Literature Review on Automatic Speech Recognition

Wiqas Ghai
Khalsa College (ASR) of Technology & Business Studies, Mohali, Punjab

Navdeep Singh
Mata Gujri College, Fatehgarh Sahib, Punjab

ABSTRACT
Automatic speech recognition, which was considered to be a concept of science fiction and which has been hit by number of performance degrading factors, is now an important part of information and communication technology. Improvements in the fundamental approaches and development of new approaches by researchers have lead to the advancement of ASRs which were just responding to a set of sounds to sophisticated ASRs which responds to fluently spoken natural language. Using artificial neural networks (ANNs), mathematical models of the low-level circuits in the human brain, to improve speech-recognition performance, through a model known as the ANN-Hidden Markov Model (ANN-HMM) have shown promise for large-vocabulary speech recognition systems. Achieving higher Recognition accuracy, low Word error rate, developing speech corpus depending upon the nature of language and addressing the issues of sources of variability through approaches like Missing Data Techniques & Convolutive Non-Negative Matrix Factorization, are the major considerations for developing an efficient ASR. In this paper, an effort has been made to highlight the progress made so far for ASRs of different languages and the technological perspective of automatic speech recognition in countries like China, Russian, Portuguese, Spain, Saudi Arab, Vietnam, Japan, UK, Sri-Lanka, Philippines, Algeria and India.

Keywords
Language Model, Hidden Markov Model, Vector Quantization, Dynamic Time Warping, Missing Data Techniques, Convolutive Non-Negative Matrix Factorization

1. INTRODUCTION
Review of literature on speech recognition systems genuinely demands the very first attention towards the discovery of Alexander Graham Bell about the process of converting sound waves into electrical impulses and the first speech recognition system developed by Davis et al. [1] for recognizing telephone quality digits spoken at normal speech rate. This effort for automatic recognition of speech was basically centred on the building up of *an electronic circuit* for recognizing ten digits of telephone quality. Spoken utterances were analyzed to get a 2-dimensional plot of formant 1 vs formant 2. For pattern matching, a circuit was designed for determining the highest relative correlation coefficient between a set of new incoming data and each of the reference digit patterns. It was also observed that circuit adjustment helps the recognition system to perform well for the speech of different speakers. An indication circuit was built to display the recognized spoken digit. The approaches to speech recognition, evolved thereafter, had a major stress on finding speech sounds and providing appropriate labels to these sounds. Various approaches and types of speech recognition systems came into existence in last five decades gradually. This evolution has lead to a remarkable impact on the development of speech recognition systems for various languages worldwide. Automatic speech recognition has been viewed as successive transformations of acoustic micro structure of speech signal into its implicit phonetic macro-structure. In other words, a speech recognition system is a speech-to-text conversion wherein the output of the system displays text corresponding to the recognized speech. Languages, on which so far automatic speech recognition systems have been developed, are just a fraction of total around 7300 existing languages. Russian, Portuguese, Chinese, Vietnamese, Japan, Spanish, Filipino, Arabic, English, Bengali, Tamil, Malayalam, Sinhala, Hindi are prominent among them. English is the language for which maximum work for recognition is done.

2. APPROACHES TO ASR

2.1 Acoustic-Phonetic approach
Hemdal & Hughes [2] took the basis of finding speech sounds and providing labels to them and proposed that there exist a fixed number of distinctive phonetic units in spoken language which are broadly characterized by a set of acoustics properties varying with respect to time in a speech signal. According this approach, the message bearing components of speech are to be extracted explicitly with the determination of relevant binary acoustic properties such as nasality, frication, voiced-unvoiced classification and continuous features such as formant locations, ratio of high and low frequencies. For commercial applications, this approach hasn't provided a viable platform. This approach is implemented in sequence: Spectral analysis, Features detection, Segmentation & Labelling, Recognising valid word. Linguistic constraints are applied to access the lexicon for word.

2.2 Pattern recognition approach
Itakura(1975) was the first to propose this approach which got a considerable support from Rabiner & Juang(1989,1993) for its further acceptance among the researchers. This approach has become the predominant method for speech recognition in the last six decades. Pattern training and pattern comparison are the two essential steps in this approach. Distinction of this approach is that it makes use of a well formulated mathematical framework and there-after establishes consistent speech pattern representations for reliable pattern comparison from a set of labelled training samples via a formal training algorithm. A speech pattern representation generally takes the form of a speech template (leading to template based approach) or a statistical model (leading to stochastic approach) which is equally applicable to a sound, a word, or a phrase. During

pattern-comparison stage of the approach, a direct comparison is made between the spoken words to be recognized with each possible pattern learned in the training stage for determining the identity of the unknown.

2.2.1 *Template bases Approach*

In this approach, a collection of prototypical speech patterns are stored as reference patterns which represents the dictionary of candidate words. An unknown spoken utterance is matched with each of these reference templates and a category of the best matching pattern is selected. Usually template for each word is constructed. This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. As a consequence, every word must have its own full reference template. Template preparation and matching become prohibitively expensive or impractical as vocabulary size increases. Watcher et al. [3] have made an attempt to overcome the key problems of HMM framework, i.e. discarding the information about time dependencies and over-generalisation, by applying template based continuous speech recognition with DTW. As a result no modelling and no training procedure are required. Explosion of search space due to DTW was kept in mind. Results were compared with that of HMM based CSR.

2.2.2 *Stochastic Approach*

This approach is based on the use of probabilistic models so that uncertain or incomplete information, such as confusable sounds, speaker variability s, contextual effects, and homophones words, can be dealt with. HMM modelling is more general and possesses firmer mathematical foundation in comparison to template based approach.

