TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

A Study of Language Models in Automatic Speech Recognition System

[1]Sundara Pandiyan S & [3] Shanthi N
[1,2&3]Assistant Professor,Department of Computer Science and Engineering,School of Engineering and Technology,CHRIST (Deemed to be University),Bangalore,India.

Article Info
Volume 82
Page Number: 16635 - 16640
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:

This paper presents a brief survey on Language Models in Automatic Speech Recognition. After years of research and development the accuracy of automatic speech recognition remains one of the important research challenges (eg. variations of the context, speakers, and environment). The design of Speech Recognition system requires careful attentions to language models used in various stages in order to improve the accuracy of the Speech Recognition system. The objective of this review paper is to summarize and compare some of the well-known language models and identify research topic and applications which are at the forefront of this exciting and challenging field.

Keywords: *component*:*Automatic Speech Recognition; Statistical Language Modeling*

I. INTRODUCTION

A. *Speech Recognition*

Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program.

B. Basic Model of Speech Recognition

Research in speech processing and communication for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine.

For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years. This report reviews some of the language models in speech recognition. Although many technological progresses have been made, still there remain many research issues that need to be tackled.

Figure.1 shows a mathematical representation of speech recognition system in simple equations which contain front end unit, model unit, language model unit, and search unit. The recognition process is shown below (Figure .1).

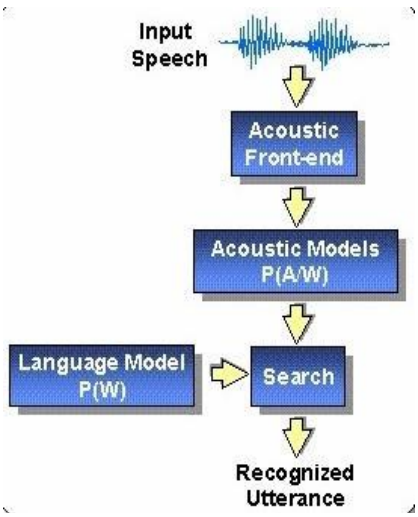

FIGURE1 .THE RECOGNITION PROCESS

1) Front End: Parameterize an input signal (e.g. audio) into a sequence of output features. It

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

performs digital signal processing (DSP) on the incoming data.

2) Acoustic Model: Contains a representation (often statistical) of a sound, often created by training using lots of acoustic data.

3) Language Model: Contains a representation (often statistical) of the probability of occurrence of words.

4) Search Graph: The graph structure produced by the linguist according to certain criteria (e.g., the grammar), using knowledge from the dictionary, the acoustic model, and the language model.

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produces an acoustic observation sequence Y, with probability P(W,Y). The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability.

$$P(W/A)= arg\ max_w\ P(W/A) \quad (1)$$

Using Baye's rule, equation (1) can be written as

$$P(W/A)=P(A/W)P(W) / P(A) \quad (2)$$

Since P(A) is independent of W, the MAP decoding rule of equation (1) is

$$W=argmax_w\ P(A/W)P(W) \quad (3)$$

The first term in equation (3) P(A/W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(A/W) is computed. The second term in equation (3) P(W), is called the language model[1]. It describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

C. *Automatic Speech Recognition Classification*

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in Figure 2.

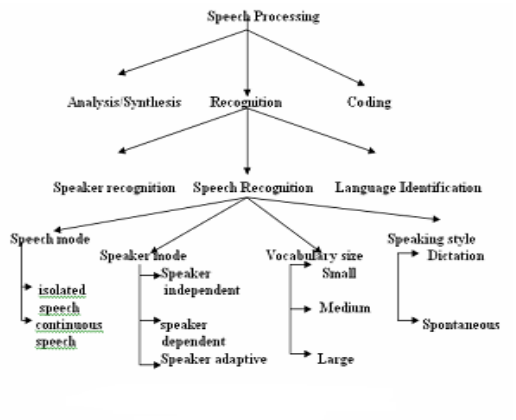

Figure2 Speech Processing Classifications

II. Language Model

Language model or grammar model which models patterns of word usage. This is normally customized for the application. Every word in the language model must be in the pronunciation dictionary. Choosing a language model depends on the application sometimes [2].

