International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.4, (2016), pp.393-404
http://dx.doi.org/10.14257/ijsip.2016.9.4.34

A Review on Automatic Speech Recognition Architecture and Approaches

Karpagavalli S[*1] and Chandra E[2]

[1]*Department of Computer Science, PSGR Krishnammal College for Women , Coimbatore 641 004, India*
[2]*Department of Computer Science, Bharathiar University, Coimbatore 641 046, India*
[*1]*karpagavalli@psgrkc.com*
[2]*crcspeech@gmail.com*

Abstract

Speech is the most natural communication mode for human beings. The task of speech recognition is to convert speech into a sequence of words by a computer program. Speech recognition applications enable people to use speech as another input mode to interact with applications with ease and effectively. Speech recognition interfaces in native language will enable the illiterate/semi-literate people to use the technology to greater extent without the knowledge of operating with computer keyboard or stylus. For more than three decades, a great amount of research was carried out on various aspects of speech recognition and its applications. Today many products have been developed that successfully utilize automatic speech recognition for communication between human and machines. Performance of speech recognition applications deteriorates in the presence of reverberation and even low levels of ambient noise. Robustness to noise, reverberation and characteristics of the transducer is still an unsolved problem that makes the research in the area of speech recognition still very active. A detailed study on automatic speech recognition is carried out and presented in this paper that covers the architecture, speech parameterization, methodologies, characteristics, issues, databases, tools and applications.

Keywords*: Automatic Speech Recognition, Acoustic Model, Language Model, Acoustic Front-end, Decoder, Generative and Discriminative Learning, Deep Learning*

1. Introduction

Speech is the most natural form of human communication and speech processing has been one of the most exciting research areas of the signal processing [1]. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal. It is a unique discipline which encompasses a broad range and variety of technologies and applications.

The applications of speech processing are mostly useful in day to day life of people. Some of the speech processing applications are Speech Coding, Text-to-Speech Synthesis, Speech Recognition, Speaker Recognition and Verification, Speech Enhancement, Speech Segmentation and Labeling (Transcription), Language Identification, Prosody, Attitude and Emotion recognition, Audio-Visual Signal Processing and Spoken Dialog Systems.

[1] Corresponding Author

ISSN: 2005-4254 IJSIP

Speech Recognition is one of the thrust research areas in speech processing and is also known as Automatic Speech Recognition (ASR). It is the process of converting a speech signal to a sequence of words (*i.e.*, spoken words to text) by means of an algorithm implemented as a computer program [2].

2. Probability Theory of Speech Recognition

The primary goal of an ASR system is to hypothesize the most likely discrete symbol sequence out of all valid sequences in the language L, from the given acoustic input O [3]. As stated above, the input is treated as a set of discrete observations, such that:

$$O = o_1, o_2, o_3, \dots, o_t \tag{1}$$

Similarly, the symbol sequence to be recognized is defined as:

$$W = w_1, w_2, w_3, \dots, w_n \tag{2}$$

The fundamental ASR system goal can then be expressed as:

$$\widehat{W} = \text{argmax}\, P(W|O) \qquad for\ W \in \mathcal{L} \tag{3}$$

This equation implies that for a given sequence *W* and acoustic input sequence *O*, the probability *P(W|O)* needs to be determined. Bayes' theorem can be applied to this probability to arrive at the following equation:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{4}$$

The quantities on the right-hand side of the equation are easier to compute than *P(W|O)*. *P(W)* is defined as the prior probability for the sequence itself. This is calculated by using the prior knowledge of occurrences of the sequence *W*. Since the P(O) is the same for each candidate sentence W, thus equation 4 can be simplified as

$$\widehat{W} = \text{argmax}\frac{P(O|W)P(W)}{P(O)} = \text{argmax}\, P(O|W)P(W) \qquad for\ W \in \mathcal{L} \tag{5}$$

The probability *P(O|W)*, which is the likelihood of the acoustic input *O*, given the sequence *W*, is defined as the observation likelihood, can be called as acoustic score. This quantity can be determined using the Hidden Markov Model.

