**International Journal of Current Advanced Research**
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: SJIF: 5.995
Available Online at www.journalijcar.org
Volume 6; Issue 9; September 2017; Page No. 6287-6297
DOI: http://dx.doi.org/10.24327/ijcar.2017.6297.0912

Research Article

# A SURVEY ON AUTOMATIC SPEECH RECOGNITION SYSTEM

**Sundarapandiyan S[1*] and Shanthi N[2]**

Department of Computer Science and Engineering, Kongu Engineering College, Erode

**ARTICLE INFO**

***Article History:***

Received 12th June, 2017
Received in revised form 3rd July, 2017 Accepted 24th August, 2017
Published online 28th September, 2017

***Key words:***

Automatic Speech Recognition, Feature Extraction, Acoustic Model, Language Model.

**ABSTRACT**

Speech is a primary mode of communication among human beings. It is natural for people to expect to be able to carry out spoken dialogue with computers. In this paper we discussed the fundamental approach and development of speech recognition in the last several year of research in Automatic Speech Recognition (ASR). The design of Speech Recognition system requires careful attentions to the following issues: Various type of speech class, Feature Extraction, Acoustic model, Pronunciation Dictionary and language model. We presented the various techniques to solve this problem existing in ASR. This paper is helpful for to review the problem in ASR research in various Speech recognition models.



## INTRODUCTION

Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means of pattern recognition algorithm. Speech is the preferred and most convenient means of conveying information. The advantage of verbal communication has become even stronger today due to convergence of computers and telecommunication systems which allows people to access information on computers located remotely. For a reason it is clear human-computer interaction is important for well formed communication. Alexander Graham Bell was the first person who 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.

Since the 1950s computer scientists have been researching ways and means to make computers able to record interpret and understand human speech. The goal of ASR is to achieve the 100% accuracy with independent of speaker, vocabulary size and environment. But last several year of research in this area can only achieve the above 90% accuracy. In these paper we discuss about the various ASR system model and there problem and Feature. In recent days Speech recognition is integrated with numerous real world applications such as telecommunications, Health care, Military, Robotics, Telecommunications, Mobile Applications, Scan soft, [2] and Robertson, [3].

**Corresponding author:* **Sundarapandiyan S**
Department of Computer Science and Engineering, Kongu Engineering College, Erode

### *Classification of ASR System*

ASR System classified into the following class depending upon the type of utterance, type of speaker, type of vocabulary .Figure 1 shows the classification of ASR System.

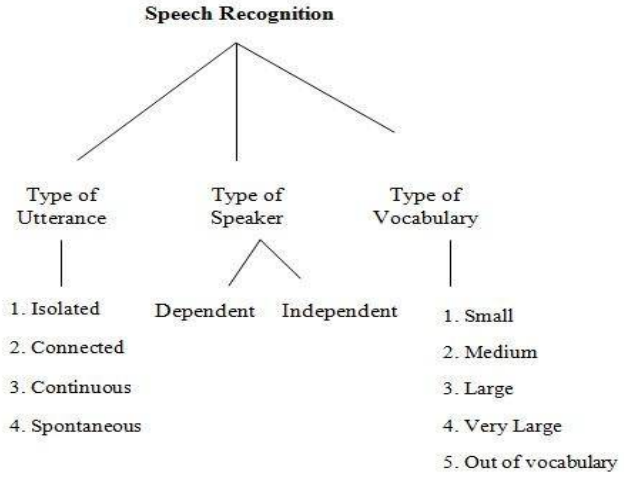


**Figure 1** Classification of ASR System

### *Type of Utterance*

Utterance is a speaking of word. This utterance may be sub-word, word or word sequence. It is classified into isolated words, connected words, continuous speech, and spontaneous speech.

1. Isolated word: Words spoken with pause (typically in

duration in excess of 200ms) before and after each word. It doesn't mean that it accepts single words, but does require a single utterance at a time. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced, which are the major advantages of this type.

2. Connected Words: Words Spoken Carefully but no explicit pause between them. Connected word systems are similar to isolated words but allow separate utterance to be run-together.
3. Continuous Speech: Words Spoken fluently as in the conversational Speech. Recognizers with continuous speech capabilities are some of the most difficult job to create because they utilize special methods to determine utterance boundaries.
4. Spontaneous Speech: This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features, such as words being run together, "ums" and "ahs" and even slight stutters.

***Types of Speaker Model***

All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into main categories based on speaker models, namely, speaker dependent and speaker independent.

1. Speaker dependent models-Designed for specific speaker, more accurate.
2. Speaker Independent Model-Speaker independent system are designed for variety of speakers. It recognizes the speech patterns of a large group of people.

***Types of Vocabulary***

Size of vocabulary affects the accuracy of ASR System. Some ASR system require few words only (Digit) some require large vocabulary (Spontaneous speech). In ASR systems the types of vocabularies can be classified as follows.

1. Small vocabulary - ten of words
2. Medium vocabulary - hundreds of words
3. Large vocabulary – thousands of words
4. Very-large vocabulary – tens of thousands of words
5. Out-of-Vocabulary – Mapping a word from the vocabulary into the unknown word

Apart from the above characteristics, the environment variability, channel variability, speaker style, sex, age, speed of speech also make the ASR system more complex. But the efficient ASR systems must cope with the variability in the signal.

***ASR System Overview***

The goal of speech recognition can be formulated as follows: for a given acoustic observation $X = X1, X2.....Xn$ find the corresponding sequence of words $W = w1, w2....wm$ with maximum a posteriori probability, Using Bayes' decision rule, this can be expressed as:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \quad (1)$$

Where

$P(W|X)$ - Maximum Posterior Probability

$\frac{P(X|W)}{P(X)}$ - Emission probability estimated from Acoustic model,

$P(W)$ - Prior probability estimated from language model.