Current systems use statistical language models to help reduce the search space and resolve acoustic ambiguity [3]. As vocabulary size grows and other constraints are relaxed to create more habitable systems, it will be increasingly important to get as much constraint as possible from language models; perhaps incorporating syntactic and semantic constraints that cannot be captured by purely statistical models.

A. Types of Language models

1) Phoneme Recognition Language Model.
2) Word Based Language Model.
3) Syllable Based Language Model.
4) Morpheme Based Language model
5) Stem-Ending-Based Model
6) Stochastic (finite-state /n gram) Language Model
7) Neural Network Language Model

1) Phoneme Recognition Language Model:

Phonetic Recognition, Language Modeling (PRLM) is based on this principle. This system uses acoustic pre-processing for feature vector extraction. Then a language specific phoneme recognizer is placed to convert speech into phoneme sequences and at the end lies the language models which

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

calculates the n-Gram probabilities [4]. Figure3 gives a graphical view of the system.

Disadvantage of this system is that it uses a single language dependent phoneme recognizer, which can make its results bias to the language in which recognizer is trained because the phones present in target languages do not always occur in the language used during training. We may wish to incorporate sounds from more than one language into a PRLM-like system. An alternative to it can be to use multiple PRLM systems in parallel, with recognizers trained in different languages.

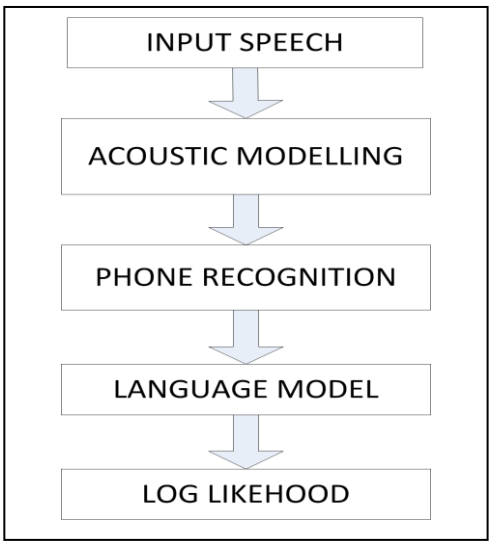

Figure 3 Phoneme Recognition Model

2) Word Based Model: In word-based language model, words are selected as base recognition units of the recognizer. Word-based system can be illustrated as in Figure 4

Here, the problem reduces to simple isolated word recognition. Words have the advantage of becoming longer units and longer language units would result in the higher performance of the acoustic processor. Consider using words as base units of the recognizer and employing a trigram language model. In this case, the dictionary size would be about several hundred thousands of words to cover general text sufficiently. Nevertheless, we would have out of vocabulary words, which are constructed through legal morphological rules. The morphological

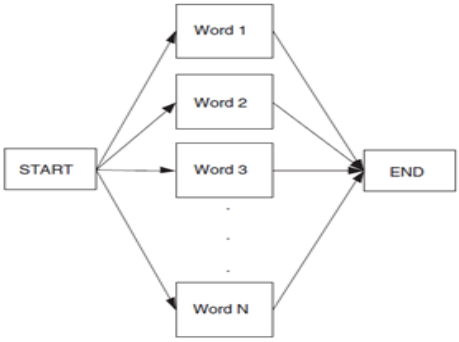

Figure4 Word-Based Model

productivity of the language makes it difficult to construct a word-based language model. The word order is determined by semantic variations rather than strict syntactic rules and the emphasized word is the one that just comes before the word. In the case of continuous speech recognition system using word-based model, bigram and trigram language models yield high perplexities. In addition, word-based model demands a huge vocabulary with a semantic freedom in word order.

3) Syllable Based Language Model: Syllable-based model is the simplest language model proposed as an alternative to the word-based model. Once the training corpus is parsed into its syllables, the n-gram models can be employed as if syllables are words. The syllable-based model is as shown in Figure. 5.

As seen from Figure. 5 the transitions are bilateral; i.e. syllables can follow each other with no constraints. The presence of the transitions between syllables is determined by the presence of the corresponding syllable bigram in the training corpus, and the weights of the transitions are the bigram probabilities.