3. Speech Recognition Architecture

A typical speech recognition system is developed with major components that include acoustic front-end, acoustic model, lexicon, language model and decoder as shown in figure 1. Acoustic front-end takes care of converting the speech signal into appropriate features which provides useful information for recognition. The input audio waveform from a microphone is converted into a sequence of fixed-size acoustic vectors is a process called feature extraction. The parameters of word / phone models are estimated from the acoustic vectors of training data. The decoder

operates by searching through all possible word sequences to find the sequence of words that is most likely to generate. The likelihood is defined as an acoustic model *P (O|W)* and *P (W)* is determined by a language model.

The functionality of automatic speech recognition system can be described as an extraction of a number of speech parameters from the acoustic speech signal for each word or sub-word unit. The speech parameters describe the word or sub-word by their variation over time and together they build up a pattern that characterizes the word or sub-word. In a training phase the operator will read all the words of the vocabulary of the current application. The word patterns are stored and later when a word is to be recognized its pattern is compared to the stored patterns and the word that gives the best match is selected. This technique is generally referred to as pattern recognition.

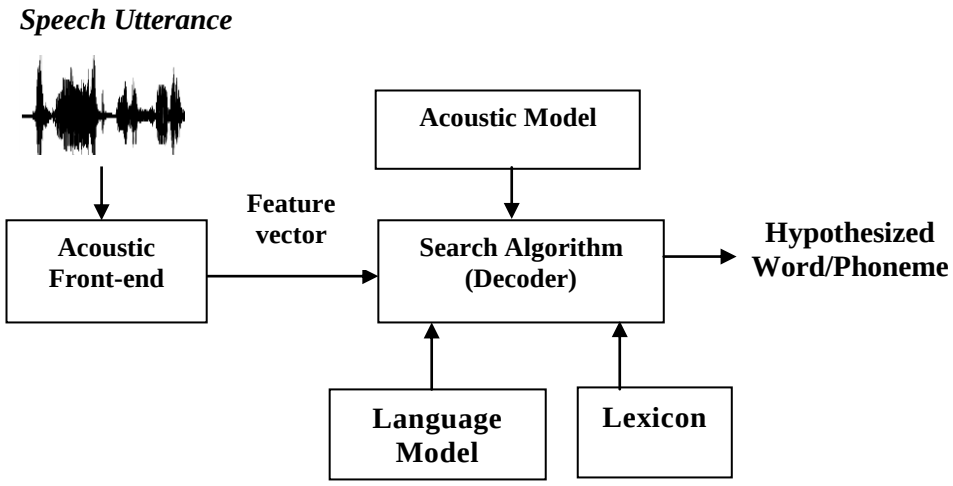

Figure 1. Speech Recognition Architecture

3.1 Acoustic Front-end

Acoustic front-end involves signal processing and feature extraction. In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal [4]. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectra temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.

There is no particular feature suitable for particular application, but the choice of features has to fulfill the following properties: they should allow an automatic system to discriminate between different through similar sounding speech sounds, they should allow for the automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environment.

In order to find some statistically relevant information from incoming data, it is important to have mechanisms for reducing the information of each segment in the audio signal into a relatively small number of parameters, or features. These

features should describe each segment in such a characteristic way that other similar segments can be grouped together by comparing their features. There are enormous interesting and exceptional ways to describe the speech signal in terms of parameters. Some of the feature extraction methods are, Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis, Mel-Frequency Scale Analysis, Filter-Bank Analysis, Mel-Frequency Cepstrum Co-efficients (MFCC), Kernal Based Feature Extraction, Dynamic Feature Extraction, Wavelet based features, Spectral Subtraction and Cepstral Mean Subtraction (CMS).

In noise robust speech recognition, many auditory-based feature extraction methods, such as zero crossing peak amplitude (ZCPA), average localized synchrony detection (ALSD), perceptual minimum variance distortionless response (PMVDR),power-normalized cepstral coefficients (PNCC), invariant integration features (IIF), amplitude modulation spectrogram, Gammatone frequency cepstral coefficients, sparse auditory reproducing kernel (SPARK), and Gabor filter bank features are effectively applied[5].

There are many feature representations in use, but the most common is the mel-frequency cepstral coefficient (MFCC) feature set. The MFCC feature extraction process has many steps which are elaborated below and the pictorial representation is given in figure 2.