Figure 2 shows the architecture of ASR system. The sound input is taken from the sound recorder and is feed to the feature extraction module. The feature extraction module generates feature vectors out of it which are then forwarded to the Decoder. The Decoder with the help of knowledge base such as the Acoustic model, Dictionary and Language model to search the most likely sequence of words. The words are considered as a recognized output. From Figure2 the ASR system consist of four basic components which are

- Feature Extraction
- Acoustic model
- Dictionary
- Language Model

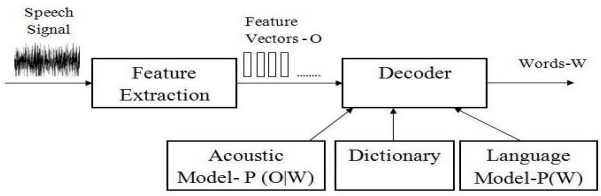


**Figure 2** Architecture of Automatic Speech Recognition

***Feature Extraction***

In the feature extraction phase speech signal is converted into a sequence of feature vectors based on spectral and temporal measurements. Typically, in speech recognition, we divide the speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors by applying feature extraction technique such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Predictive (PLP) and Rasta PLP(RPLP) .Then this Feature vectors are converted into low dimensional acoustic vectors by applying Dimensionality reduction technique such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS) and Linear Discriminant Analysis(LDA) and Locally-Linear Embedding (LLE) then these vectors are transferred to the classification stage. This Feature Extraction method removes the noise in the speech signal and also removes the redundant data. Table1 shows the various feature extraction, dimensionality reduction technique and their properties.

***Feature Extraction Technique***

Feature Extraction is the most important part of Speech recognition. Because every speech has different individual characteristics embedded in utterances. Several techniques was proposed for feature extraction such as MFCC, PLP, LPC ,RPLP.MFCC was introduced by Davis and Mermelstein [4] and became the standard front-end, much effort has been taken to improve its efficiency in real-world environment. Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum. This feature is well known in the field of speech recognition and it is used for variety of ASR application [5] [6] [7].

PLP is widely used in speech recognition systems as a feature extraction method. PLP similar to LPC analysis is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, PLP modifies the short-term spectrum of the speech by several psychophysically based transformations [8] [9]. LPC [10][11] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. RPLP was obtained from combination of MFCC and PLP. The objective of modeling technique is to generate speaker models using speaker-specific feature vectors. Such models will have enhanced speaker-specific information at reduced data rate [12]. Among these the most commonly used cepstral coefficients are MFCCs and LPCCs, because of less intra-speaker variability and also availability of spectral analysis tools .Figure 3 shows the extraction of MFCC Features.

**Table 1** Feature Extraction Techniques and their properties

| Techniques | Properties |
|---|---|
| MFCC | MFCCs are coefficients that collectively make up an MFC. MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform. |
| LPC | LPC is to predict the current value of the signal using a Linear combination of previous samples. |
| PLP | PLP analysis is computationally efficient and yields low dimensional representation of speech. |
| RPLP | Combination of MFCC and PLP. |
| PCA | Linear transformation, convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. |
| MDS | Non-linear transformation, It is used for exploring similarities or dissimilarities in data. |
| LLE | Including faster optimization when implemented to take advantage of sparse matrix algorithms, and better results with many problems. |
| LDA | Linear Transformation, Supervised algorithm, LDA easily handles the case where the within-class frequencies are unequal. |

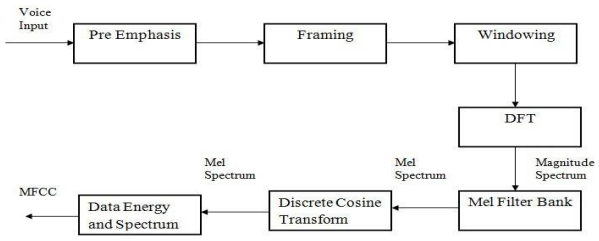


**Figure 3** Extracting the MFCC Features

### *Dimensionality Reduction Technique*

Dimensionality Reduction Technique converts the high dimensional acoustic vector into the low dimensions. There are a couple of benefits to using the dimensionality reduction technique. First of all, it can dramatically reduce the word error rate. Second, it also makes the decoder faster since it reduces the dimensionality of the features, and also reduces the size of the acoustic model. Many dimensionality reduction methods have appeared which is Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS) and Linear Discriminant Analysis (LDA) and Locally-Linear Embedding (LLE).

PCA to map the variance of the speech material in a database into a low-dimensional space, followed by clustering and a selection technique [13]. A unified algorithmic frame work for solving many variants of MDS [14].LLE was presented including faster optimization when im- plemented to take advantage of sparse matrix algorithms [15].The largest improvements in speech recognition could be obtained when the classes for the LDA transform were defined to be sub-phone units [16].

### *Acoustic Model*

The Acoustic Model module provides a mapping between units of speech incoming features provided by the Frontend. Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. Acoustic model classified into two main categories 1) Generative Model 2) Discriminative Model .Generative model learns the joint probability distribution of observed acoustic features and the corresponding speech class. Generative speech recognizers such as those based on Gaussian Mixture Models, Hidden Markov Model, and Stochastic Segment Models .Discriminative model learns the conditional probability distribution of observed acoustic features and the corresponding speech class.

Discriminative speech recognizers, such as those based on Maximum Entropy models, Neural networks, and Conditional random fields. Generative model converted into discriminative model by applying the base rule. Table 2 shows the techniques of Acoustic model and their properties

### *Gaussian Mixture Model*

In 1995 GMM was successfully applied to the speech recognition system [19][20].GMM has been widely used in statistical speaker recognition [21][22][23].GMM are interpreted to represent the broad acoustic classes. It Provide a smooth approximation to the underlying long term sample distribution of observations obtained from utterances by a given speaker. The Gaussian model is a probability density function. Parameters of the GMM are mean, standard deviation and component weights. GMM parameters are derived from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model [1]. A Gaussian mixture model is a weighted sum of M component Gaussian densities, is given by the equation,

$$P(X|\lambda) = \sum_{i=1}^{M} w_i \, g(x|\mu_i, \sum i) \qquad (2)$$

Where
$g(x|\mu i,)$ - Gaussian function,
$\mu i$ −Mean,
$wi, i = 1 \ldots\ldots M$ are the mixture weights,
$\sum i$ −Co variance,
$\lambda$–Gaussian mixture model Parameter, It contains mean vectors, covariance matrices and mixture weights.