2.3 Knowledge based approach

This approach focuses on to mechanize the speech recognition process according to the way a person applies intelligence in visualising, analyzing, and characterizing speech based on a set of measured acoustic features. The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. Both acoustic phonetic and template based approach failed at their own to explore considerable insight into human speech processing. As a result, error analysis and knowledge based system enhancement couldn't get strength. In traditional Knowledge based approach, the production rules are created heuristically from empirical linguistic knowledge or from the observations from the speech spectrogram. Knowledge helps the algorithm to perform better and also in the selection of a suitable input representation, the definition of units of speech and the design of the recognition algorithm itself. Samouelian [4] proposed a data driven methodology for CSR, in which the knowledge about the structure and characteristics of the speech signal is acquired explicitly from the database by using inductive inference. This approach was found to have advantages of solving the problem of inter and intra speaker speech variability and also ability to generate decision trees. The recognition performance of this approach fell short and that may be attributed to the very small number of speakers taken. Tripathy et al. [5] proposed a knowledge based approach using a fuzzy inference algorithm for the classification of spoken English vowels. This technique gave better results over the standard MFCC feature analysis.

2.4 Connectionist Approach

This approach focuses on the representation of knowledge and integration of knowledge sources. Connectionist modelling of speech is the youngest development in speech recognition. In connectionist models, knowledge or constraints are distributed across many simple computing units rather than encoded in individual units, rules, or procedures. Uncertainty is also modelled by the pattern of activity in many units and not as likelihoods or probability density functions of a single unit. The computing units are simple in nature and knowledge lies in the connections and interactions between linked processing elements. The style of computation, performed by networks of these units, bears resemblance to the style of computation in the nervous system. Connectionist learning seeks to optimize or organize a network of processing elements. The simplicity and uniformity of the underlying processing element makes connectionist models attractive for hardware implementation, which enables the operation of a net to be simulated efficiently. Connectionism appears to hold great promise as plausible model of cognition. For instance, a new technique has been designed by Savage et al. [6] where an ANN block has been added to the output of each VQs. Global distances of each VQ representation for each word, has been fed to ANN. This technique resulted in the increase of recognition rate in comparison to the results obtained with VQ alone. Getting a 100% recognition rate with the combined application of VQ & ANN is great but the increase in recognition rate is not considerable. To improve upon the conventional VQ/HMM which suffers from quantization error, DVQ-distributed vector quantization technique has been designed to improve Discrete HMM based isolated word ASR system by Debyeche et al. [7]. Hybrid implementation of this technique involves K-mean DVQ and NN-DVQ and out of these two, NN-DVQ has given encouraging results with due regards to error reduction. Hatulan et al. [8] took an initiative for Filipino language and developed a speech to text converter using ANN/HMM approach. Feed-forward ANN were used in training the networks to a specified target. The outputs were fed as input to HMM for predicting the most probable phoneme sequence.

2.5 Support Vector Machines

SVM is one of the powerful state-of-the art classifiers for pattern recognition which uses a discriminative approach. Optimised margin, between the samples and the classifier border, helps to generalise unseen patterns. SVMs use linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The variable length data has to be transformed to fixed length vectors before SVMs can be used. It is a generalized linear classifier with maximum-margin fitting functions. This fitting function provides regularization which helps the classifier generalized better. SVM controls the model complexity by controlling the VC dimensions of its model rather than controlling model complexity by using a small number of features. This method is independent of dimensionality and can utilize spaces of very large dimensions spaces, which permits a construction of very large number of non-linear features and then performing adaptive feature selection during training. By shifting all non-linearity to the features, SVM can use linear model for which VC dimensions is known. Sendra et al. [9] have worked on a pure SVM-based continuous speech recogniser by applying SVM for making decisions at frame level and a Token Passing algorithm to obtain the chain of recognized words. The Token Passing Model is an extension of the Viterbi algorithm meant for continuous speech recognition so as to manage the uncertainty about the number of words in a sentence. The results achieved from the experiments have concluded that with a small database, recognition accuracy improves with SVMs but with the large database, same result is obtained at the expense of huge computational effort.

2. PHASES OF ASR

Automatic speech recognition system involves two phases: Training phase and recognition phase. A rigorous training procedure is followed to map the basic speech unit such as phone, syllable to the acoustic observation. In training phase, known speech is recorded, pre-processed and then enters the first stage i.e. Feature extraction. The next three stages are HMM creation, HMM training and HMM storage. The recognition phase starts with the acoustic analysis of unknown speech signal. The signal captured is converted to a series of acoustic feature vectors. Using suitable algorithm, the input observations are processed. The speech is compared against the HMM's networks and the word which is pronounced is displayed. An ASR system can only recognize what it has learned during the training process. But, the system is able to recognize even those words, which are not present in the training corpus and for which sub-word units of the new word are known to the system and the new word exists in the system dictionary.

3. MODULES OF ASR

Modules identified for a speech recognition system (Fig. 1) are

i. Speech Signal acquisition
ii. Feature Extraction
iii. Acoustic Modelling
iv. Language & Lexical Modelling
v. Recognition

Two of these modules Speech acquisition and Feature extraction are common to both the phases of ASR.

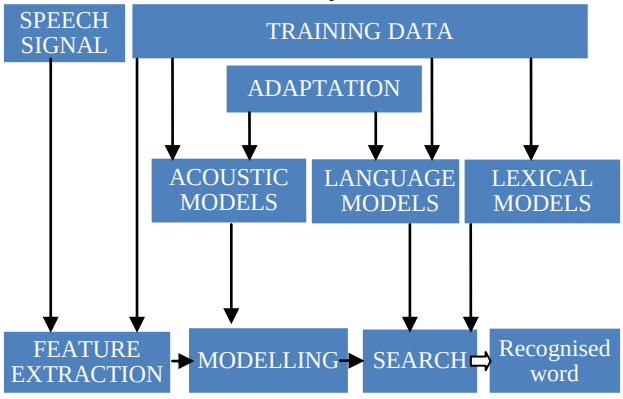

Figure 1: ASR BLOCK DIAGRAM

Model adaptation is meant for minimizing the dependencies on speakers' voice, acoustic environment, microphones and transmission channel, and to improve the generalization capability of the system.

