International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

SPEECH RECOGNITION SYSTEMS – A REVIEW

Sania Iqbal[1]

[1]*Department of Information Technology, Central University of Kashmir, (India)*

ABSTRACT

We as humans speak and listen to each other in human-human interface, similar attempts have been made to develop vocally interactive computers, computer that can give speech as output (speech synthesizer), given speech as (speech recognizer). Speech recognition allows the machine to turn the speech signal into text or commands through the process of identification and understanding, and also makes the function of natural voice communication. Developing and understanding Speech Recognition systems is an inter-disciplinary activity, taking expertise in linguistics, computer science, and electrical engineering and its ultimate goal is to achieve natural language communication between man and machine. This paper provides a survey on speech recognition and discusses the techniques that enables computers to accept speech as input and shows the major developments in the field of speech recognition. This paper highlights the speech recognition techniques and provides a brief description about the four stages in which the speech recognition techniques are classified.

Keywords: artificial neural networks, filterbank, HMM, human-computer interface, mel filtering, natural language processing, speech recognition

INTRODUCTION

Speech is a primary mode of communication as it is the most natural and efficient form of exchanging information among human beings. We as humans communicate each other using speech as interface. So, it is more logical that human-computer interface is implemented as natural language speech recognition wherein speech is a medium of interaction between man and machine. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means an algorithm implemented as a computer program. The goal of speech recognition is to develop technique and system that allows the machine to turn the voice signal into the appropriate text or speech or command [1]. However, speech is a complex phenomenon as the human vocal tract and articulators, being the biological organs, are not under our conscious control. As such speech is greatly affected by accents, articulation, pronunciation, roughness, emotional state, gender, pitch, speed, volume, background noise and echoes, thus making speech recognition a cross-disciplinary field, involving wide range of principles. It is related with acoustics, phonetics, linguistics, information theory, pattern recognition theory and neurobiology disciplines.

With the rapid development of computer hardware and software and information technology, speech recognition technology is gradually becoming a key technology in the computer information processing technology.

1. Linguistics of human speech in Speech Recognition

Phonetics is the part of linguistics that focuses on the study of the sounds produced by human speech. It encompasses their production of speech by human vocal apparatus, their acoustic properties, and perception. The atomic unit of speech sound is called a *phoneme*. Words are comprised of one or more phonemes in sequence. The acoustic realization of a phoneme is called a *phone*. One major way to categorize phonemes is into vowels and consonants.

Vowels can be distinguished by two attributes. First, they are voiced sounds i.e. the airflow from the vocal chords into the mouth cavity is created by the vibration of the vocal chords at a particular pitch. Second, the tongue does not form a constriction of air flow during production. The placement of the tongue, lips, and jaw distinguishes different vowel sounds from each other. These different positions form different resonances inside the vocal tract called *formants* and the resonant frequencies of these formants characterizes the different vowel sounds.

Consonants are characterized by significant constriction of air flow in the airway or mouth. Like vowels, some consonants can be voiced, while others are unvoiced. Unvoiced phonemes do not engage the vocal cords and therefore do not have a pitch. Some consonant phonemes occur in pairs that differ only in whether they are voiced or unvoiced but are otherwise identical. For example, the sounds /b/ and /p/ have identical articulatory characteristics but the former is voiced and the latter is unvoiced.

One important aspect of phonemes is that their realization can change, depending on the surrounding phones. This is called phonetic context and it caused by a phenomenon called *coarticulation*. The process of producing these sounds in succession changes their characteristics. Modified versions of a phoneme caused by coarticulation are called allophones. Speech recognition systems use this context-dependent nature of phonemes to create a detailed model of phonemes in their various phonetic contexts.

Syntax describes how sentences can be put together given words and rules that define allowable grammatical constructs. Semantics refers to the way a meaning is attributed to the words or phrases in a sentence. Both syntax and semantics are a major part of natural language processing but neither play a major role in speech recognition.