**Table 2** Acoustic model Techniques and their properties

| Techniques | Properties |
|---|---|
| Hidden Markov Model | A HMM is an extension of a Markov chain in which the input symbols are not same as the states. Initial probability Distribution, Transition probability, Observation probability are the elements of HMM |
| GMM | A Gaussian is a probability density function; It is parameterized by mean and variance. |
| Stochastic Segment Model | It is a better frame work for modeling the dynamics of the Speech production mechanism. Stochastic segment model which provides a joint Gaussian model for a sequence of observation. |
| Maximum entropy direct Model | This model attempts to model the posterior probability directly, make decoding simpler. Asynchronous and overlapping feature can be incorporated formally. |
| Support Vector Machine | The SVMs are effective discriminant classifiers capable of maximizing the error margin. It is capable of to deal with samples of very high dimensionality. |
| Dynamic Time Warpping | DTW is used to compute the best possible alignment warp, between test and reference pattern. It allows system to find an optimal match between two given sequence. |
| Artificial Neural Network | ANN applied to the speech recognition in the form of Hybrid ANN/HMM. The goal Hybrid systems for ASR to take the advantage from the properties of both HMMs and ANNs. Three major class of neural network were proposed namely Time Delay Neural Network, Multi Layer Perceptron, and Recurrent Neural Network. |
| Deep Belief Network | DBNs are MLP But DBN use a greedy layer by layer pre-training algorithm to initialize the network weights. Unsupervised contrastive divergence algorithm is used to maximize the marginal probability. |

In 2010 Daniel Povey describe an acoustic modeling approach in which all phonetic states share a common Gaussian Mixture Model structure, and the means and mixture weights vary in a subspace of the total parameter space. We call this a Subspace Gaussian Mixture Model (SGMM) [24].

***Hidden Markov Model***

In the late 1960 and early 1970's Baum and his colleagues was implemented the HMM for speech recognition [25][26][27].From 1970's HMM algorithm widely used in all speech recognition system and become increasingly popular in 1970's.Because First this models are very rich in mathematical structure and hence can form the theoretical basis for use in wide range of application. Second the models, when applied properly work well in practice for several important applications [28].

Hidden Markov Model (HMM) is defined to be a state machine. The states of the model are represented as nodes and the transitions are represented by edges. The elements of Hidden Markov model are Transition probability, observation probability, and Initial Probability Distribution. Figure4 shows the representation of HMM for a sentence.

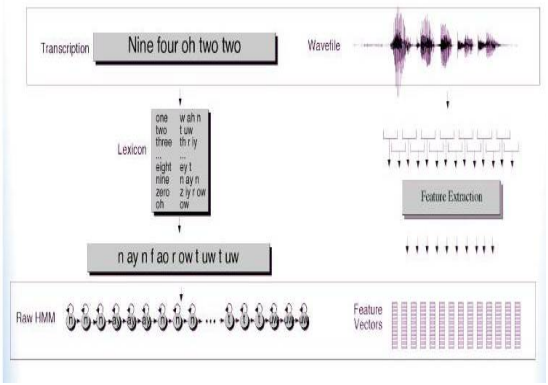


**Figure 4** Representation of HMM for a sentence

Forward-Backward algorithm, Baum- Welch algorithm, Viterbi algorithm are the three al- gorithms used in the Hidden Markov Model. In Hidden Markov Model, Forward-Backward and Baum-Welch algorithms works as a learning algorithm for finding the posteriori probability. Viterbi decoding algorithm is used to find the maximum posterior probability of words or sub words sequence. The Emission probability of Hidden Markov model is expressed as

$$b_j(k) = \frac{\sum_{s.to_t=v_k}^{T} \gamma t(j)}{\sum_{t=1}^{T} \gamma t(j)} \quad (3)$$

Where

$\sum_{s.to_t=v_k}^{T} \gamma t(j)$ - expected number of times in state j and observing symbol

$\sum_{t=1}^{T} \gamma t(j)$ - expected number of times in state j

***Dynamic Time Wrapping***

The DTW is a well-known algorithm in many areas. While first introduced in 1960s [29] and extensively explored in 1970s by application to the speech recognition [30], [31]. DTW is a much more robust distance measure for time series, allowing similar shapes to match even if they are out of ns iphase in the time axis. Dynamic time warping has been shown to be an effective method of handling variation the time scale of polysyllabic words spoken in isolation. This class of techniques has applied to connected word recognition with high degrees of success [31].

***A Stochastic Segment Model***

In 1987 a new direction in speech recognition via statistical methods is to move from frame based models such as HMM to segment based model that provide a better frame work for modeling the dynamics of speech production mechanism [32] . The Stochastic Segment Model is a joint model for sequence of observations which provides a explicit modeling of time correlation as well as formalism for incorporating segmental feature .Most of the existing ASR system today utilize the frame based in HMM .Although this approach has been very successful, it has some drawbacks, For example Temporal dynamics can only be modeled through the use of additional derivative features [33].

***Maximum Entropy Direct Model***

Maximum Entropy Direct Model has been used for statistical natural language understanding [34], for information extraction and segmentation [35], and only recently for acoustic modeling [36]. The direct model can potentially make decoding simpler. It can also be a joint acoustic and language model. speech recognition using maximum entropy

direct modeling, where the probability of a state or word sequence given an observation sequence is computed directly from the model. Asynchronous and overlapping features can be incorporated formally, unlike the case for HMMs [37].

***Support Vector Machines***

In 1999 SVM was applied to speech recognition system [ 38] .Support Vector Machines (SVMs) are state-of-the-art classifiers. SVMs solution relies on maximizing the distance between the samples and the classification border. This distance is known as the margin and, by maximizing it, they are able to generalize unseen patterns. This maximum margin solution allows the SVM to outperform most nonlinear classifiers in the presence of noise, which is one of the long standing problems in ASR. Also, SVMs don't have the convergence and stability problems typical of other classifiers as Neural Networks (NNs) [39][40][41][42]. SVM have attained prominence due to their inherent discriminative learning and generalization capabilities [40].

***Artificial Neural Network***

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach (ANN) and pattern recognition approach (HMM) .In particular recognition systems based on HMMs are effective under many circumstances, but do suffer from some major limitations that limit applicability of ASR Technology in real word environments. Attempts were made to overcome the limitations with the adoption of ANNs as an alternative paradigm for ASR, but ANNs were unsuccessful in dealing with long time sequence of speech signals. Between end of 1980s and beginning 1990s, some researchers began exploring new research area by combining HMMs and ANNs with single hybrid architecture.