3.1 Feature Extraction

Feature extraction (Fig. 2) requires much attention because recognition performance depends heavily on the feature extraction phase. LPC, MFCC, AMFCC, RAS, DAS, ΔMFCC, Higher lag autocorrelation coefficients, PLP, MF-PLP, BFCC, RPLP are the different techniques for feature extraction. It has been found that noise robust spectral estimation is possible on the higher lag autocorrelation coefficients. Therefore, eliminating the lower lags of the noisy speech signal autocorrelation leads to removal of the main noise components. Jain and Saxena [10] applied different techniques such as RAS, DAS, MFCC, AMFCC and Higher lag auto correlation coefficients to evaluate the extraction and implementation of features from the input speech signal. During testing, they knowingly added different types of noises such as White noise, F16 noise, Babble noise and factory noise. Recognition rates were compared, for different feature extraction algorithms and noise types at various noise levels, with the help of graphs.

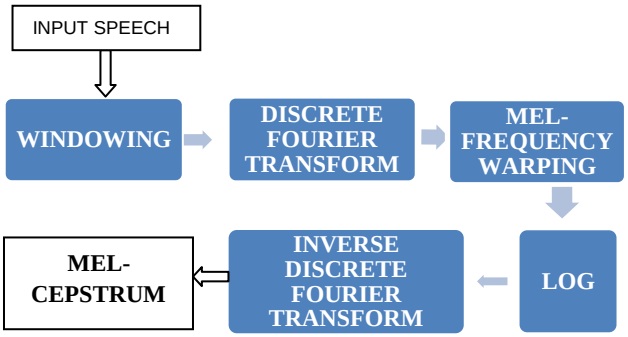

Figure 2: FEATURE EXTRACTION BLOCK DIAGRAM

It was found that higher lag autocorrelation algorithms gave the better results. Out of above list of techniques for feature extraction BFCC, RPLP, MF-PLP are being used for robust feature selection. Secondly, it has been found that F0 contour is the most essential characteristic to differentiate various tones and MFCC & PLP fail to provide it. Bi-gram language model is also not capable to provide a solution to tonal problem in spite of the improvement in the accuracy. PRAAT tool helps in extracting F0 contours. Thirdly, during feature extraction module, frequency domain features are obtained using Fourier transform or wavelet transform. Wavelet transform provides better time-frequency localisation for tracking sudden changes in speech signals. Wavelet transform uses wavelets in place of sine waves of different frequencies and provides different resolution for each scale. Other aspects, which bear very great influence on creation of ASR, are: Acoustic Model, Lexical Model and Language Model.

3.2 Acoustic model

Acoustic model is the main component for an ASR and it accounts for most of the computational load and performance of the system. It is used to link the observed features of the speech signals with the expected phonetics of the hypothesis sentence. The Acoustic model is developed for detecting the spoken phoneme. Its creation involves the use of audio recordings of speech and their text scripts and then compiling them into a statistical representation of sounds which make up words. Aggarwal and Dave [11] have made an attempt to provide a comprehensive overview of acoustic modelling in the context of ASR. At present Gaussian mixture models are the dominant technique for modelling the emission distribution of hidden Markov models for automatic speech recognition. HMM suffers two major drawbacks. Strong Independency assumption in HMM states that frames are independent, given a state. As a result, it lacks an ability to deal with a feature which straddles over several frames. Features such as delta coefficient, segmental statistics and modulation spectrum have been developed which can deal with phenomena of straddling. Aggarwal and Dave [12] have reviewed the variety of modifications and extensions adopted for the HMM based acoustic models in the form of refinements such as variable duration models, discriminative techniques, connectionist approach (HMM+ANN) to overcome the limitations of traditional HMM and advancements such as margin based methods, wavelets and dual stream approach. Ostendorf et al. [13] also came up with segmental models to overcome this weakness of HMM. Secondly, HMM is a generative model and fails to discriminate sequences. This weakness has aroused due the maximisation of maximum likelihoods (MLI) instead of maximum a posteriori probabilities (MAP). Some alternative

algorithms have been developed for training such as MMI-Maximum mutual information, MCE-Minimum classification error, MPE-Minimum phone error have been developed. Hidden Conditional Random Fields (HCRF) is also capable of overcoming the above two major drawbacks of HMM while preserving the merits of HMM such as efficient algorithms including forward-backward algorithm and Viterbi decoding. But HCRF has a drawback not considering non-linearity among features which would be crucial for speech recognition. HCRF approach becomes difficult to use when the number of features increases. Yasuhisa et al. [14] have proposed Hidden Conditional Neural Fields- HCNF which can easily consider non-linearity between features by introducing gate function into HCRF. It has also been found that HCNF can be trained without any initial model and incorporate any kinds of features. It was found that results of HCRFs were inferior to the results of HMMs because of the implementation of HCRF without using mixtures. HCNFs have clearly outperformed HCRFs and the result showed the effectiveness of incorporating the gate function into HCRF. This result was superior to the results of HMMs and comparable with the best ones of previous results in mono-phone setting. Mohamed et al [15] have proposed a technique in which Gaussian mixture models have been replaced by multi layer feed-forward neural networks where multiple layers of features have been generatively pre-trained. It has been the very first application of neural networks to acoustic modelling. This approach provides a hierarchical framework where each layer is designed to capture a set of distinctive feature landmarks. A specialized acoustic representation is constructed for each feature in which the corresponding feature is easy to detect. Discriminative fine tuning was performed using back propagation to slightly adjust the features so as to make them better at predicting a probability distribution over the states of mono-phone hidden Markov models.