2. Speech Recognition System

Speech Recognition System consists of vocally interactive computers i.e. computer that can give speech as output and recognize speech which is given as input. It is structured in four main blocks as depicted in Fig. 1. The signal is processed by each block in a linear chain, and its content changes gradually from an acoustic to a symbolic description of the message.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

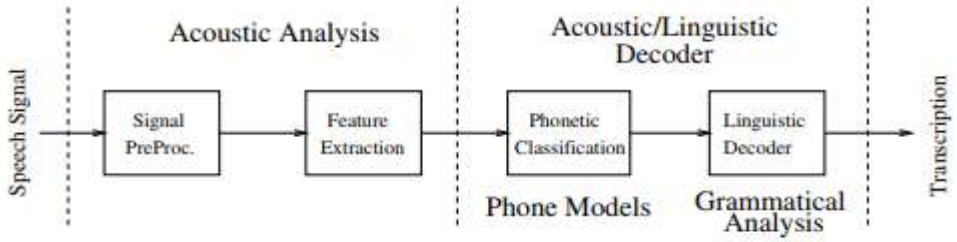

Figure 1: structure of a speech recognition system

The first two blocks, corresponding to the "Acoustic analysis" block in Fig. 1, provide the acoustic observations. These observations are vectors of coefficients that are considered to represent the speech signal characteristics. The aim of this process is not only to reduce the data flow, but also to extract information from the signal in an efficient form for the following processing (Section 2.1 & 2.2). The last two blocks provide the acoustic classification and the linguistic analysis (Section 2.4 & 2.5)

2.1 Speech Signal Processing

Speech sound waves propagate through the air and are captured by a microphone which converts the pressure wave into electrical activity which can be captured. The electrical activity is sampled to create a sequence of waveform samples that describe the signal. Speech signals have less high frequency (only up to 8000 Hz) information so a sampling rate of 16,000 Hz is typically used

Two physiological factors contribute to the characteristics of the speech waveform:

1) the excitation from the vocal chords that drives the air through the vocal tract and out the mouth and 2) the shape of the vocal tract itself when making a particular sound.

As such the speech production process is usually modeled in signal processing using a *source-filter model*. According to source filter model theory of speech production, speech sounds are produced by the action of a filter, the vocal tract, on a sound source, either the glottis or some other constriction within the vocal tract. The model assumes that source and filter properties are independent and one can be modified without affecting other. This assumption may not be strictly true in all cases but provides us with a very useful and largely accurate model of speech production.

Before recognition of speech signal, it undergoes pre-processing. When the speaker speaks, the speech includes different types of information. The information is different because of the vocal tract, the source of excitation, the behavior feature, Also, the acoustic environment and transduction equipment have an effect on the speech generated, the background noise or room reverberation along with the speech signal is completely undesirable. Speech pre-processing is intended to solve such problems. This plays an important role in eliminating the irrelevant sources of variation and noise. It ultimately improves accuracy of the speech recognition process. The speech pre-processing generally involves noise filtering, smoothing, end point detection, framing, windowing, reverberation cancelling and echo removing.

Speech signal also has information that identifies a speaker. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The information about the behavior feature can be used for speaker recognition. The speech analysis stage deals with forming suitable frame size for segmenting speech signal for further analysis and extracting [6]. The speech analysis stage can be further classified into three analyses:

a. Segmentation Analysis: Here the testing to extort the information of speaker is done by utilizing the frame size and the shift which lies in between 10 to 30 milliseconds. This analysis extracts vocal tract information of speaker.

b. Sub-segmental Analysis: Here the testing to extract the information of speaker is done by utilizing the frame size and the shift which lies between 3 to 5 milliseconds. This analysis extracts and analyses the features of excitation state.

c. Supra-segmental Analysis: Here the analysis to extract the behavior features of the speaker is done by using the frame size and the shift size that lies in between 50 to 200 milliseconds. [5]

2.2 Feature extraction for speech recognition

The speech recognition system is essentially a pattern recognition system, including feature extraction, here the source is the excitation signal generated by the vocal chords that passes through the vocal tract, modeled as a time-varying linear filter. Since the phoneme classification is largely dependent on the vocal tract shape, and hence, the filter portion of the source-filter model. The excitation or source signal is largely ignored or discarded. Thus, feature extraction process for speech recognition is largely designed for capturing the time-varying filter shapes over the course of an utterance. The feature extraction is needed because the raw speech signal contains information besides the linguistic message and has a high dimensionality, rendering the raw speech signal unfeasible for the classification of sounds and as such will result in a high word error rate. Thus, the feature extraction algorithm aims to derive a characteristic feature vector with a lower dimensionality, which is used for the classification of sounds. [2,3]. Broadly the feature extraction techniques can be classified as temporal analysis and spectral analysis technique [4]. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