- **Pre-emphasis** – This stage is used to amplify energy in the high-frequencies of the input speech signal. This allows information in these regions to be more recognizable during HMM model training and recognition.
- **Windowing** – This stage slices the input signal into discrete time segments. This is done by using a window of *N* milliseconds wide and at offsets of *M* milliseconds long. A Hamming window is commonly used to prevent edge effects associated with the sharp changes in a rectangular window.
- **Discrete Fourier Transform** – DFT is applied to the windowed speech signal, resulting in the magnitude and phase representation of the signal.
- **Mel Filter Bank** – While the resulting spectrum of the DFT contains information in each frequency, human hearing is less sensitive at frequencies above 1000 Hz. This concept also has a direct effect on performance of ASR systems; therefore, the spectrum is warped using a logarithmic Mel scale. A Mel frequency can be computed using equation 6. In order to create this effect on the DFT spectrum, a bank of filters known as triangular filters is constructed with filters distributed equally below 1000 Hz and spaced logarithmically above 1000 Hz. The output of filtering the DFT signal by each Mel filter is known as the Mel spectrum.

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \quad (6)$$

- **Log** –Taking logarithm of this provides Mel spectrum co-efficients.
- **DCT** –The final step in obtaining MFCC is performing discrete cosine transform on the Mel spectrum co-efficients. The output of DCT is Mel-cepstral coefficients of 13th order.
- **Delta MFCC Features** – In order to capture the changes in speech from frame-to-frame, the first and second derivative of the MFCC coefficients are also calculated and used [6].

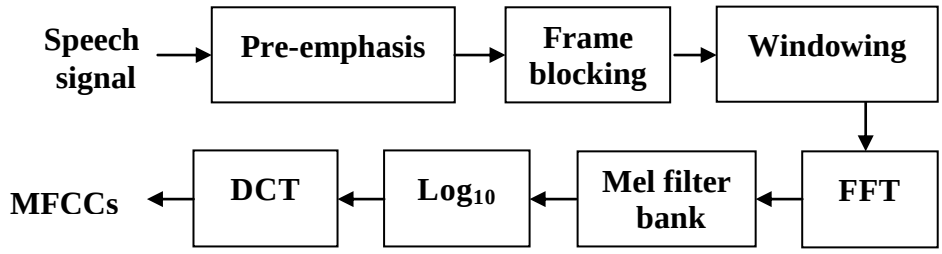

Figure 2. MFCC Feature Extraction

3.2 Acoustic Model

Acoustic model is one of the most important knowledge sources for automatic speech recognition system, which represents acoustic features for phonetic units to be recognized. In building an acoustic model, one fundamental and important issue is choosing of basic modeling units. Generally speaking, when the target language of the speech is specified, there is several types of sub word unit can be used for acoustic modeling. Different acoustic modeling unit can make a dramatic difference on the performance of the speech recognition system.

Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. Hidden Markov Model (HMM) is one of the most commonly used statistical models to build acoustic models. Other acoustic models include segmental models, super-segmental models (including hidden dynamic models), neural networks, maximum entropy models, and (hidden) conditional random fields, *etc.* An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a phoneme acoustic model is created by taking a large database of speech called a speech corpus and using special training algorithms to create statistical representations for each phoneme in a language. Each phoneme has its own HMM. The speech decoder listens for the distinct sounds spoken by a user and then looks for a matching HMM in the acoustic model. Each spoken word *w* is decomposed into a sequence of basic sounds called base phones. The acoustic model describes the probability of a specific observation given a base phone.

3.3 Language Model

A language model is a collection of constraints on the sequence of words acceptable in a given language. These constraints can be represented, for example, by the rules of a generative grammar or simply by statistics on each word pair estimated on a training corpus. Although there are words that have similar sounding phone, humans generally do not find it difficult to recognize the word. This is mainly because they know the context, and also have a fairly good idea about what words or phrases can occur in the context. Providing this context to a speech recognition system is the purpose of language model. The language model specifies what are the valid words in the language and in what sequence they can occur.