3.3 Lexical Models

Lexicon is developed to provide the pronunciation of each word in a given language. Through lexical model, various combinations of phones are defined to give valid words for the recognition. Neural networks have helped to develop lexical model for non-native speech recognition.

3.4 Language Models

Language model is the single largest component trained on billion of words, consisting of billions of parameters and developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary. ASR systems utilise *n*-gram language models to guide the search for correct word sequence by predicting the likelihood of the *n*th word on the basis of the $n-1$ preceding words. The probability of occurrence of a word sequence W is calculated as:

$$P(W) = P(w_1, w_2, \ldots\ldots\ldots\ldots, w_{m-1}, w_m)$$
$$= P(w_1).\ P(w_2|w_1).\ P(w_3|w_1w_2)\ldots\ldots\ldots\ldots$$
$$P(w_m|w_1w_2w_3\ldots\ldots\ldots\ldots w_{m-1}).$$

During the construction of n-gram language models for large vocabulary speech recognizers, two problems are being faced. Large amount of training data generally leads to large models for real applications. Second is the sparseness problem, which is being faced during the training of domain specific models. Language models are cyclic and non-deterministic. Both these features make it complicated to compress its representations. Sorensen and Allauzen [16] proposed a technique based on unary degree sequences/ LOUDS which have a capability to avoid the use of indices and pointers. Unlike previous proposed systems for compressing language models, it does so while simultaneously improving access time.

4. SPEECH DATABASE FOR ASR

To develop a speech database for a speech recognizer, following is the general methodology adopted:

4.1 Text Corpus

Generation of optimal set of textual sentences is the first step. This set will be made available for recoding speech by the native speakers. Its creation involves:

a. Text Corpus Collection
b. Phonetizing Text Corpus
c. Optimal Text Collection

4.2 Speech Data Collection

Text corpus is used finally to record the words/sentences through a single speaker or number of speakers depending upon the requirements. Precisely it involves the following steps:

a. Selection of Speaker
b. Data Statistics
c. Transcription Correction

Lot of research has been done on clear speech using written text material which is generally read by a talker in conversational or clear styles. It has been found that read clear speech and spontaneously elicited clear speech is not same acoustic phonetically. Kain et al. [17] have proposed a communicative setting which allows talkers to naturally hear themselves while speaking and allows listening sound pressure levels to be controlled. They compared the words spoken in two different conditions of normal hearing and simulated hearing loss respectively. Acoustic phonetic properties of keywords in the first format were found to correspond to conversational speech and keywords spoken in the second format were found to correspond to moderately clear speech. There are languages which have regional and social variations with regard to pronunciation. As a result, all or majority of the dialect regions are to be considered while going for the design of speech corpus. Modern standard Arabic is one of such languages in Algeria. While designing speech corpus Algerian Arabic speech database by Hamdani [18], six regions from southern and northern Algeria so as to cover all the regional and social variations of MSA. It has been found that these regions are having maximum homogeneous phonetic features and minimum phonological differences and this fact helped the researchers to get higher recognition rate in their speech recognizer.

5. TOOLS FOR ASR

PRAAT: It is free software with latest version 5.3.04 which can run on wide range of OS platforms and meant for recording and analysis of human speech in mono or stereo

AUDACITY: It is free, open source software available with latest version of 1.3.14(Beta) which can run on wide range of OS platforms and meant for recording and editing sounds.

CSL: Computerised Speech Lab is a highly advanced speech and signal processing workstation (software and hardware). It possesses robust hardware for data acquisition and a versatile suite of software for speech analysis.

HTK: The basic application of open source Hidden Markov Toolkit (HTK), written completely in ANSI C, is to build and manipulate hidden Markov models. This tool kit has been originally designed for to recognize English, so the characters are stored as 8-bit ASCII standard code. But there are languages which don't have this format. For example, Vietnamese language is stored in UTF-8 format. In this case

Uni-key software was used to convert Vietnamese character format from UTF-8 to VIQR code which is a convention for writing Vietnamese using ASCII 7 bit format. Nguyen et al. [19] kept this point in mind and developed an automatic speech recognition system using HTK.

SPHINX: Sphinx 4 is a latest version of Sphinx series of speech recognizer tools, written completely in Java programming language. It provides a more flexible framework for research in speech recognition.

JULIUS: It is an open source high performance two-pass large, two-pass large vocabulary continuous speech recognition decoder software which works well on Linux OS. During its last revision, a grammar based recognition parser "Julian" has been integrated it. Mathur et al. [20] used Julius to develop a domain specific speaker independent continuous speech recognizer. In spite of creating just a base line recognizer, the results were encouraging.

SCARF: It is a software toolkit designed for doing speech recognition with the help of segmental conditional random fields.

MICROPHONES: They are being used by researchers for recording speech database. Sony and I-ball has developed some microphones which are unidirectional and noiseless.

6. TYPES OF ASRs

Type of speech or speaking mode, dependence on speaker, size of vocabulary and bandwidth are the different basis on which researchers have worked. Among them, speaking mode is one of the main criteria which lead to the evolution of following ASRs. Rest of the factors have been dealt along with.