Steps involved before actual feature extraction are:

Step 1: Windowing

Speech Waveform is a *non-stationary* signal, i.e. its statistical properties change over time. So the speech signal is examined in chunks called windows or frames that are small enough such that the speech can be assumed to be stationary within those windows. In the process of windowing the disruptions which are present at the start as well as at the end of the frame are minimized. Thus, the analysis is done on a series of short, overlapping frames of audio. In speech recognition, we typically use windows of length 0.025 sec (25 ms) with an overlap of 0.01 (10 ms). This corresponds to a frame rate of 100 frames per second.

Step 2: Discrete Fourier Transformer

Each frame of data is transformed into the frequency-domain using a discrete Fourier transform, which converts the windowed frames into magnitude spectrum. The Fourier transform represents both the spectral magnitude (absolute amplitude) and phase for each frame and frequency. For feature extraction purposes, the phase information is of no use, only the magnitude is considered.

Step 3: Mel filtering

To remove variability in the spectrogram caused by the harmonic structure in the voiced regions and the random noise in the unvoiced regions, spectral smoothing operation is performed on the magnitude spectrum. A filterbank is applied which is based on the processing done by the auditory system. The filterbank applies an approximately logarithmic scale to the frequency axis. That is, the filters become wider and farther apart as frequency increases. The most common filterbank used for feature extraction is the Mel filterbank. A Mel filterbank of 40 filters is shown in Fig. 2. Each filter will average the power spectrogram across a different frequency range.

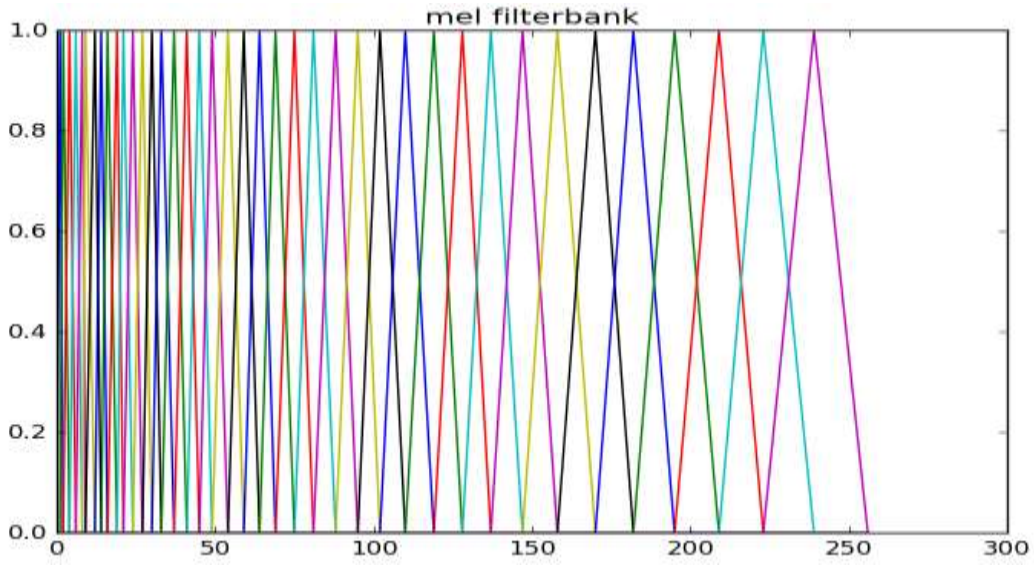

Figure 2: mel filterbank with 40 filters

Step 4: Log compression

In this step a logarithm operation is applied to compress the dynamic range of the signals. It closely models a nonlinear compression operation that occurs in the auditory system. The output of this logarithm operation is referred to as *"filterbank" coefficients*. Compared to the original spectrogram of speech, the filterbank coefficients are a much smoother version, where both the high frequency noise variability and pitch/harmonic structure are removed.