The very first in 1987 Multi Layer Perceptron (MLP) neural network was applied for speech recognition technology [43]. Then MLP with different techniques was applied to improve the performance of ASR system [44] [45]. In 1989 Waibel, Alex was proposed the time-delay neural networks for ASR system [46] [47] [48]. In 1990s Recurrent Neural Network was introduce by Robinsion.T [49] [50]. Recently, Deep Belief Networks (DBNs) were introduced as a newly powerful machine learning technique. Generally, DBNs are MLPs with many hidden layers. DBNs use a greedy Layer-by-layer pre-training algorithm and totally unsupervised [51]. In 2010 DBN was successfully applied for phoneme recognition [52] [53].

***Pronunciation Dictionary***

Responsible for determining how a word is pronounced. The pronunciation dictionary plays a role in determining the phonetic transcription of words uttered by speakers in the training corpus. It is a text file with an entry on each line, consisting of a word followed by a phoneme sequence. Accuracy of ASR system also depends on the pronunciation Dictionary [54] [55]. The pronunciation dictionary used to train the acoustic models was compiled from two different sources: the transcription information belonging to the corpus and a dictionary file. Classification of dictionary depending upon the sound unit. It is classified as

- Word Level Dictionary,
- Phoneme Level Dictionary,
- Tri phone Level Dictionary,
- Morpheme Level Dictionary,
- Syllable Level Dictionary.

***Language model***

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. Language modeling is the task of estimating the probability distribution of linguistic units such as words and sentences. The probability distribution itself is referred to as a language model.

The most prominent use of language models is in Automatic Speech Recognition (ASR), where the language model assigns a prior probability to help differentiate words that have similar acoustical properties .Language models have been used in a variety of NLP tasks including speech recognition, document classification, optical character recognition, and statistical ma- chine translation. The design of speech recognition system requires careful attention to language models of various stages in order to improve the accuracy of the speech recognition system.

***Types of language model***

The main classification of language models are unigram language model, n-gram language model and neural network based language models. Sometimes, choosing a language model depends on the application [66].

In unigram language model the probability of generation of a term is an independent event and does not depend on the previous terms being generated [67]. The words bigram and trigram language model denoted as n-gram language models with n=2 and n=3, respectively, which defined the probability of occurrence of an ordered sequence of n words, was the most frequently used variant. The second type is the n gram model which assigns different probabilities according to the likelihood of n phones or n syllables appear together in the written Text [68].

The next type of language model is Neural Network (NN) based language modeling architecture can be divided into two types: recurrent and non-recurrent networks. An important non-recurrent neural network consists of architectures in which cells are organized into layers, and only unidirectional connections are permitted between adjacent layers. This is known as a feed forward multi-layer perceptron (MLP) architecture. On the other hand; recurrent neural networks are characterized by both feed forward and feedback paths between the layers. The feedback paths are enable the activation at any layer either to be used as an input to a previous layer or return to that same layer after one or more time steps [69].

***Unigram language model***

The language models unigrams are simple: it assumes that the probability of a token is independent of the surrounding tokens, given the grade language model. A unigram language model is defined by a list of type's words and their individual probabilities. Although this is a weak model, it can be trained from less data than more complex models, and turns out to give good accuracy for our problem [70]. Unigram language model approximation is

$$P_{uni} = (t_1 t_2 t_3) = P(t_1)P(t_2)P(t_3) \quad (4)$$

***N gram language model***

ASR systems utilize n-gram language models to guide the search for correct word sequence by predicting the likelihood of the nth word on the basis of the n−1 preceding words. Since 1980's, n-gram language models, and its variants, idea can be traced to an experiment by Claude Shannon for large vocabulary speech recognition systems[71].In n-gram model, the probability of observing the sentence *w*1, ..., *wm* is approximated as

$$P(w_1, w_2 \dots . w_m) = \prod_{i=1}^{m} P(w_i | w_1, w_2 \dots . w_{i-1}) = P(w_i | w_{i-(n-1)}, \dots . . w_{i-1}) \quad (5)$$

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. The most popular language models in use are the N-grams. Although they are effective for some applications, their predictive power is limited. Usually N is of the order of 2 or 3, which greatly restricts the predictive power of the N-grams. Higher order N-grams (of the order of N = 6,7,...) have been tried, but they have been found to be unreliable, the main reason for this being data sparseness. Even higher order N-grams cannot capture the long range dependencies of natural language, which humans can easily identify. Several attempts have been made to overcome this limitation. The earliest attempts in this regard were made in the form of variable length word-category based N-grams [72].

***Neural Network Based Language Model***

A series of papers [Schwenk Gauvain 2002, 2003, 2004a, 2004b, Schwenk 2004c] has examined a connectionist approach to statistical language modeling. Statistical language models (LM) play an important role in state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems.

Back-off n-gram and class-based LMs [73][74][75] are the dominant language models used in LVCSR systems. However, when only limited amounts of text data is available in training and adaptation, the generalization ability of these discrete, non-parametric models remain limited. To handle this data sparsity problem, a range of language modeling techniques based on a continuous vector space representation of word sequences have been proposed [76][77].

Among these one of the most successful schemes is the neural network LM (NNLM). Due to their inherently strong generalization and discriminative power, they have become an increasingly popular choice for LVCSR tasks. Neural networks in language modeling offer the following advantages over competing approaches: In contrary to commonly used n-gram language models, there is no necessity of smoothing in cases of sparse training data. Due to the projection of the entire vocabulary into a small hidden layer, semantically similar words get clustered [78].