6.1 Isolated word speech recognition: IWR

A speech recognition system, where a discrete utterance is dealt with two implicit assumptions, is specifically considered as isolated word speech recognition. First assumption is that speech to be recognised consists of a single word/phrase and it is going to be recognised as a complete entity with no explicit knowledge for the phonetic content of the word/phrase. Second assumption is that each spoken word/phrase has a clearly defined beginning and ending point. Command and Control systems are one of the applications where this type of speech recognition can be applied. Kumar and Aggarwal [21] have developed a Hindi isolated word speech recognizer using HTK on Linux platform which recognizes isolated words using acoustic word model. Vocabulary used for this system used just 30 words. Gupta [22] has developed a speaker independent isolated Hindi word speech recognizer for recognizing the ten digits in Hindi, using continuous HMM which can support a vector as an observation, for the same. Hamming window was used for feature extraction keeping the amount of distortion into consideration. The system used acoustic word model as a knowledge model. Frame rate of 10 ms and hamming window of 25 ms were applied in both the systems. Venkataramani [23] has worked on the development of an on-line 'speech to text engine' for isolated word recognition on a vocabulary of 10 words (digits 0-9) which was implemented as a system on a programmable chip (SOPC). Speech was acquired at run time through a microphone and processes the sampled speech to recognize the uttered text. Hidden Markov model (HMM) for speech recognition has been used to convert the speech to text which gets stored in a file that connects to an FPGA on a development board using a standard RS-232 serial cable.

Existing ASR systems are mainly focussed on the acoustic signal patterns whose performance degrades in the presence of ambient noise. Lee [24] proposed one more source for speech information i.e. EMG – **Electromyogram signals** which are emitted by the articulatory muscles. This approach is based on the fact that there are different phonemes for each vocal articulation. EMG signals have been applied to classify the phonemes and words. To create a relationship between each facial muscle and recognition performance, a trial and error approach was used heuristically in which surface EMG signals were obtained from three articulatory muscles: the levator anguli oris, the zygomaticus major, and the depressor anguli oris. Recognition was performed on isolated words. It was found in results that HMMs derived from the dependent model produced better recognition accuracy than the independent model.

6.2 Connected word recognition: CWR

Connected word speech recognition is the system where the words are separated by pauses. Connected word speech recognition is a class of fluent speech strings where the set of strings is derived from small-to-moderate size vocabulary such as digit strings, spelled letter sequences, combination of alphanumeric. Like isolated word speech recognition, this set too has a property that the basic speech-recognition unit is the word/phrase to much extent. Rabiner et al. [25] analysed three algorithms designed for connected word recognition: Two level DP approach, Level Building approach and One Pass approach are three algorithms and found them to be providing the identical best matching string with the identical matching score for connected word recognition. But they differ in computational efficiency, storage requirement and ease of realisation in real time hardware. Garg et al. [26] have developed a speaker dependent connected digits recognition system by applying unconstrained Dynamic time warping technique in which they recognised each digit by calculating distance w.r.to matching of input spoken digit with stored template. Mishra et al. [27] have developed a connected Hindi digits recognition system using robust features extraction techniques and HTK as recognition engine. SYAMA [28] built an isolated word and speaker independent speech recognition system for Malayalam. Microsoft Visual Studio was used for compiling HTK and Active Perl as interpreter. He evaluated the system for both isolated as well as connected word recognition. Accuracy of this system is just 62%. An improved language model and a spell-checker acting as a linguistic resource may enhance the %age of recognition accuracy. Kumar, Ravinder et al. [29] 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 was used to transform signal parameters to codebook indices and Dynamic time warping techniques was used for finding the lowest distance path through the matrix, with some modification to noise and word detection algorithms. VC++ was used to program sound blaster card with MCI commands. Accuracy of this ASR was just 61% and still more work is said to be done on Punjabi language.

6.3 Continuous speech recognition: CSR

Continuous speech recognition deals with the speech where words are connected together instead of being separated by pauses. As a result unknown boundary information about words, co-articulation, production of surrounding phonemes and rate of speech effect the performance of continuous speech recognition systems. It has been found that there are 3 approaches to speech recognition w.r.to the choice of sub-word units: Word based, phone based, syllable based.

6.3.1 Word Based

Word unit is acoustically well defined and the acoustic variation generally occurs in the beginning and the end of the word. The problem with the choice of word unit is that each word has to be trained individually and as a result no sharing of the parameters is possible. This leads to the need of setting up a very large training set and growing memory requirement. In spite of this issue, word models have been successfully applied for building limited vocabulary ASR but it's not practically applicable for LVCSR.

6.3.2 Phone Based

Phone model have the capability to overcome this problem by incorporating the sharing of parameters and thereby saving the computing resources. But the phones are highly context dependent and their aspiration varies too across the word. Over-generalization by phone models in comparison to no generalization by word models is a strange observation for the researchers. Phone-in-context and tri-phone have provided solution to over-generalisation by adding right and left contexts. Tri-phone sub-word models have provided better acoustic modelling and considerable reduction in word error rate for LVCSR systems. Thangarajan et al. [30] have developed a small vocabulary word-based and medium vocabulary tri-phone based continuous speech recognizer for Tamil language. Larger memory requirement is a limitation of tri-phone sub-word model approach.

6.3.3 Syllable Based

Syllable is a larger sub-word unit which is being employed as a model for acoustic modelling. Out of its three constituents: Onset, Nucleus, Coda, only nucleus has no contextual dependencies. Contextual effect of coda of current syllable with the onset of the following syllable is the issue to be dealt further. There are languages where pronunciation mainly depends on the syllable and as a result, syllable is being used as a sub-word model. There are strong linguistic rules to form syllables in phonetic languages where as in languages like English, syllabication is fuzzy. Thangarajan et al. [31] have worked for developing a continuous speech recognizer for Tamil language using syllable as a sub-word unit for building acoustic model. A syllable based lexicon was created through an algorithm where each word was segmented into its constituent prosodic syllables. Recognition accuracy was fairly good but around 10% increase in WER has been attributed to large number of syllables to be modelled with the available limited training set. Abushariah et al. [32] worked for the development of continuous speech recognition system on Arabic Language using Sphinx as well as HTK tools. Five-state Hidden Markov Models (HMM) having 3 emitting states for tri-phone acoustic modelling were used. Their statistical Language model contained uni-grams, bi-grams, and tri-gram. This system was tested for different combinations of speakers and sentences. To ensure and validate the pronunciation correctness of the speech data, a manual human classification and validation of the correct speech data was conducted. A round robin technique was applied for fair testing & evaluation of this system and to make this system speaker independent. The word recognition accuracy was best for different speakers but similar sentences and was least for different speakers and different sentences.