Other pre-processing steps can also be applied prior to feature extraction based on requirement. These include: 1) Dithering: adding a very small amount of noise to the signal to prevent mathematical errors during feature computation 2) Pre-emphasis: applying a high pass filter to the signal prior to feature extraction to counteract the

fact that the voiced speech at the lower frequencies has much high energy than the unvoiced speech at high frequencies, it is performed with a simple linear filter.

Step 6: Feature normalization

Chances are that the communication channel will introduce some bias (constant filtering) on the captured speech signal, resulting in variations in signal gain which therefore cause differences in the computed filterbank coefficients even though the underlying signals represent the same speech. These channel effects can be modeled as a convolution in time, which is equivalent to elementwise multiplication in the frequency domain representation of the signal. Thus, we model the channel effects as a constant filter. For convenience normalization is performed on filterbank features directly, after the log operation.

Step 7: Mel Cepstrum

This filterbank spectrum is then passed to inverse of discrete Fourier transform which produces the final result as *Mel-Cepstrum*. The Mel-Cepstrum consists of the features that are required for speech identification.

Few feature extraction techniques and their comparison are given in the TABLE below.

Table: Comparison of various feature extraction techniques used in speech recognition

Technique	Characteristics	Advantages	Disadvantages
a) Linear Predictive coding	Provides autoregression based speech features. [8] Formant estimation and static technique. [5] Residual sound is very close to the vocal tract input signal. [7]	Reliable, accurate and robust technique for providing parameters which describe the time varying linear system to represent the vocal tract. [9] Good computation speed and accurate parameters of speech Useful for encoding speech at low bit rate.	Unable to distinguish the words with similar vowel sounds [10]. Assumes that signals are stationary and hence is not able to analyze the local speech events accurately. Generates residual error as output
b) Mel-frequency cepstrum (MFFCs)	Used for speech processing tasks. Mimics the human auditory system	The recognition accuracy is high i.e. performance rate is high. Captures main characteristics of phones in speech. Low Complexity.	In case of background noises, it does not give accurate results. [10] The filter bandwidth is not an independent design parameter Performance is affected by the number of filters. [12]
c) RelAtive SpecTrAl (RASTA Filtering)	It is a band pass filtering technique Designed to lessen impact of	Removes the slow varying environmental variations as well as the fast variations in	Causes a minor deprivation in performance for the clean information but it also slashes

	noise as well as enhance speech. Widely used for the speech signals that have background noise or simply noisy speech.	artefacts. [14] Independent of the choice of microphone or the position of the microphone to the mouth, hence it is robust. [15] Captures frequencies with low modulations that correspond to speech. [16]	the error in half for the filtered case. [15] RASTA combined with PLP gives a better performance ratio. [16]
d) Probabilistic Linear Discriminate Analysis (PLDA)	Based on i-vector extraction. The i-vector is one which is full of information and is a low dimensional vector having fixed length. Uses the state dependent variables of HMM.	Is a flexible acoustic model which makes use of variable number of interrelated input frames without any need of covariance modelling. [17] High recognition accuracy	Based on the Gaussian assumption which are on the class conditional distributions. The generative model is also a disadvantage. The objective was to fit the date which takes class discrimination into account. [18]

2.3. Modelling: Fundamental equation of speech recognition

Speech recognition is cast as a statistical optimization problem. For a given sequence of observations O = {O_1, … O_N}, the most likely word sequence W = {W_1, … W_M} is sought, i.e. the word sequence which maximizes the posterior probability P(W|O).

Mathematically, this is expressed as:

$$\widehat{W} = argmax_W P(W|O)$$

Solving this expression using Bayes rule

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

Because the word sequence is independent of the marginal probability of the observation P(O), it can be ignored. Thus, the expression becomes:

$$\widehat{W} = argmax_W P(W|O)_{P(W)}$$

This is known as the *fundamental equation of speech recognition*. The speech recognition problem can now be considered as a search over this joint model for the best word sequence. The component P(*O*|*W*) is known as an **acoustic model**, that describes the distribution over acoustic observations *O* given the word sequence *W*. It models how sequences of words are converted into acoustic realizations, and then into the acoustic observations presented to the speech recognition system. The component P(*W*) is called a **language model** based solely on the

word sequence *W*. The language model assigns a probability to every possible word sequence and trained on sequences of words that are expected to be like those the system will encounter.