The architecture of the neural network language model is shown in Figure5.The neural network language model is implemented as a standard fully connected multilayer perceptron with three layers, termed the projection, hidden and output layers. Only the hidden and output layers have a non-linear activation function [79]. The inputs to the network are the indices of the previous words that define the n-gram context

$$h_j = w_{j+n+1}, w_{j+n+2} \dots . . , w_{j-1} \quad (6)$$

Where

$h_j$ - Context
$w_j$-Taget Word
The outputs of the network are the posterior probabilities of all words in the vocabulary given the history:

$$P(w_j = i | h_j), i = 1 \dots . . N \quad (7)$$

Where $i = 1 \dots . . N$

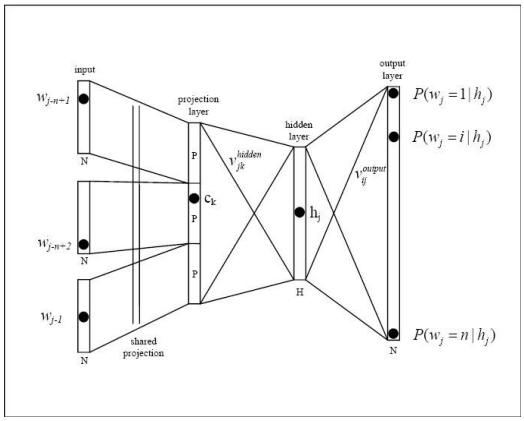


**Figure 5** Architecture of Neural Network Language Model

The projection layer maps the discrete word indices of an n-gram context to a continuous Vector space. The hidden layer processes the output of the projection layer and is also created with a number of neurons specified in the topology configuration file. The output layer processes the output from the hidden layer and is created with a number of neurons equal to the size of the vocabulary (or, if enabled, the size of the shortlist) and a number of weights equal to the dimensionality of the hidden layer output. There are also outputs for Out Of Vocabulary and the sentence end token. When training w i'th a shortlist there is no output for OOV. The purpose of the output layer is to compute the posterior probabilities of each word *wj* in the vocabulary given the n-gram context *hj* that was fed forward through the network [80].

The major classifications of neural network based language model are feed forward neural network based language model and recurrent neural network based language model (RNNLM). The RNN is similar approaches to feed-forward networks, except that recurrency between hidden and input layer is being added. RNN has an input layer, hidden layer (also called context layer or state) . At each time, the both neural network receives an input, updates its hidden state, and makes a prediction [81].

***Evaluation Metrices for Language model***

In the literature, two primary metrics are used to estimate the performance of language models in speech recognition systems. First, they are evaluated by the word error rate (WER) (Discussed in 9.1.1) yielded when placed in a speech recognition system [82]. Second, and more commonly, they

are evaluated through their perplexity on test data, an information theoretic assessment of their predictive power. While word error rate is currently the most popular method for rating speech recognition performance, it is computationally expensive to calculate. Perplexity is a measure of how well the model predicts the occurrence of words in a given text[83]. The perplexity of a language model X is given by

$$PP(X) = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i | H_{q-1}) \quad (8)$$

Where
N-Total Number of words in the test set.

Some of the language modeling toolkits are The CMU-Cambridge Statistical Language Modeling Toolkit[84],Random Forest Language Model Toolkit[85][86], RNNLM - Recurrent Neural Network Language Model Toolkit[87][88], SRILM - An Extensible Language Modeling Toolkit[89][90], The MIT Language Modeling (MITLM) toolkit[91][92].IRST Language modeling toolkit[93] .Language models are useful in a large number of areas, including speech recognition, handwriting recognition, machine translation, information retrieval, context-sensitive spelling correction, and text entry for on small input devices.Table3 shows some of the Language Modeling toolkits.

***Tools of ASR***

***CMU Sphinx***

CMU Sphinx is the general term to describe a group of speech recognition systems developed at Carnegie Mellon University. In 2000, the Sphinx group at Carnegie Mellon committed to open source several speech recognizer components, including Sphinx 2, Sphinx3 and Sphinx4 [57]. The speech decoders come with acoustic models and sample applications [56] .

***HTK***

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition. HTK consists of a set of library modules and tools available in C source form. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis [58] [59].

***JRTk***

Janus Recognition Toolkit (JRTk), sometimes referred to as Janus, is a general purpose speech recognition toolkit developed and maintained by the Interactive Systems Laboratories at Carnegie Mellon University. The JRTk provides a flexible Tcl/Tk script based environment which enables researchers to build state-of-the-art speech recognizers and allows them to develop, implement, and evaluate new methods[60][61].

**Table 3** Language Model Toolkit

| Name of The Toolkit | Released year | Description |
|---|---|---|
| CMUSLM | 1994 | N-gram based Language Model |
| SRILM | 1995 | N-gram based language Model, written in C++. |
| RFLM | 1997 | C++ software package based on the SRI LM Toolkit. It is a collection of randomized decision tree language models. |
| MITLM | 2008 | statistical n-gram language models involving iterative parameter estimation |
| RNNLM | 2011 | Based on Recurrent Neural Network |
| IRSTLM | 2011 | Suitable to estimate, store, and access very large LM. |

***Speech Corpus preparation for ASR***

Speech corpus for ASR consists of speech audio files and text transcription. Transcriptions contain the sequence of words and non speech sounds are written exactly as they occurred in a speech signal. This transcription is used to record the word/sentence through a single speaker or number of speaker. Some of the standard speech corpus providers are Linguistic Data Consortium (LDC), the European Language Resources Association (ELRA), the Japanese Language Resource Consortium(JLRC) and Evaluations and Language resources Distribution Agency(ELDA).During the preparation of speech corpus for speech recognition a lot of manual effort is spent on the preparation .i.e. Recording of speech data, Segmentation, Dictionary generation. The tool for preparation of the speech corpus is presented in the Table 4.

***Dragon Naturally Speaking***

Dragon NaturallySpeaking is a speech recognition software package developed and sold by Nuance Communications. The software has three primary areas of functionality: dictation, text-to-speech and command input. The user is able to dictate and have speech transcribed as written text [62] [63].

***Kaldi***

Kaldi is a toolkit for speech recognition and licensed under the Apache License v2.0. Kaldi is similar in aims and scope to HTK. The goal is to have modern and flexible code, written in C++, which is easy to modify and extend [64] [65].