6.3.4 Additional Morpheme Level

To have much more variety on word form level in a language, leads to increase in the size of vocabulary and decrease in the speed and quality of processing. These problems have a solution in the additional morpheme level of speech signal representation. Ronzhin and Karpov [33] kept these factors into consideration while developing a large vocabulary continuous speech recognition system. Incorporation of morpheme level, the size of needed vocabulary got reduced. HMM with mixture Gaussian probability density function was used as an acoustic model. They studied the peculiarities of Russian language on to a great depth such as longer size of Russian words, set of accents & dialects, strict grammatical constructions, diverse phonetic structure. They applied dynamic warping of sentences to create a search of optimal matching of two sentences. Recognition accuracy of morpheme based recognizer came about 95% which was found be 1.7 times faster than word based recognizer. Vietnamese is a syllabic tonal language with six tones where each syllable has only one tone. The meaning of the word depends on the tone. Keeping this factor into consideration, Thang tat et al. [34] developed a LVCSR for Vietnamese language and applied the combination of MFCC & F0 features and bigram language model to improve the accuracy of their ASR. Incorporation of F0 gave a significant increase of around 10% in recognition accuracy and decrease of around 36% in error.

6.3.5 Sub-syllable Based

For better speech recognition on small size of trained speech data, another smaller component has been used i.e. sub-syllable. For example, Chinese is a mono-syllable and tonal language in which each syllable of a character is composed of an initial and a final. Huang feng-long [35] used sub-syllable for generating features while developing an independent speech recognition system using HMM for small vocabulary. To improve the performance, they applied keyword-spotting criterion. This criterion has a basis that in spite of the ASRs being guided by grammatical constraints, speaking natural sentences and noise lowers the performance of an ASR. Sinhala is one of the less-resourced non-Latin language for which speaker dependent continuous speech recognizer have been developed using HTK by Nadungodage and Weerasinghe [36]. A considerable increase in the size of vocabulary for continuous speech recognizer due to the differences between written and spoken Sinhala, pushed them to take only written Sinhala vocabulary.

6.4 Spontaneous speech recognition

A spontaneous speech is a speech which is natural sounding and not rehearsed. An ASR for such a speech handles a variety of natural speech features i.e. words being run together, "ums", "ahs" and slight stutters. The unavailability of sufficient amount of transcribed spontaneous speech data pushed Raza et al. [37] to work on it and train a single speaker, medium vocabulary spontaneous speech recognition system for Urdu language with a varying mixture of read and spontaneous speech data. This mixture contained tokens in the ratio 1:2, unique words in the ratio 5:2 and phoneme occurrence in the ratio 5:9. Variation in the mixture of the two types of data brought changes in the Language model too. Reduction in WER with the incorporation of read data up to 1:1 ratio and there after increase in WER with further increase in read data, is the interesting result obtained.

7. DEVELOPING ASR

Steps to develop a general automatic speech recognizer have been observed as:

i. First of all the speech is acquired through a uni-directional and noiseless microphone.
ii. Signal parameterisation using a suitable feature extraction technique
iii. Acoustic analysis: The training waveforms are converted into some series of coefficient vectors.

iv. Definition of the models: A prototype of HMM is defined for each element of the task vocabulary.
v. Training of the models: Each HMM is initialised and trained with the trained data.
vi. Definition of the task: Here the grammar for the speech recognizer is defined.
vii. Recognition: The unknown input speech signal is recognized at this stage.

There have been some alterations at one or more steps applied by the researchers due to improvement in existing techniques, advent of new techniques and variation in acoustic structure of different languages from time to time.

8. PERFORMANCE OF ASR

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system.

8.1 Accuracy

8.1.1 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.

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*).

8.1.2 Single Word Error Rate (SWER)

8.1.3 Command Success Rate (CSR)

8.2 Speed

Real Time Factor is parameter to evaluate speed of automatic speech recognition. If it takes time *P* to process an input of duration *I*, the real time factor is defined as

$$RTF = \frac{P}{I}$$

e. g. Real time factor is 2 if it takes 6 hours of computation time to process a recording of duration 3 hours. RTF $\leq$ 1 implies real time processing.

9. ROBUSTNESS OF ASR

For achieving the true robustness in ASR, all the sources of variability are to be handled at a priority. These are sources which are part of inevitable critical environments in real world applications and accordingly automatic speech recognition suffers degradation in recognition performance.