2.4 Acoustic Modeling

An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these representations is assigned a label called a phoneme. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 40 different phonemes. An 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. It is a hybrid model which uses deep neural networks to create frame-level predictions and then a hidden Markov model to transform these into a sequential prediction.

2.4.1 Hidden Markov Model

Hidden Markov models are used to model the acoustic observations (feature vectors) at the subword level, such as phonemes. Each phoneme is modeled with 3 states (in speech recognition), to separately model the beginning, middle and end of the phoneme. Each state has a self-transition and a transition to the next state. Word HMMs can be formed by concatenating its constituent phoneme HMMs. Thus, a high-quality pronunciation dictionary which "spells" each word in the system by its phonemes is formed. For example, the HMM word "cup" can be formed by concatenating the HMMs for its three phonemes.

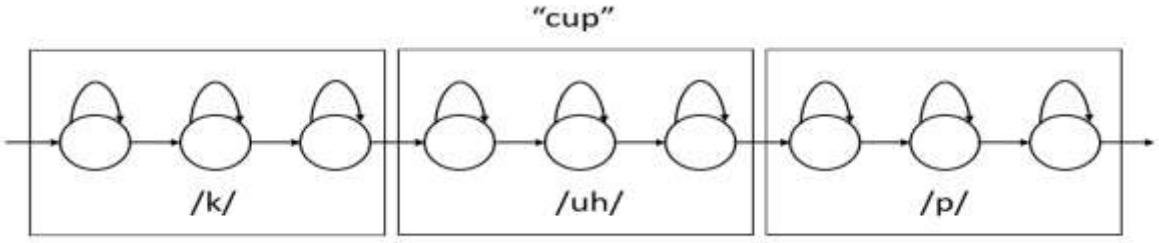

Figure 3: HMM word.

In a *hidden* Markov model each state is defined by a probability distribution over events or observations. This makes the model *doubly stochastic*. The transitions between states are probabilistic and so are the observations in the states themselves. There are three fundamental problems for hidden Markov models each with well-known solutions: 1) The Evaluation Problem (the probability that these observations were generated by the model, solved by *forward algorithm*), 2) The Decoding Problem (the most likely sequence of states that can explain the observations, solved using *Viterbi algorithm), 3)* The Training problem (how can we adjust the model parameters, solved using the *Baum-Welch* algorithm).

Earlier each state in the HMM had a probability distribution defined by a Gaussian Mixture Model (GMM), thus, each state of the model has its own GMM. Now speech recognition systems use a single deep neural network that has output labels that represent the state labels of all HMMs states of all phonemes. Such acoustic models are called

"hybrid" systems or DNN-HMM systems. This acoustic- phonetic approach involves pooling data associated with multiple context-dependent states that have similar properties and combining them into a single "tied" or "shared" HMM state. This tied state, known as a *senone*, is then used to compute the acoustic model scores for all of the original HMM states whose data was pooled to create it. Grouping a set of context-dependent triphone states into a collection of senones is performed using a *decision-tree clustering* process. A decision tree is constructed for every state of every context-independent phone.

2.4.2 Deep Neural Network Acoustic Models

One of the most significant advances in speech recognition in recent years is the use of deep neural network acoustic models. [19] The most common objective function used for training neural networks for classification tasks is *cross entropy*. For a *M*-way multi-class classification task such as senone classification, the objective function for a single sample can be written as:

$$E = -\sum_{i=1}^{M} t_m \log y_m$$

Where t_m is the label (1 if the data is from class *m* and 0 otherwise) and y_m is the output of the network which is a softmax layer over the output activations. Thus, for each frame M-dimensional 1-hot vector is generated, that consists of all zeros except for a single 1 corresponding to the true label i.e. every frame of every utterance is assigned to a senone in order to generate these labels. To label all frames of the training data with a corresponding senone label, a process known as *forced alignment* is used, wherein HMM decoding is performed but constrain search to be done along all paths that will produce the correct reference transcription. Forced alignment generates the single most-likely path, and thus, the senone label for every frame in the utterance.

