94.65% was obtained using MFCC where as 73.25% was obtained using DTW.

Broadening the scope of Marathi speech recognition to the benefit for society, Gaikwad et al. [10] have developed a Polly clinic inquiry system using IVR, which can respond to wide range of health care services. It is a commendable effort which can help the patients to interact for their health related queries. Three sets of database have been created. First one of 50 isolated words corresponds to doctor name, second of 1920 sentences corresponds to queries and the last one of 60 samples corresponds to call back/solution. CSL – Computerised Speech Lab has been used for speech data collection. Size of vocabulary for training is 1000 sentences. Time for training has been 4' 55" where for testing and relevant answer, time has been 1' for 670 continuous sentences. Overall accuracy has been found to be around 88%.

E. Oriya Speech Recognition System

Oriya, mainly spoken in Orissa & West Bengal, is an official language of Orissa and 2nd official language of Jharkhand. It has number of dialects such as Midnapori, Singhbhumi, Baleswari, Ganjami etc. Mughalbandi Oriya is considered as proper or Standard Oriya due to literary traditions. Oriya has 28 consonants and 6 vowel phonemes. Unlike Hindi, Oriya has retained most of the cases of Sanskrit. Oriya Automatic recognition of continuously spoken digits is one area where speech recognition can serve the speakers of any language and thereby help them to exploit the benefits of modern technology. Mohanty and Swain [11] have made such effort for Oriya language. For this experiment, they tried Bakis model of HMM which allows the states to transit to themselves or to successive states but side by side, it restricts transition to earlier states. The likelihood probability was calculated using Gaussian Mixture model and prior probability was determined with N-gram grammar. Speech corpus to be used for training and testing in the ratio 3:1, was collected at sampling frequency 16 KHz, Bit rate 16 bits, mono channel and contained 2000 sentences and around 8000 words which were obtained from 50 speakers = 35 female +15 male. Attributes like gender, age and dialect were captures to great extent. Feature extraction was applied using MFCC. Trigram Language model was used because there is a strong dependence of most of the words on previous words for Oriya language. Word accuracy obtained was 94.72% for seen data and 78.23% for unseen data where as sentence accuracy obtained was lesser than these %ages for both types of data. Oriya is Eastern Indo-Aryan Language.

Mohanty and Swain [12] have come forward to apply the benefit of automatic speech recognition systems to society by developing an isolated speech recognizer for Oriya language so that visually impaired students can attempt the closed ended questions such as fill-in-blanks, dichotomous, ranking scale, multiple choice and rating scale questions, well during their exams. Oriya isolated words were recorded using high quality directional microphone in laboratory environment. A desktop microphone was placed to contain noticeable noise. MFCCs were extracted at sampling rate of 16 KHz with 16 bit quantization. HMM was used for pattern recognition. Viterbi decoding was used as decoding engine. Performance was measured by computing recognition accuracy at word level. Word accuracy for seen data and unseen data was found to be 76.23% and 58.86% respectively.

F. Sinhala Speech Recognition System

Sinhala is mother tongue of Sinhalese people and as well as national language of Sri-Lanka. Complete alphabet of Sinhala has 18 vowels and 36 consonants. For writing colloquial spoken Sinhala, only 12 vowels and 24 consonants are used. In contrast to spoken Sinhala, grammar for written Sinhala depends on number, person, gender and tense. Nadungodage & Weerasinghe [13] developed a continuous speech recognizer using written Sinhala vocabulary only. Sinhala sentences were recorded with the help of Praat at sample frequency 16 KHz using a mono channel. A bi-gram language model was created. HTK was used for developing Sinhala continuous speech recognizer. System was trained from only a single female. Sentence recognition accuracy obtained was 75% and word recognition accuracy obtained was 96%. It was observed that most of the incorrectly identified utterances differed from the correct utterances only by 1 or 2 syllables.

G. Urdu Speech Recognition System

Urdu, being national language of Pakistan, is one of 22 scheduled languages in India and official language of 5 states. Urdu is mutually intelligible with standard Hindi spoken in India. Urdu has recognised dialects such as Dakhni, Rekhta and Modern Vernacular Urdu. It has 28 consonants and 10 vowels. Raza et al. [14] worked for the development of a HMM based Large Vocabulary Automatic Spontaneous Urdu speech recognition system with the help of Sphinx 3 trainer and decoder. The training data used for this work contained a phonetically rich sentence based corpus read out by native Urdu speakers and Spontaneous conversational data from recorded interviews of native speakers.