**Table 4** Speech Corpus Preparation Tools

| Techniques | Properties |
|---|---|
| Audacity | Used for recording the speech signal at different frequency and sample rate. |
| Praat | Analysis of speech in phonetics. |
| Emu | To find various speech segments based on the sequential and hierarchical structure of the utterance in which they occur. |
| ToBI | ToBI(Tones and Break Indices) is a set of conventions for transcribing and annotating the prosody of speech. |
| DAMSL | Dialog Act Markup in Several Layers. A set of primitive communicative actions that can be used to analyze dialogs. |
| MATE | Support for the annotation of speech and text. |
| Transcriber | Tools for segmenting, labeling and transcribing speech. |
| LIUM | for speech recording and segmentation |

**Table 5** Growth of ASR System

| Author | Year | Research Work | Description |
|---|---|---|---|
| K.H.Davis R.Biddulh and S.Balashek | 1952 | Automatic Recognition of spoken digits | Bell labs spoken Digit recognizer, The system relied on Measuring spectral resonances during the vowel region of each Digit. |
| Velickko etal | 1970 | Isolate word recognition | Pattern recognition ,Dynamic time wrapping linear predictive coding ideas were applied to speech recognition |
| C.SMyers and L.R.Rabiner | 1981 | A Comparative Study of several Dynamic Time wrapping algorithm for connected word recognition | The Two Level Dynamic Programming Matching (TLDPM) algorithm attempts to find the best concatenation of reference patterns |
| Teuvo Kohonen | 1988 | The Neural Phonetic type writer | Vector quantization, neural network Model, Short learning algorithm used is described. |
| Steve Young | 1996 | A Review of Large Vocabulary Continuous Speech Recognition | MFCC Based front end, HMM Based pattern recognition approach. |
| Dan Ellis | 1998 | Speech recognition at ICSI; Broadcast news and beyond. | Feature adaptation technique was used such as VTLN, MLLR, and SAT. |
| SaruwatariH | 2009 | Hands free speech recognition for real world speech dialogue systems | Hands free speech dialogue system which is used for railway station guidance. |
| J.Fiscus and J.Garofolo | 2002 | RT-2002 Evaluation plan NIST | Rich transcription of meetings, Very large vocabulary, controlled environment. |

***State of the Art of ASR System***

Recent years have seen a considerable growth in the development of practical systems for automatic speech recognition (ASR). Building a speech recognition system becomes very much complex because of the criterion mentioned in the previous section. Now the research in ASR is concentrating in the following feature which is speaker independence, large vocabulary, 100% accuracy, developed for many languages, Speech capabilities. Table 5 shows the growth of ASR system in the last 60 years.

***Performance Measure of ASR***

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

***Accuracy***

***Word Error Rate***

Word error rate is a common metric of the performance of a speech recognition or machine translation system. 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.

$$WER = \frac{S+D+I}{N} \quad (9)$$

Where
5
S is the number of substitutions,
D is the number of the deletions,
I is the number of the insertions,
N is the number of words in the reference.

***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

$$RTF = \frac{P}{I} \quad (10)$$

## CONCLUSION

In this survey, we have discussed the technique developed in each stage of speech recognition system.

We also presented the list of technique with their properties for Feature extraction, Acoustic model and Language Model. Through this review it is found that MFCC is used widely for feature extraction of speech, LDA is suitable for Dimensionality reduction technique, DBN is an appropriate model for acoustic model technique and RNNLM is acceptable for Language model.

## References


1. Davis, K., Biddulph, R., and Balashek, S., Automatic Recognition of Spoken Digit *J. Acoust. Soc. Am.* 24: Nov 1952, p. 637.
2. Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn An Evaluation Of Audio-Visual person Recognition on the XM2VTS corpus using the Lausanne protocol MIT Lincoln Laboratory, 244 Wood St., Lexington MA.
3. W. M. Campbell, D. E. Sturim W. Shen D. A. Reynolds.J Navratily The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition MIT Lincoln Laboratory IBM 2006.
4. Davis, S.B., Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences IEEE Trans. on Acoustic, Speech and Signal Processing, 1980, pages: 357-366.
5. L. Muda, M. Begam, I. Elamvazuthi Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques *Journal of Computing*, 2(3), 2010, 138-143.
6. A. A. M. Abushariah, T. S. Gunawan, O. O. Khalifa English digits speech recognition system based on Hidden Markov Models *Int. Conf. on Computer and Communication Engineering*, 11-13 May 2010, 1-5.
7. B. Kotnik, D. Vlaj, Z. Kacic, B. Horvat Robust MFCC feature extraction algorithm using efficient addictive and convolutional noise reduction procedures ICSLP'02 Proceed- ings, 2002, 445-448.
8. Hermansky, H., Hanson, B. and Wakita, H. Low dimensional representation of vowels based on all-pole modelling in the psychophysical domain Speech Communication, (1985) , 4(13):181-187.
9. Hermansky, H. Perceptual Linear Predictive (PLP)