2.4.3 Recurrent Neural Networks

Another class of neural network is a recurrent neural network used in acoustic modelling which unlike feedforward DNNs, process data as a sequence and have a temporal dependency between the weights. Also, in contrast to a feedforward layer, a recurrent layer's output has dependence on both the current input and the output from the previous time step. In applications, where latency is not a concern, it is possible to perform the recurrence in both the forward and backward directions. These networks are *bidirectional* neural networks, where each layer has a set of parameters to process the sequence forward in time and a separate set of parameters to process the sequence in reverse. These two outputs are concatenated to form the input to the next layer. RNNs are appealing for acoustic modeling because they can learn the temporal patterns in the feature vector sequences, which is very important for speech signals. However, in order to train RNNs, the sequential nature of the training sequences must be preserved. Thus, rather than frame-based randomization performed in feedforward networks, *utterance-based randomization*, where the ordering the utterances is randomized but the sequential nature of the utterances themselves is preserved is performed. [11]

2.5 Language Modelling

Language model is the component of a speech recognition system that estimates the prior probabilities P(W) of possible spoken utterances. These prior probabilities are combined with the acoustic model likelihoods P(O|W) in the Fundamental Equation of Speech Recognition to arrive at the overall best hypothesis. Thus, the language model (or LM) embodies the recognizer's knowledge of what probable word sequences are, even before it has heard any actual speech sounds. The LM should assign high probabilities to likely utterances and low probabilities to unlikely ones, without ruling anything out completely. Furthermore, the assignment of probabilities is not done by estimating a parameterized model from data, thus allowing the statistics of actually observed training data determine what words are likely to be heard in a language, scenario, or application. In language modeling *sentence* is the word sequence corresponding to an entire speech utterance and can be anything a speaker would utter in the context of a speech application.

2.5.1 Vocabulary

The *vocabulary* of the LM consists of a finite set of words, which is also the vocabulary of the speech recognizer. A word that is not recognized, is not considered possible by the LM (i.e., it's probability would be zero). Words outside the vocabulary are called *out-of-vocabulary words*, or *OOVs*. When an OOV occurs in the input data, it will incur (at least) one-word recognition error, thus it is important to choose the vocabulary so as to minimize the chances of OOVs. General strategy is to pick the words that have the highest prior probability of occurring, i.e. we choose the most frequently occurring words in a corpus of training data. For example, we can pick the N most frequent words, or all words occurring more than K times in the data. An optimal vocabulary size is essential, as it represents a good tradeoff between recognizer speed since, a larger vocabulary means more computation in recognition and accuracy as by reducing OOVs but adding very rare words will have negligible effect on accuracy, and might reduce accuracy, due to search errors and greater acoustic confusability within the vocabulary.

2.5.1 Markov factorization and N-grams

Even with a finite vocabulary, probable set of word sequences is infinite, so making it impossible to parameterize the LM by listing the probability of every possible sentence. Borrowing the concepts from acoustic modeling, the idea is to use the chain rule to factor the sentence probability into a product of *conditional word probabilities*, and then apply the Markov assumption to limit the number of states, and thereby, parameters. The Markov state of the LM is called the *context* or *history*.

$$P(W)=P(w_1) \times P(w_2|w_1) \times P(w_3|w_1w_2) \times \ldots \times P(w_n|w_1 \ldots w_{n-1})$$

Thus, each word is predicted by only the preceding ones, for example, if the Markov state is limited to one word (first-order Markov model) each word is predicted by only the immediately preceding one.

$$P(W)=P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times \ldots \times P(w_n|w_{n-1})$$

Such a model is a *bigram* model, in language modeling this model is not usually, because it uses only statistics of two adjacent words at a time. Similarly, a second-order Markov model is called a *trigram model*, as it predicts each word based on the preceding two. The generalization of this scheme is the *N-gram model*, i.e. each word is

conditioned on the previous N−1 words. However noticeable improvement is seen when going from bigrams to trigrams, but little improvement as N is increased further. Therefore, LMs beyond 4-grams and 5-grams are rarely used.