Language models are usually trained that is, the n-gram probabilities are estimated by observing sequences of words in corpora of text that contain, typically, millions of word tokens and by reducing perplexity on training data[3]. It has been observed however that reduced perplexity does not necessarily lead to better speech recognition results. Therefore algorithms that improve language models based on their effect on speech recognition are particularly appealing a language model that specifies the probability distribution of words the speaker may utter next, given a history of uttered words. Common language models are bigram and trigram models. These models contain

computed probabilities of groupings of two or three particular words in a sequence, respectively. There are tools for language modeling like CMU Statistical Language Modeling (SLM) Toolkit, Stanford Research Institute Language Modeling Toolkit.

3.4 Decoder

In the decoding stage, the task is to find the most likely word sequence W given the observation sequence O, and the acoustic-phonetic-language model. The decoding problem can be solved using dynamic programming algorithms. Rather than evaluating likelihoods of all possible model paths generating O, the focus is on finding a single path through the network yielding the best match to O. To estimate the best state sequence for the given observation sequence, the Viterbi algorithm is frequently used [7]. In the case of larger vocabulary recognition tasks, it would be challenging to consider all possible words during the recursive part of the Viterbi algorithm. To address this, a beam search can be used for Viterbi iteration, only the words with path probabilities above a threshold are considered when extending the paths to the next time step. This approach speeds up the searching process at the expense of decoding accuracy. The Viterbi algorithm assumes that each of the best paths at time t must be an extension of each of the best paths ending at time $t - 1$, which is not generally true. The path that seems to be less probable than others in the beginning may turn into being the best path for the sequence as a whole (*e.g.*, the most probable phoneme sequence does not need to correspond to the most probable word sequence). This issue is addressed by extended Viterbi and forward-backward algorithms [3].

4. Speech Recognition Methodologies

ASR methodologies are broadly classified into three approaches, namely, acoustic-phonetic approach, pattern-recognition approach and artificial intelligence approach.

Acoustic-Phonetic Approach: It is based on acoustic phonetics that postulates that there exist finite, distinctive phonetic units in spoken language. The phonetic units are characterized by a set of acoustic properties that are manifested in the speech signal, or its spectrum, over time. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling [7-8].

Pattern Recognition Approach: The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns.
The essential feature of this approach is that it uses a well-formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (*e.g.*, Hidden Markov Model) and can be applied to a sound smaller than a word, a word, or a phrase. The pattern-matching approach has become the predominant method for speech recognition in the last six decades.

Artificial Intelligence Approach: The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of acoustic phonetic and pattern recognition methods.

The main methodologies that made significant change in the speech recognition area are elaborated below. Two main approaches to pattern matching have been widely used in ASR – deterministic pattern matching based on dynamic time warping (DTW) [9], and stochastic pattern matching employing hidden Markov models (HMMs) [10]. In DTW, each class to be recognized is represented by one or several templates. Using more than one reference template per class may be preferable in order to improve the pronunciation/speaker variability modeling. During recognition, a distance between an observed speech sequence and class patterns is calculated. To eliminate the impact of the duration mismatch between test and reference patterns, stretched and warped versions of the reference patterns are also employed in the distance calculation. The recognized word corresponds to the path through the model that minimizes the accumulated distance. Increasing the number of class pattern variants and loosening warping constrains may improve DTW-based recognition performance at the expense of storage space and computational demands. In state of the art systems, HMM-based pattern matching is preferred instead of DTW due to better generalization properties and lower memory requirements.

Generative Learning Approach -HMM-GMM

In ASR, the most common generative learning approach is based on Gaussian-Mixture-Model based Hidden Markov models. Conventional speech recognition systems utilize Gaussian mixture model (GMM) based hidden Markov models (HMMs) to represent the sequential structure of speech signals. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scale, speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes. Typically, each HMM state utilizes a mixture of Gaussian to model a spectral representation of the sound wave. A GMM-HMM is parameterized by $\lambda = (A, B, \pi)$. π is a vector of state prior probabilities; A=(a_{ij}) is a state transition probability matrix; B={(b_1,…b_n}and is a set where b_j represents the Gaussian mixture model of state j. The state is typically associated with a sub-segment of a phone in speech [7-11].