A normal home and office environment, where ambient noise exists, was utilized for carrying out recordings. A tri-gram language model was derived from actual training data. After studying in detail all the existing options for phonetic transcription such as IPA, SAMPA, X-SAMPA, ARPABET etc., case insensitive SAMPA was applied because a case insensitive notation free from special characters was required. A mapping mechanism was developed to map Unicode to ASCII, because Urdu script uses Unicode characters. It was observed from the experiments that WER was least for 1:1 ratio of read speech & spontaneous speech data and maximum for read speech based training data.

Sarfraz et al. [15] have worked on the development of LVCSR for Urdu language using CMU Sphinx Open Source Toolkit. They used a corpus of training data recorded in noisy environment so as to improve the robustness of speech recognition. Speech data was recorded at 16 KHz. Tri-phone Language model corresponding to each training data set was created using SLM toolkit. Trigram models accompanied with Witten Bell discounting. Refining process was carried out to identify diacritical marks in words so that proper diacritical marks can be inserted and phonemic transcription was created. Fine tuning of the transcriptions was also included in refining process to match the pronunciations. Increase in training data, for each speaker, gave a significant improvement was found in test results in case of male speakers. For multi speakers data, improvement was found in WERs for the data sets which were retrained and tested after the review and refinement of transcription data.

H. Punjabi Speech Recognition System

1. Introduction: Punjabi is 10th most widely spoken language in the world which got emerged as an independent language in 11th century. There are around 107 million native speakers of the Punjabi language with around 77 million in Pakistan and 30 million in India. Punjabi with Gurumukhi script is one of the 22 languages with official status in India and more specifically first official language in Indian Punjab in comparison to no official status for Punjabi language in Pakistan Punjab. Spoken Punjabi in India relies more heavily on Sanskrit vocabulary through Hindi and has having many dialects such as Majhi, Potwari, Dhani, Hindko, Malwi etc. Majhi is a prestige dialect of Punjabi language.

2. Features: Punjabi is the only tonal language in Indo-Aryan languages family with its phoneme inventory containing 10 vowels, 25 consonants, 7 diphthongs and three tones whose production has neither friction nor stoppage of air in the mouth. Patterns of pitch variation are employed in a tonal language to distinguish between different meanings of word which have the same pattern of consonants and vowels. Tone features are segmental and phonemic in function. Punjabi is also considered as fusion language because it has a capability of fusing morphemes. Punjabi words are usually ordered as head final SOV - Subject Object Verb. Punjabi verbs inflect for tense, aspect (perfective, imperfective), mood (indicative, imperative, subjunctive, conditional), voice (active, passive), person, number, and gender (male, female).

3. Work: Kumar [16] developed a speaker dependent, real time, an Isolated and connected word recognition system for Punjabi language using acoustic template matching technique. It was designed for medium sized dictionary. Vector quantization and Dynamic warping techniques were used with some modification to noise and word detection algorithms. The recordings were done at a Sampling frequency of 16 KHz, 8 bit Sample size and Mono channel. LPC analysis was used for feature extraction. LPC analysis included computation of Autocorrelation Coefficients (with 10 LPC coefficients), LPC Coefficients (using 10 LPC coefficients), Cepstral Coefficients (10 Cepstral coefficients) and Delta Cepstrum (using 5 frame windows). DTW-Dynamic Time Warping has been used for recognition. In training mode, 1500 isolated words of a single Punjabi speaker were taken as a knowledge base acting as reference template. The performance was evaluated with 500 isolated Punjabi words. Recognition accuracy has been found to be 61% for isolated words and lesser for connected words recognition. More precisely, vowels of Punjabi language have found to be more accurate than consonants. Kumar [17] also compared the performance of DTW based speech recognition and HMM based speech recognition for Punjabi isolated words. Performance was in favour of DTW based recognizer but that is just because of insufficiency of HMM training data. Time and space complexity has been found to be less in HMM based recognizer for same size of code book.

4. Prospects: Being a widely spoken language, Punjabi language has number of speakers in the world. The dimensions of automatic speech recognition, which have been discussed in this paper, and the techniques, which have been applied for other languages in its Indo-Aryan language family, can be helpful in providing a great boost to the field of automatic speech recognition for this language. Use of advanced speech recognition tools such as HTK and SPHINX, advanced feature extraction techniques such as MFCC and Extended MFCC and use of ANN can support the Punjabi ASRs research and development.