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

9.1 Efforts for Compensation

Following are the efforts made by few researchers to deal with the sources of variability mentioned above.

i. The varied nature of **background noise** always creates a challenging task to learn the noise bases in a multi-source noise environment and thereby to suppress it from the speech signal. A large amount of training data is required to reliably capture noise variation. Vipperla et al. [38] addressed the problem of recognizing the speech of target speaker, under typical ambient noise conditions recorded in a home environment came up, with the help of a computationally efficient online CNMF- Convolutive Non-Negative Matrix Factorization approach. They approached the problem from a signal enhancement perspective. This approach has been found to be extremely beneficial in the scenario where global noise bases are learnt from over several hours of data.

ii. Room acoustics, Speaker specific behaviour, background noise and microphone characteristics are the unavoidable sources which corrupt the real world speech. Reason behind this declined performance is that the observed acoustic features no longer match the acoustic models. Gemmeke et al. [39] made an attempt to handle above problem and thereby to improve the performance of ASR by combining MDT-Missing Data Techniques with three conventional methods: multi-condition training, de-reverberation and spectral subtraction.

iii. There are variations in the uttered speech by different individuals due to different geographical boundaries, social background, age, gender, occupation etc. An attempt has been made by Paul & Parekh [40] for modelling the same isolated words spoken by different individuals by using four features: first three formant frequencies and zero-crossing rate of the audio signal. They used the multi-layer neural networks for the classification. Results were better in comparison to wavelet based features.

iv. Pronunciation variation is another major cause of performance degradation for a variety of ASR tasks. Modelling of pronunciations in a speech recognizer is done by a phonemic dictionary accompanied by a set of rewrite rules to account for phonological variation. It has been observed that even a well defined lexicon some time or most of the times, fails to support all variations in human's pronunciation. Sharing Gaussian densities across phonetic models and decision tree have proved to be efficient and that also without dictionary modification but Kanokphara et al. [41] have addressed this issue through an alternative method where a re-label strategy has been modified to have rule based pronunciation so that real phonetic acoustic models can be developed. Different combinations of initial phonemic transcriptions and training strategy were tried. It has been found that pronunciation variation system initialised from phonemic transcriptions generated by re-label training gave good results.

v. To compensate the sources of variability such as inter-speaker, channel, environmental and transducer variability, Potamianos & Rose [42] made an attempt to apply frequency warping for implementing linear model transformation and speaker normalisation. It was found that on combining frequency warping and spectral shaping, there was a considerable improvement in the WER reduction.

10. AREAS OF APPLICATION

Growing interest researchers in the development of new techniques for different phases of automatic speech recognition systems, have lead to the applications of ASRs in different fields such as:

i. Command and Control Systems
- Voice Repertory Dialer [43]
- Automated Call-Type Recognition
- Call Distribution by Voice commands
- Directory Listing Retrieval [44],[45]

- Credit Card Sales Validation

ii. Call Routing: e.g. I would like to make a collect call.

iii. Transcription system using ASR in Japanese Parliament [46]

iv. Demotic appliance control and content based spoken audio search: e.g. find a pod cast where particular words were spoken.

v. Automated data entry: e.g. entering a credit card numbers.

vi. Preparation of structured documents: e.g. radiology report.

vii. Speech to text processing: e.g. word processors or emails.

viii. Direct voice input: e.g. In aircraft cockpits

ix. Home automation

x. Transcription of speech to mobile text message.

xi. Court Reporting

xii. Vehicle Navigation System

11. CONCLUSION

Researchers, working on the very promising and challenging field of automatic speech recognition, are collectively heading towards the ultimate goal i.e. Natural Conversation between Human beings and machines, are applying the knowledge from areas of Neural Networks, Psychoacoustics, Linguistics, Speech Perception, Artificial Intelligence, Acoustic-Phonetics etc.. The challenges to the recognition performance of ASR are being provided concrete solutions so that the gap between recognition capability of machine and that of a human being can be reduced to maximum extent. An attempt has been made through this paper to give a comprehensive survey and growth of automatic speech recognition over the last six decades through the never ending efforts of researchers in countries like China, Russian, Portuguese, Spain, Saudi Arab, Vietnam, Japan, UK, Sri-Lanka, Philippines, Algeria and India.

12. REFERENCES

[1] Davis, K., Biddulph, R., and Balashek, S., "Automatic Recognition of Spoken Digit," J. Acoust. Soc. Am. 24: Nov 1952, p. 637.

[2] Hemdal, J.F. and Hughes, G.W., A feature based computer recognition program for the modeling of vowel perception, in Models for the Perception of Speech and Visual Form, Wathen-Dunn, W. Ed. MIT Press, Cambridge, MA.

[3] Watcher, M. D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., "Template Based Continuous Speech Recognition", IEEE Transaction on Audio, Speech, & Language Processing, 2007.

[4] Samoulian, A., "Knowledge Based Approach to Speech Recognition", 1994.

[5] Tripathy, H. K., Tripathy, B. K., Das, P. K., "A Knowledge based Approach Using Fuzzy Inference Rules for Vowel Recognition", Journal of Convergence Information Technology Vol. 3 No 1, March 2008.

[6] Savage, J., Rivera, C., Aguilar, V., "Isolated word speech recognition using Vector Quantization Techniques and Artificial Neural Networks", 1991.

[7] Debyeche, M., Haton, J.P., Houacine, A., "Improved Vector Quantization Technique for Discrete HMM speech recognition system", International Arab Journal of information Technology, Vol. 4, No. 4, October 2007.

[8] Hatulan, R. J. F., Chan, A. J. L., Hilario, A. D., Lim, J. K. T., and Sybingco, E., "Speech to text converter for Filipino Language using Hybrid Artificial Neural Network and Hidden Markov Model", ECE Student Forum December 1, 2007 De La Salle University.

[9] Sendra, J. P., Iglesias, D. M., Maria, F. D., "Support Vector Machines For Continuous Speech Recognition", 14th European Signal Processing Conference 2006, Florence, Italy, Sept 2006.

[10] Jain, R. And Saxena, S. K., "Advanced Feature Extraction & Its Implementation In Speech Recognition System", IJSTM, Vol. 2 Issue 3, July 2011.

[11] Aggarwal, R.K. and Dave, M., "Acoustic Modelling Problem for Automatic Speech Recognition System: Conventional Methods (Part I)", International Journal of Speech Technology (2011) 14:297–308.

[12] Aggarwal, R. K. and Dave, M., "Acoustic modelling problem for automatic speech recognition system: advances and refinements (Part II)", International Journal of Speech Technology (2011) 14:309–320.