State-of-the-art systems use hidden markov models to achieve good levels of performance. One of the reasons for the popularity of HMMs is that they readily handle the variable length data sequences which result from variations in word sequence, speaker rate and accent. Even though the HMM-GMM approach had become the standard tool in ASR, it has its own advantages as well as disadvantages. HMMs-based speech recognition systems can be trained automatically and are simple and computationally feasible to use. However, one of the main drawbacks of Gaussian mixture models is that they are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space.

Discriminative Learning – HMM-ANN

The paradigm of discriminative learning involves either using a discriminative model or applying discriminative training to a generative model. The use of neural networks in the form of Multilayer Perceptron (MLP) with the softmax nonlinear function at the final layer was popular in 1990's. Since the output of the MLP can be interpreted as the conditional probability [12], when the output is fed into an HMM, a good discriminative sequence model, or hybrid MLP-HMM, can be created.

Due mainly to the difficulty in learning MLPs, this line of research has been switched to a new direction where the MLP simply produces a subset of feature vectors in

combination with the traditional features for use in the generative HMM [13]. Neural networks trained by back-propagation error derivatives emerged as an attractive acoustic modeling approach for speech recognition in the late 1980s. In contrast to HMMs, neural networks make no assumptions about feature statistical properties.

Neural networks allow discriminative training in a natural and efficient manner when used to estimate the probabilities of a speech feature segment, However, in spite of their effectiveness in classifying short-time units such as individual phones and isolated words, neural networks are rarely successful for continuous recognition tasks [14-15], largely because of their lack of ability to model temporal dependencies. These kind of shallow architectures have been shown effective in solving many simple or well-constrained problems, but their limited modeling and representational power can cause difficulties when dealing with more complicated real-world applications involving human speech. Thus, one alternative approach is to use neural networks as a pre-processing *e.g.* feature transformation, dimensionality reduction for the HMM based recognition.

Deep Learning -HMM DNN

Deep learning sometimes referred as representation learning or unsupervised feature learning, is a new area of machine learning. Deep learning is becoming a mainstream technology for speech recognition and has successfully replaced Gaussian mixtures for speech recognition and feature coding at an increasingly larger scale. The first type consists of generative deep architectures, which are intended to characterize the high-order correlation properties of the data or joint statistical distributions of the visible data and their associated classes. Use of Bayes rule can turn this type of architecture into a discriminative one. Examples of this type are various forms of deep auto-encoders, deep Boltzmann machine, sum-product networks, the original form of Deep Belief Network (DBN) and its extension to the factored higher-order Boltzmann machine in its bottom layer.

The second type of deep architectures are discriminative in nature, which are intended to provide discriminative power for pattern classification and to do so by characterizing the posterior distributions of class labels conditioned on the visible data. Examples include deep-structured CRF, tandem-MLP architecture, deep convex or stacking network and its tensor version, and detection-based ASR architecture.

In the third type, or hybrid deep architectures, the goal is discrimination but this is assisted with the outcomes of generative architectures. The generative component is mostly exploited to help with discrimination as the final goal of the hybrid architecture [16][17][18].

5. Characteristics of Speech Recognition Systems

Automatic speech recognition systems are designed to solve a particular problem. There are numerous parameters that affect the design of the ASR system. There are number of issues that need to be addressed in order to define the operating range of each speech recognizing systems that is built. Some of them are, modeling units like word, syllable, phoneme used for recognition, vocabulary size like small, medium and large, task syntax like simple to complex task using N-gram language models, task perplexity, speaking mode like isolated, connected, continuous, spontaneous, speaker mode like speaker trained, adaptive, speaker independent, dependent, speaking environment as quiet room, noisy places, transducers may be high quality microphone, telephones, cell phones, array microphones, and also transmission channel [19].

.