analysis of speech *J. Acoust.Soc. Am*., (1990), 87(4): 1738–1752.
10. Corneliu Octavian DUMITRU, Inge GAVAT A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.
11. DOUGLAS O'SHAUGHNESSY Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis Proceedings of the IEEE, September 2003, VOL. 91, NO. 9, 00189219/03 2003 IEEE.
12. Rok Gajsek and France Mihelic Comparison of speech parameterization techniques for Slovenian language 9th International PhD Workshop on Systems and Control, October 2008, 1. - 3.
13. D.A. Reynolds Experimental evaluation of features for robust speaker identification IEEE Trans. Speech Audio Process, Oct 1994 vol. 2 (4), pp. 639-43.
14. Arvind Agarwal, Jeff M. Phillips, Suresh Universal Multi-Dimensional Scaling ACM, July 25-28, 2010, 978-1-45030055-110/07.
15. S. T. Roweis and L. K. Saul Nonlinear Dimensionality Reduction by Locally Linear Embedding Science, 22 December 2000, Vol 290, 2323-2326.
16. Haeb Umbach.R Linear discriminant analysis for improved large vocabulary continuous speech recognition IEEE International Conference on Mar 1992 23-26.
17. Aleix M. Martinez, Aleix M. Martinez, Avinash C.Kak PCA versus LDA IEEE Transactions on Pattern Analysis and Machine Intelligence 2001.
18. Haeb-Umbach Linear discriminant analysis for improved large vocabulary continuous speech recognition IEEE Transaction - Acoustic, Speech and Signal Processing on 23-26 Mar 1992, Volume: 1, Page(s): 13 - 16 vol.1.
19. D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models IEEE Trans. *Speech and Audio Processing*, 1995, vol. 3, pp. 72-83.
20. D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models Speech Communication, 1995, vol. 17, no. 1-2, pp. 91-108 .
21. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, Speaker verification using adapted Gaussian mixture models Digital Signal Processing, 2000, vol. 10, pp. 19-41,.
22. J. Fortuna, P. Sivaaaran, A. Ariyaeeinia, and A. Malegaonkar "Open-set speaker identification using adapted Gaussian mixture models in Proc. Interspeech, Lisbon, Portugal, Sep. 2005, pp. 1997-2000.
23. A. Stergiou, A. Pnevmatikakis, and L. C. Polymenakos Enhancing the performance of a gmm-based speaker identification system in a multi-microphone setup in Proc. Interspeech, Pittsburgh, Pennsylvania, Sep. 2009, pp. 1463-1466.
24. Daniel Povey1, Luk ´ s Burget2 SUBSPACE GAUSSIAN MIXTURE MODELS FOR SPEECH RECOGNITION Computer Speech and Language, Vol. 25, No. 2, Amsterdam,Volume 25, Issue 2, April 2011,Pages 404-439.
25. F.Jelinek A fast sequential decoding algorithm using a stack IBM *J.Res Develop*, 1969, Vol 13, PP 675-685.
26. L.R Bahl and F.Jelinek Decoding for channels with insertions, deletions and substitutions with applications to speech recognition IEEE Trans , 1975, Vol 1T-21,pp 404-411.
27. L.R Bahl and F.Jelinek Design a Lingustic statistical decoder for the recognition of continuous speech IEEE Transaction vil 1T-21, 1975, pp 250-256.
28. Lawrence R. Rabiner A Tutorial on Hidden Markov Model and Selected Applicaiton in Speech Recognition IEEE, 1989, vol 77.
29. R. Bellman and R. Kalaba, on adaptive control processes," Automatic Control IRE Transactions on 1959, vol. 4, no. 2, pp. 1-9.
30. C. Myers, L. Rabiner, and A. Rosenberg Performance tradeoffs in dynamic time warping algorithms for isolated word recognition Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 1980, vol. 28, no. 6, pp. 623-635.
31. H. Sakoe and S. Chiba Dynamic programming algorithm optimization for spoken word recognition IEEE Transactions on Acoustics, Speech and Signal Processing, IEEE Transactions on, 1978, vol. 26, no. 1, pp. 43-49.
32. M.A Bush and G.K Kopec Network based connected digit recognition IEEE Tran Acoustic Speech, Signal Processing, 1987, Vol ASSP-35 PP 1401-1413.
33. Mari OSTENDORF and SALIM ROUKOS A Stochastic Segment-based Model for Phoneme based Continuous speech recognition IEEE Tran, Acoustic Speech, Signal Processing, 1989, vol 37.
34. K. A. Papineni, S. Roukos, and R. T. Ward Feature-based language understanding in Proc. Eurospeech-97, (Rhodes, Greece), Sept. 1997.
35. McCallum, D. Freitag, and F. Pereira Maximum entropy Markov models for information extraction and segmentation in Machine Learning: Proceedings of the Seventeenth International Conference (ICML 2000), (Stanford, California), 2000, pp. 591–598.
36. A. Likhododev and Y. Gao Direct models for phoneme recognition in Proc. ICASSP 2002, vol. I, (Orlando, FL), May 2002, pp. 89–92.
37. Hong-Kwang Jeff Kuo, Yuqing Gao MAXIMUM Entropy Direct Models For Speech Recognition IEEE, 2003.
38. P. Clarkson On the use of support vector machines for phonetic classification in ICASSP99, 1999, pp. 585–588.
39. E. Osuna, *et. al.* "An Improved Training Algorithm for Support Vector Machines Proceedings of the IEEE NNSP'97, pp. 24-26, Amelia Island, FL, USA, September 1997.
40. B. Scho¨lkopf Support Vector Learning., Ph.D. Thesis, R. Oldenbourg Verlag Publica- tions, Munich, Germany, 1997.
41. Ganapathiraju, J. Hmaker, and J. Picone Hybrid SVM/HMM architectures for speech recognition in Proc. of the International Conference on Spoken Language Processing, 2000, vol. 4, pp. 504-507.
42. N. Smith and M. Gales, Speech recognition using SVMs Advances in Neural Information Processing Systems 14. MIT Press, 2002.
43. R.P Lippmann and B. Gold Neural Classifier useful for