IV. CONCLUSION

Lot of research in the field of Automatic speech recognition is being carried out for Hindi, Oriya, Malayalam, Bengali, Assamese, Marathi, Urdu and Sinhala languages of Indo-Aryan languages family. But there is a long way to go so as to enhance the performance standards set for other languages. It has been observed that use of techniques Cooperative Heterogeneous ANN Architecture, Maximum Likelihood Linear Regression, Extended MFCC and Learning Vector quantization are helping the researchers to

get improved recognition performance of speech recognition systems. Computerised Speech Lab has also helped in speech acquisition process. Punjabi, being a widely spoken Indo-Aryan language, is still trailing in the research and development for the field of automatic speech recognition. So far the work done for Punjabi language is Isolated word speech recognition using Acoustic template matching technique on MATLAB. In this paper, almost all the efforts made by various researchers for the research and development of ASR for Indo-Aryan languages have been analyzed and the applicability of techniques applied for other Indo-Aryan languages has been discussed for Punjabi language.

REFERENCES

1. Sarma, M. P.; Sarma, K. K., "Assamese Numeral Speech Recognition using Multiple Features and Cooperative LVQ – Architectures", International Journal of Electrical and Electronics 5:1, 2011.
2. Sarma, M.; Dutta, K.; Sarma, K. K., "Assamese Numeral Corpus for Speech Recognition using Cooperative ANN Architecture", International Journal of Electrical and Electronics Engineering 3:8 2009.
3. Chowdhury, S. A., "Implementation of Speech Recognition System for Bangla", BRAC University, DHAKA, Bangladesh, August 2010.
4. Hasnat, M. A., Molwa, J.,Khan, M., "Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective",2007.
5. Samudravijaya, K., "Computer Recognition of Spoken Hindi". Proceeding of International Conference of Speech, Music and Allied Signal Processing, Triruvananthapuram, pages 8-13, 2000.
6. Kumar, K.; Aggarwal, R.K., "Hindi Speech Recognition System Using HTK", International Journal of Computing and Business Research, ISSN (Online): 2229-6166, Volume 2 Issue 2, May 2011.
7. Aggarwal, R.K. and Dave, M., "Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 2, No. 4, December 2011.
8. Sivaraman. G.; Samudravijaya, K., "Hindi Speech Recognition and Online Speaker Adaptation", International Conference on Technology Systems and Management: ICTSM-2011, IJCA.
9. Gawali, Bharti W., Gaikwad, S., Yannawar, P., Mehrotra Suresh C., "Marathi Isolated Word Recognition System using MFCC and DTW Features (2010)", Int. Conf. on Advances in Computer Science 2010, DOI: 02.ACS.2010. 01.73.
10. Gaikwad, S.; Gawali, B.; Mehrotra, S. C.; "POLLY CLINIC INQUIRY SYSTEM USING IVR IN MARATHI LANGUAGE", International Journal of Machine Intelligence, ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 3, 2011, pp-142-145.
11. Mohanty, S.; Swain, B. K., "Continuous Oriya Digit Recognition using Bakis Model of HMM", International Journal of Computer Information Systems, Vol. 2, No. 1, 2011.
12. Mohanty, S.; Swain, B. K., "Markov Model Based Oriya Isolated Speech Recognizer-An Emerging Solution for Visually Impaired Students in School and Public Examination", Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; International Conference On Communication Technology-2010.
13. Nadungodage, T.; Weerasinghe, R., "Continuous Sinhala Speech Recognizer", Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011.
14. Raza, A., Hussain, S., Sarfraz, H., Ullah, I., and Sarfraz, Z., "An ASR System for Spontaneous Urdu Speech", Proceedings of O-COCOSDA'09 and IEEE Xplore, 2009.
15. Sarfraz, H.; Hussain, S.; Bokhari, R.; Raza, A. A.; Ullah, I.; Sarfraz, Z.; Pervez, S.; Mustafa, A.; Javed, I.; Parveen, R., "Large Vocabulary Continuous Speech Recognition for Urdu", International Conference on Frontiers of Information Technology, Islamabad, 2010.
16. Kumar, R., Singh, C., Kaushik, S., "Isolated and Connected Word Recognition for Punjabi Language using Acoustic Template Matching Technique", 2004.
17. Kumar, R., "Comparison of HMM and DTW for Isolated Word Recognition System for Punjabi Language", International Journal of Soft Computing 5(3):88-92, 2010.

AUTHOR PROFILE

Mr. Wiqas Ghai, B.E. (E&CE), MCA, M.Phil(CS) and presently pursuing Ph. D. from Punjabi University. He has around nine years of teaching experience including two years of industrial experience. His areas of interest are Automatic speech recognition, RDBMS, Visual Basic, SQL and C++.

Mr. Navdeep Singh, MCA from Thapar University, Ph.D. from Punjabi University. He has around nine years of teaching experience. His areas of interests are Automatic Speech Recognition, Software Engineering and Computer Networks.