6. Challenges in Speech Recognition

Robustness of an ASR system is the system's ability to successfully deal with different aspects of variability in the speech signal. There are a number of well-known factors that determine the accuracy of a speech-recognition system. The most noticeable ones are speaker variability, pronunciation variability, region variability, speech rate variability, context variability, channel variability and environment variability. In the design of speech recognition systems, these challenging factors must be considered and effective models to be created to provide good recognition accuracy irrespective of these variabilities [20]. In higher level, speech recognition system design requires the availability of algorithms or processes for automatic generation of word lexicons, automatic generation of language models for new tasks, automatic speech segmentation algorithms, optimal utterance verification-rejection algorithm, achieving or surpassing human performance on ASR tasks**.**

7. Applications of Speech Recognition

More recently, with the exponential growth of big data and computing power, ASR technology has advanced to the stage where more challenging applications are becoming a reality. Examples are voice search and interactions with mobile devices (*e.g.*, Siri on iPhone, Bing voice search on winPhone, and Google Now on Andriod), voice control in home entertainment systems (*e.g.*, Kinect on xBox), and various speech-centric information processing applications capitalizing on downstream processing of ASR outputs [21]. Some of these typical applications include dictation systems, voice user interfaces, voice dialling, call routing, domestic appliance control, command and control, voice enabled search, simple data entry, hands and eyes free applications and learning system for disabled people.

8. Databases and Tools for Speech Recognition

8.1 Speech Databases

There are many speech databases are available to carry out research in automatic speech recognition in American and European languages. Some of the commonly used databases are TIMIT, GlobalPhone, Aurora, Wall Street Journal, AN4, TI Digits, TI46, NTIMIT, RM1, RM2, Switch Board *etc.* The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT - Texas Instruments and Massachusetts Institute of Technology), contains recordings of phonetically-balanced prompted English speech. It was recorded using a Sennheiser close-talking microphone at 16 kHz rate with 16 bit sample resolution [22-23]. TIMIT contains a total of 6300 sentences, consisting of 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. All sentences were manually segmented at the phone level.

GlobalPhone, a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. GlobalPhone [24] was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers GlobalPhone supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages.

The first Aurora database is Aurora 2, a task of recognizing digit strings in noise and channel distorted environments. The evaluation data is artificially corrupted. The Aurora 3 task consists of noisy speech data recorded inside cars as part of the SpeechDatCar project. Although still a digit recognition task, the utterances in Aurora 3 are collected in real noisy environments [21]. The Aurora 4 task is a standard large vocabulary continuous speech recognition task which is constructed by artificially corrupting the clean data from the Wall Street Journal (WSJ) corpus. Aurora 5 was mainly developed to investigate the influence of hands free speech input on the performance of digit recognition in noisy room environments and over a cellular telephone network. The evaluation data is artificially simulated.

8.2 Speech Recognition Tools

Researchers on automatic speech recognition have several potential choices of open-source toolkits for building a recognition system. Notable among these are: HTK, Julius (both written in C), Sphinx-4 (written in Java) of the Carnegie Mellon University and Kaldi, a free, open-source toolkit for speech recognition research. Kaldi provides a speech recognition system based on finite-state transducers (using the freely available OpenFst), together with detailed documentation and scripts for building complete recognition systems [25].

Some of the other less popular open-source systems and kits are RWTH Aachen Automatic Speech Recognition System (RASR), Segmental Conditional Random Field Toolkit for Speech Recognition (SCARF), Improved ATROS (iATROS), SRI International's Decipher, idiap's Juicer and SHoUT speech recognition toolkit [26].

9. Measures of Performance

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR). The performance of the speech recognizer is measured in terms of Word Error Rate (WER) and Word Recognition Rate (WRR) [3]. Word errors are categorized into number of insertions, substitutions and deletions. Finally, the word error rate and word recognition rate are computed by the following equations.

$$\text{Word Error Rate}(\%) = \frac{\text{Insertion(I)} + \text{Substitution (S)} + \text{Deletion (D)}}{\text{No. of Reference Words (N)}} * 100 \tag{7}$$

$$\text{Word Recognition Rate (WRR)=1 - WER} = \frac{N - S - D - I}{N} \tag{8}$$

10. Conclusion

This review paper presented the speech recognition architecture, speech parameterization, methodologies, characteristics, issues, databases, tools and applications. Building automated systems to perform spoken language understanding as well as recognizing speech, as human being do is a complex task. The goal of automatic speech recognition research is to address the various issues relating to speech recognition. Robust speech recognition, Multimodal speech recognition, Multilingual speech recognition are some of the research areas gaining momentum. ASR is for languages like English, French and Czech has attained well maturity level and many developments are happening in

Chinese and Japanese. In Indian languages, there are very few ASR research is going on, it must change and many ASR tasks to be attempted in Indian languages to bring out effective interfaces in native languages to enable the people to use the technology.