2.5.3 Class-based LMs

Major drawback of N-gram model is that all words are treated as completely distinct, consequently, the model needs to see a word sufficiently many times in the training data to learn N-grams it typically appears in. Class-based language models exploit the concept of *word similarity* i.e. words share many properties, both syntactically and meaning-wise, it therefore groups (some) words into *word classes*, and then collect N-gram statistics involving the class labels instead of the words. The probability of the pure *word* string is expressed as a product of two components- the probability of the string containing the class labels, multiplied by the *class membership probabilities*. The membership probabilities are estimated from data or set to a uniform distribution. There are two basic approaches to come up with word classes. One involves prior knowledge, typically from an application domain, even though the training data is unlikely to have usage samples of all the possible instantiations. The other way to define word classes is in a purely data-driven way, without human or domain knowledge. It involves searching the space of all possible word/class mappings and pick one that maximizes the likelihood of the resulting class-based model on the training data.

2.5.4 Neural Network LMs

Neural network-based machine learning methods have taken over in many areas, including in acoustic modeling for speech recognition. Similarly, artificial neural networks (ANNs) have also been devised for language modeling, and, given sufficient training data, give superior performance compared to N-gram methods. Much of the success of ANNs in language modeling stems from overcoming two specific limitations of N-gram models. The first limitation is the lack of generalization across words. The ANN feedforward language model, addressed the word generalization problem by including a *word embedding* layer that maps the discrete word labels upon input to a dense vector space. In other words, context words that affect the next word similarly, will be encoded as nearby points in space, and the network can then exploit this similarity when encountering the words in novel combinations. The second limitation of N-grams was the truncation of the context, which so far was limited to the previous N−1 words. This is a problem because language allows embedded clauses, arbitrarily long lists of adjectives, and other constructs that can put arbitrary distance between related words that would be useful in next-word prediction. Any reasonable value of N would be insufficient to capture all predictive words in a context. This limitation is overcome in recurrent networks, which feed the activations of a hidden layer at time $t-1$ as extra inputs to the next processing step at time t, as shown in the Fig. 4:

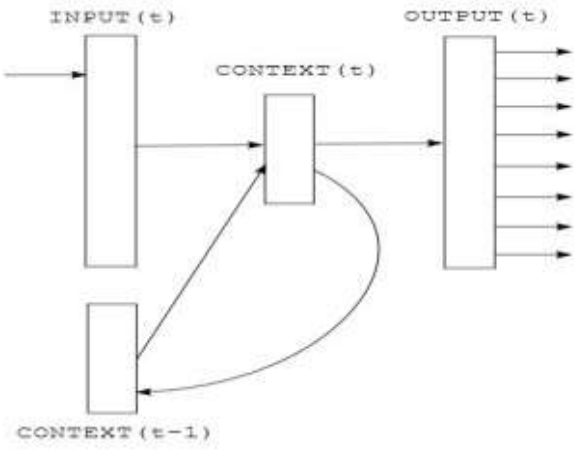

Figure 4: recurrent ANN-LM

This allows the network to pass information from one word position to the next, repeatedly, without a hard limit on how far back in time information originates and can be used to predict the current next word. However, there are practical issues with the trainability of such recurrent networks because the mathematical rules governing ANN activations lead to an exponential dilution of information over time. (and can been solved with mechanisms to gain information flow from one time step to the next).

III CONCLUSION

Speech recognition represents one of the most important techniques to endow a machine with simulated intelligence to recognize user or user-voiced commands and to facilitate human interface with the machine. It also represents a key technique for human speech understanding and is becoming increasingly important for safety reasons. For example, SR may be used to replace the manual task of punching text. In this paper we discussed the present scenario of speech recognition systems- the four stages of speech analysis and identification. We analyzed various feature extraction techniques applied. Accuracy being a key issue in speech recognition, we conclude the Mel frequency cepstrum is a feature extraction technique that is able to mimic the human auditory system and gives a better performance rate. ANN are one of the promising field for the future speech recognition systems, overcoming the conventional methods of modelling and are proving to be useful in speech signal classification. From the different types of ANN's discussed, RNN have achieved better speech recognition rates, but the training algorithm is more complex and dynamically sensitive, which can cause issues. Both acoustic and language model have benefited from general improvements in ANN technology, such as deeper stacking of network layers 'deep learning' and better training methods. The methods in language modeling are constantly evolving, and research in this area is very active. But still the perplexity of state-of-the-art LMs is still two to three times worse than the predictive powers of humans. So, for speech recognition systems to be widely used we still have a lot of areas for improvement.