speech recognition In IEEE Proc. First Intl Conf on Neural Networks 1987, Vol 4 pages 417-422.
44. Cosi, Piero, Bengio, Yoshua, De Mori, Renato Phonetically-based multi-layered networks for acoustic property extraction and automatic speech recognition Speech Communication on 1990 Volume-9, Pages 15-30
45. Bourlard.H Links between Markov models and multilayer perceptrons IEEE Transaction-Pattern analysis and Machine Intelligence on December-1990, Volume: 12, Page(s): 1167 - 1178
46. Waibel, Alex Modular construction of time-delay neural networks for speech recognition Neural Computation, 1989, Vol 1(1), 39-46.
47. Waibel.A Phoneme recognition using time-delay neural networks IEEE Transaction - Acoustic, Speech and Signal Processing on March -1989 Volume: 37, Page(s): 328 - 339 .
48. Waibel.A Modularity and scaling in large phonemic neural networks IEEE Transaction-Acoustic, Speech and Signal Processing on December -1989 Volume: 37, Page(s): 1888 - 1898 .
49. Tan LEE, P.C CHING, L.W.CHAN, Recurrent Neural Network for Speech Modeling and Speech Recognition IEEE International Conference -Acoustic, Speech and Signal Processing on May 1995 Volume: 5, Page(s): 3319 -3322
50. Robinsion.T A real-time recurrent error propagation network word recognition system International Conference -Acoustic, Speech and Signal Processing on March-1992,Volume: 1 Page(s): 617 - 620 vol.1
51. G. Hinton, S. Osindero, and Y. Teh, A fast learning algorithm for deep belief nets Neural computation, 2006, vol. 18, no. 7, pp. 1527-1554,.
52. G. Dahl, A. MarcAurelio Ranzato, and G. Hinton Phone Recognition with the Mean-Covariance Restricted Boltzmann Machine Advances in Neural Information Processing Systems, 2010, vol. 24.
53. Van Hai Do, Xiong Xiao, Eng Siong Chng Comparison and Combination of Multilayer Perceptrons and Deep Belief Networks in Hybrid Automatic Speech Recognition Systems APSIPA ASC 2011 Xi'an.
54. R. Thangarajan· A.M. Natarajan · M. Selvam Phoneme Based Approach in Medium Vocabulary Continuous Speech Recognition in Tamil *International Journal of Computer Processing of Oriental Languages.*
55. R. Thangarajan · A.M. Natarajan · M. Selvam Syllable modeling in continuous speech recognition for Tamil language *Int J Speech Technol* (2009) vol :12,pages: 47–57
56. Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel Sphinx-4: A Flexible Open Source Framework for Speech Recognition SMLI TR2004-0811 c 2004 SUN MICROSYSTEMS INC.
57. http://cmusphinx.sourceforge.net/
58. S. Young The HTK hidden Markov model toolkit: Design and philosophy Cambridge University Engineering Department, UK, Tech. Rep. CUED/FINFENG/TR152, Sept. 1994.
59. http://htk.eng.cam.ac.uk/
60. Finke, Michael, *et.al.* The JanusRTk Switch-board/Callhome 1997 Evaluation System
61. Proceedings of the LVCSR Hub5-e Workshop, Baltimore, USA, 1997.
62. Zeppenfeld, Torsten, *et. al.* Recognition of Conversational Telephone Speech Using the Janus Speech Engine IEEE International Conference on Acoustics, Speech, and Singal Processing, Germany, 1997. "Dragon NaturallySpeaking TM Creating Voice Commands Dragon Systems. August 1998. Version 3.0
63. http://www.nuance.com/naturallyspeaking/
64. Daniel Povey, Arnab Ghoshal The Kaldi Speech Recognition Toolkit Microsoft Research, USA -2011.
65. http://kaldi.sourceforge.net/
66. S.Saraswathi,T.V.Geetha Design of language models at various phases of Tamil speech recognition system , 2010, Vol. 2, No. 5, pp. 244-257
67. Xiaoyong Liu and W. Bruce Croft Statistical Language Modeling For Information Retrieval ARIST (2005), 39(1): 1-31.
68. Chin-Hui Lee, Biing-Hwang Juang A Survey on Automatic Speech Recognition with an Illustrative Example on Continuous Speech Recognition of Mandarin Computational Linguistics and Chinese Language Processing August 1996, vol.1, no.1.
69. M.M.El Choubassi, H.E.El Khoury, C.E.Jabra Alagha, J.A.Skaf and M.A.AlAlaoui Arabic Speech Recognition Using Recurrent Neural Networks IEEE transaction 2003.
70. Kevyn Collins-Thompson Jamie Callan A Language Modeling Approach to Predicting Reading Difficulty *JASIST*, (2005), 1448-1462.
71. B.H. Juang Lawrence R. Rabiner Automatic Speech Recognition – A Brief History of the Technology Development Encyclopedia of Language and Linguistics Elsevier Citeseer, 2005
72. T. Niesler and P. Woodland, A variable-length category-based n-gram language model in Proc.ICASSP '96, Atlanta, GA, 1996, pp. 164-167.
73. S.M. Katz, Estimation of probabilities from sparse data for thelanguage model component of a speech recognizer IEEE Trans.Acoustics, Speech and Signal Processing, 1987, Vol. 35, No. 3.
74. P.F. Brown et al Class-based n-gram models of natural language", Computational Linguistic 1992, Vol. 18, No. 4, pp. 467-479.
75. A.Paccanaro and G.E. Hinton Extracting distributed representations of concepts and relations from positive and negative propositions In Proc. IJCNN2000.
76. D.Mrva and P.C. Woodland "A PLSA-based language model for conversational tele- phone speech In Proc. Interspeech2004.
77. T.Brants Test data likelihood for PLSA models Information Retrieval, 2005, Vol. 8, No. 2, pp. 181-196.
78. Stefan Kombrink, Toma´˘s Mikolov Recurrent Neural Network Lan- Guage Modeling Applied To The Brno Ami/Amida 2009 Meeting Recognizer Setup Proceedings of the 17th Conference STUDENT EEICT, 2011, p. 527-531.
79. Robert D. Dony, Simon Haykin Neural Network Approaches to Image Compression PROCEEDINGS OF THE IEEE, FEBRUARY 1995, VOL. 83, NO. 2.
80. Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas

Burget,Jan Honza Cernocky Strategies for Training Large Scale Neural Network Language Models IEEE Signal Processing Society 2011.
81. Ilya Sutskever, James Martens, Georey Hinton Generating Text with Recurrent Neural Networks ICML 2011.
82. Stanley Chen, Douglas Beeferman, Ronald Rosenfeld Evaluation Metrics For Language Models In Darpa Broadcast News Transcription and Understanding Workshop, 1998.
83. Kumaran, R., Gowdy, J. N., and, Narayanan, K. Language Modeling Using Independent Component Analysis for Automatic Speech Recognitionof the European Signal Processing Conference (EUSIPCO), Antalya, Turkey, September 2005.
84. P.R. Clarkson and R. Rosenfeld Statistical Language Modeling Using the CMU- Cambridge Toolkit Proceedings ESCA Eurospeech 1997.
85. Peng Xu and Frederick Jelinek Random Forests in Language ModelingIn Proceedings of EMNLP 2004, pages 325-332.
86. http://old-site.clsp.jhu.edu/ yisu/rflm.html
87. Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukas Burget, Jan Cernocky. RNNLM - Recurrent Neural Network Language Modeling Toolkit IEEE - ASRU 2011.
88. http://www.fit.vutbr.cz/ imikolov/rnnlm/
89. Andreas Stolcke SRILM —An Extensible Language Modeling Toolkit" Proc. Intl. Conf. on Spoken Language Processing, 2002, vol. 2, pp. 901-904.
90. http:/www.speech.sri.com/projects/srilm/
91. Bo-June (Paul) Hsu and James Glass. Iterative Language Model Estimation: Efficient Data Structure Algorithms in Proc. Interspeech, 2008.
92. http://code.google.com/p/mitlm/
93. http://sourceforge.net/projects/irstlm

**How to cite this article:**

Sundarapandiyan S and Shanthi N (2017) 'A Survey on Automatic Speech Recognition System ', *International Journal of Current Advanced Research,* 06(09), pp. 6287-6297. DOI: http://dx.doi.org/10.24327/ijcar.2017.6297.0912

*******