References

[1] X. Huang and L. Deng, "An Overview of Modern Speech Recognition , in Handbook of Natural Language Processing", Second Edition, Chapter 15, Chapman & Hall/CRC, **(2010)**, pp. 339-366.
[2] X. Huang, A. Acero and H.-W. Hon, "Spoken Language Processing: a guide to theory, algorithm, and system development", Prentice Hall, **(2001).**
[3] D. Jurafsky and J. H. Martin, "Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Prentice Hall, **(2009).**
[4] M. A. Anusuya and S. Katti, "Front end analysis of speech recognition: a review", Int. J. Speech Technology, vol. 14, no. 2, **(2011)**, pp. 99–145.
[5] J. Li, L. Deng, Y. Gong and R.H.-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4,**(2014),** pp. 745 – 777.
[6] S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, **(1980)**, pp. 357–366.
[7] R. Lawrence and B.-H. Juang "Fundamentals of Speech Recognition", Prentice-Hall, Inc., (Engelwood, NJ), **(1993).**
[8] M. A. Anusuya and S. K. Katti, "Speech Recognition by Machine:A Review", International Journal of Computer Science and Information Security, vol. 6, no. 3, **(2009)**, pp. 181 -205.
[9] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no. 1, **(1978)** pp.43–49.
[10] J. K. Baker, "The Dragon System-An Overview',' IEEE Trans. on Acoustics Speech Signal Processing, Vol. ASSP-23, no. 1, **(1975)**, pp. 24-9.
[11] J. Bilmes, "What HMMs can do," *IEICE Trans. Inf. Syst.*, vol. E89-D,no. 3,**(2006)**, pp. 869–891.
[12] S. Renals, N. Morgan, H. Boulard, M. Cohen and H. Franco, "Connectionist probability estimators in HMM speech recognition", IEEE Trans. Speech Audio Processing , vol. 2, no. 1, **(1994)**, pp. 161–174.
[13] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki,M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard and M. Athineos, "Pushing the envelope—Aside [speech recognition]", IEEE Signal Process. Mag., vol.22, no. 5, **(2005)**, pp. 81–88.
[14] H. A. Bourlard and N. Morgan, "Connectionist Speech Recognition- A Hybrid Approach", kulwer Academic Publishers, **(1994)**.
[15] N. Smith and M. J. F. Gales, "Using SVM's and discriminative models for speech recognition", Proc. ICASSP, vol. 1, **(2002)**, pp.77 -80.
[16] D.Yu and L. Deng, "Automatic Speech Recognition - A Deep Learning Approach", Springer-Verlag London, **(2015)**.
[17] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21 no. 5, **(2013)**, pp.1060-1089.
[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal Process. Magazine, vol. 29, no. 6, **(2012)**, pp. 82-97.
[19] B. Jacob, M.M Sondhi and H.Yiteng, "Springer Handbook of Speech Processing", Springer, **(2008)**.
[20] M. Forsberg, "Why Is Speech Recognition Difficult?", Chalmers University of Technology, Citeseer, **(2003)**.
[21] J. Li, L. Deng, R. H.-Umbach and Yifan Gong, "Robust Automatic Speech Recognition: A Bridge to Practical Applications", Academic Press, **(2015).**
[22] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM", Tech. rep., U.S. Dept. of Commerce, NIST, Gaithersburg, MD, **(1993).**
[23] V. Zue, S. Seneff and J. Glass,"Speech database development at MIT: TIMIT and beyond", Speech Communication, vol. 9, no. 4, **(1990)**, pp. 351-356.
[24] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text &speech database in 20 languages, " in Proc. of ICASSP, **(2013)**.
[25] D. Povey and A. Ghoshal, "The Kaldi Speech Recognition Toolkit", Proceedings of ASRU, Hawaii US, **(2011)**, pp 1-4.
[26] C. Gaida, P. Lange, R. Petrick, P. Proba, A. Malatawy and A. D. Suendermann-Oeft,, "Comparing open-source speech recognition toolkits", DHBW Stuttgart Technical Report ,**(2014)**.