There has been a lot of research in the field of speech recognition but still the systems developed so far have limitations: there are a limited number of vocabularies in the current systems and work needs to be done towards expanding this vocabulary, there is problem of overlapping speech that is the systems cannot identify speech from multiple users, also the user needs to be in a place such that background is noise free for an accurate recognition, there are issues with the accent and the pronunciation of the user or speaker. However, it is foreseeable in the near future that, with as the artificial intelligence and machine learning technology continues to progress, the speech recognition system will be more in-depth, the application of speech recognition systems will be more extensive [13]

REFERENCES

[1] R. Klevansand R. Rodman, *Voice Recognition, Artech House, Boston, London 1997.*

[2] D. O'Shaughnesssy, *Automatic speech recognition: History, methods and challenges Pattern Recognition, vol. 41, no. 10,* pp. 2965–2979, 2008.

[3] H. Niemann, Klassifikation von Mustern, 2nd ed., Berlin, New York, Tokyo: Springer, 2003.

[4] J. W. Picone, *Signal modelling technique in speech recognition*, Proc. Of the *IEEE, vol. 81, no.9*, pp. 1215-1247, Sep. 1993

[5] Santosh K.Gaikwad and Pravin Yannawar, *A Review on Speech Recognition Technique,* A Review, International Journal of Computer Applications, *Volume 10– No.3*, November 2010

[6] Gin-Der Wu amd Ying Lei, *A Register Array based Low power FFT Processor for speech recognition*, Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan

[7] Celso Auguiar, *Modelling the Excitation Function to Improve Quality in LPC's Resynthesis*, CCRMA - Center for Computer Research in Music and Acoustics. Stanford University

[8] Tomyslav Sledevic,˘ Artu¯ras Serackis, Gintautas Tamulevici˘ us, Dalius Navakauskas, *Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System*, International Journal of Electrical, Computer, Electronics and Communication *Vol:7 No:12*, 2013

[9] Shanthi Therese Chelpa Lingam, *A Review of Feature Extraction Techniques in Automatic Speech Recognition,* International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) , *Volume No.2, Issue No.6*, pp : 479-484 1 June 2013

[10] Navnath S Nehel and Raghunath S Holambe, *DWT and LPC based feature extraction methods for isolated word recognition*, Journal on Audio, Speech, and Music Processing, 2012

[11] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, *Neural Network used for Speech Recognition,* Journal of Automatic Control, University of Belgrade, *Vol. 20*, pp. 1-7, 2010.

[12] Nilu Singh, R.A Khan and Raj Shree, *A Comparative Study on MFCC and Prosodic Feature Extraction Techniques*, International Journal of Computer Applications (0975 – 8887) , *Volume 54– No.1*, September 2012

[13] Yangshang Guo, Yang Jinlong. *The speech recognition technology overview*, 2006.

[14] Chia-Ping Chen Jeff Bilmes and Daniel P. W. Ellis, *WA Speech Feature Smoothing for Robust ASR,* Department of Electrical Engineering University of Washington Seattle

[15] H. Hermansky and N. Morgan, *Rasta processing of speech*, IEEE Trans. on Speech and Audio Processing, *vol. 2, no. 4*, pp. 578{589, 1994.

[16] Yuxuan Wang, Kun Han, and DeLiang Wang, Fellow, *Exploring Monaural Features for Classification-Based Speech Segregation*, IEEE Trans. On audio, speech and language processing, 2012.

[17] Liang Lu, Member, IEEE and Steve Renals, Fellow, IEEE, *Probabilistic Linear Discriminant Analysis for Acoustic Modelling*, IEEE signal processing letters, 2014

[18] Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng and Frank Soong, *-vectors, a Discriminative Scoring for Speaker Recognition Based*, Center for Speaker and Language Technologies (CSLT), 2014

[19] Sonali B. Maind, Priyanka Wankar, *Research Paper on Basic of Artificial Neural Network*, International Journal on Recent & Innovation Trends in Computing & Communication, *Vol. 1, Issue 1*, pp. 96-100.