[13] Ostendorf, M., Digalakis, V., & Kimball, O. A. (1996). From HMM's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, *4*(5), 360–378.

[14] Yasuhisa Fujii, Y., Yamamoto, K., Nakagawa, S., "AUTOMATIC SPEECH RECOGNITION USING HIDDEN CONDITIONAL NEURAL FIELDS", ICASSP 2011: P-5036-5039.

[15] Mohamed, A. R., Dahl, G. E., and Hinton, G., "Acoustic Modelling using Deep Belief Networks", submitted to IEEE TRANS. On audio, speech, and language processing, 2010.

[16] Sorensen, J., and Allauzen, C., "Unary data structures for Language Models", INTERSPEECH 2011.

[17] Kain, A., Hosom, J. P., Ferguson, S. H., Bush, B., "Creating a speech corpus with semi-spontaneous, parallel conversational and clear speech", Tech Report: CSLU-11-003, August 2011.

[18] Hamdani, G. D., Selouani, S. A., Boudraa, M., "ALGERIAN ARABIC SPEECH DATABASE (ALGASD): CORPUS DESIGN AND AUTOMATIC SPEECH RECOGNITION APPLICATION", The Arabian Journal for Science and Engineering, Volume 35, Number 2C, Dec 2010.

[19] NGUYEN Hong Quang, TRINH Van Loan, LE The Dat, "Automatic Speech Recognition for Vietnamese using HTK", 2004.

[20] Mathur, R., Babita, Kansal, A., "Domain specific speaker independent continuous speech recognizer using Julius", Proceedings of ASCNT – 2010, CDAC, Noida, India, pp. 55 – 60.

[21] Kumar, K. and 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.

[22] Gupta, R., and Sivakumar, G., "*Speech Recognition for Hindi Language*", IIT BOMBAY, 2006.

[23] Venkataramani, B., "*SOPC-Based Speech-to-Text Conversion*", 2006.

[24] Lee, K.S., "EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables" IEEE Transactions on Biomedical Engineering, vol. 55, issue-3, pp: 930-940, March 2008.

[25] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.

[26] Garg, A., Nikita, Poonam, "Connected digits recognition using Distance calculation at each digit", IJCEM International Journal of Computational Engineering & Management, Vol. 14, October 2011, ISSN (Online): 2230-7893.

[27] Mishra, A. N., Biswas, A., Chandra, M., Sharan, S. N., "Robust Hindi connected digits recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 2, June, 2011.

[28] Syama, R. and Mary Idicula, S., "*Speech Recognition for Malyalam Language*", 2008.

[29] Kumar, R., Singh, C., Kaushik, S., "Isolated and Connected Word Recognition for Punjabi Language using Acoustic Template Matching Technique",2004.

[30] Thangarajan, R., Natarajan, A.M., Selvam, M., "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language", March 2008.

[31] Thangarajan, R., Natarajan, A.M., Selvam, M., "Syllable based Continuous Speech Recognition for Tamil", Jan 2008.

[32] Mohammad A. M. Abushariah, Moustafa Elshafei, Othman O. Khalifa, "*Natural Speaker-Independent Arabic Speech Recognition System Based on Hidden Markov Models Using Sphinx Tools*", May 2010.

[33] Ronzhin, A. I., Karpov, A. A., "*Large Vocabulary Automatic speech recognition for Russian Language*", 2004.

[34] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom, "Vietnamese Large Vocabulary continuous speech recognition", 2004.

[35] Huang Feng-Long, "An Effective approach for Chinese speech recognition on small size of vocabulary", Signal & Image Processing: An International Journal (SIPIJ) Vol.2, No.2, June 2011.

[36] Nadungodage, T. and Weerasinghe, R., "Continuous Sinhala Speech Recognizer", Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

[37] Raza, A., Hussain, S., Sarfraz, H., Ullah, I., and Sarfraz, Z., "An ASR System for Spontaneous Urdu Speech " in *Proceedings of O-COCOSDA'09 and IEEE Xplore*, 2009.

[38] Vipperla, R., Bozonnet, S., Wang, D., Evans, N. "Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization", CHIME Workshop on Machine Listening in Multisource Environments, Sept 2011.

[39] Gemmeke, J. F., Segbroeck, M. V., Wang, Y., Cranen, B., Hamme, H. V., "Automatic speech recognition using missing data techniques: Handling of real-world data", 2011.

[40] Paul, D., and Parekh, R., "Automatic Speech Recognition of Isolated Words Using Neural Networks", Vol. 3 No. 6, IJEST-2011.

[41] Kanokphara, S., Tesprasit, V., Thongprasirt, R., "Pronunciation Variation Speech Recognition without Dictionary Modification On Sparse Database",2002.

[42] Potamianos, A., and Rose, R.C., "On Combining Frequency Warping and Spectral Shaping for HMM based Speech Recognition", IEEE international conference on acoustics, Speech, & Signal Processing, April1997.

[43] Rabiner, L. R., Wilpon, J. G., Rosenberg, A. E., "A Voice Controlled Repertory-Dialer System", The Bell System Technical Journal Vol. 59, No. 7, 1980.

[44] Aldefeld, B. Rabiner, L.R. Rosenberg, A.E. Wilpon, J.G., "Automated Directory Listing Retrieval System based on Isolated Word Recognition", Vol 68, issue 11, Nov 1980.

[45] Myers, C. S. And Rabiner, L. R., "Automated Directory Listing Retrieval System based on recognition of connected letter strings, Journal of the Acoustical Society of America, Vol. 71, No. 3, Mar 1982.

[46] Kawahara, T., "New Transcription System using ASR in Japanese Parliament", Academic Center for Computing and Media Studies, Kyoto University, 2011.