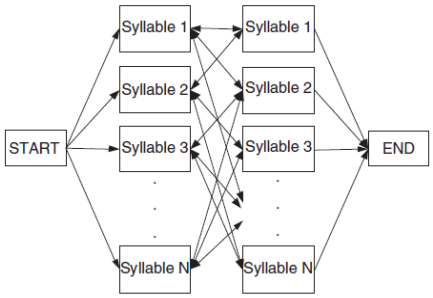

Figure5 Syllable-Based Model

Recognition of some syllable sequences is not possible due to error in recognized phonemes. They are detected using the articulator-acoustic features of the phonemes.

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

Based on the previous syllable pattern and the generation of syllable from the phoneme sequence the current one was identified and the ambiguities were resolved. The syllables were combined to form words and the errors in word level were detected and corrected using morpheme-based language model as discussed in the next section.

4) Morpheme based language models: The morpheme-based language model utilizes morphemes as base units of the recognizer. The morpheme-based lattice is illustrated in Figure.6.

Words are modeled as a stem followed by suffixes in accordance with the morphotactic of the language and with the spelling rules. The decomposition of words into stems and morphemes is performed with the morphological parser. The transitions between morphemes are weighted with the bigram probabilities. Stems are categorized under two groups, nominal and verbal. There are also homonyms; a stem can be both verbal and nominal. Stems are linked with both verbal and nominal suffix lattices. There are transitions between verbal and nominal suffix lattices. In the lattice, units are represented with their surface realizations. The suffix lattices are constructed using the phonetic rules; i.e. all possible suffix sequences obtained through the network obey the phonetic rules. The links between the stems and suffix lattices are also constructed using the phonetic rules. The suffixes that may follow a particular stem are determined by the last vowel and last phoneme of the stem. Once this mapping is defined, new stems can be added to the lattice automatically.

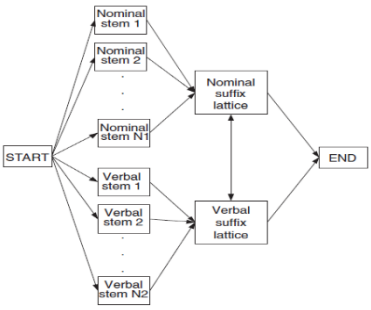

Figure6 Morpheme-Based Model

5) Stem-Ending-Based Model: The stem-ending based language model is also a morphological model. The base units of this model are stems and the group of suffixes following them, which are called ‘‘endings’’. This model can be viewed as a solution to the short units of morpheme-based model and to the problem of very large vocabulary caused by the word-based model. In terms of acoustic models, short recognition units will not be a crucial problem if the acoustic models are trained to cover the cross word co articulation. In that case, short and long recognition units will be equally good. However, short units allow more freedom and larger search space for the recognizer. At one extreme, if we use phonemes as our recognition units, we have to decide the most probable hypothesis from a huge space of n-gram phonemes and that leads to a poor performance of the decoder. Stem-ending based modeling is proposed for agglutinative languages by Kanevsky [5]. Mengusoglu and Deroo [6] used the stem-ending approach for modeling Turkish.

The stem-ending based word model can be Illustrated in Figure7. Not all stems are connected to all endings. The connections are present if the corresponding bigrams appeared in the text corpus.

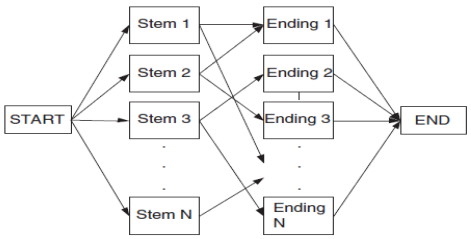

Figure7 Stem-ending Based Model

Finite State/n-gram Models: A finite state (also called stochastic n-gram) language model provides estimates of the prior probabilities of the sequence of any particular series of words. The entire set of possibilities for a four-word sentence in a three-word language model can be described in a tree as shown in Figure 8

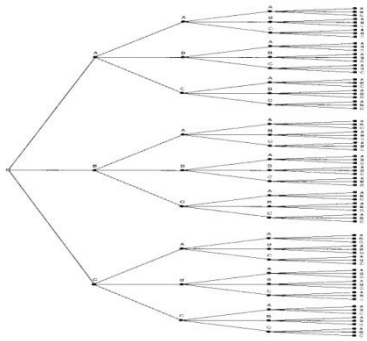

Figure 8: Sentence Tree

These sentences will be characterized by a range of probabilities, with some being improbable (having a probability near 0), and others being highly likely (with probabilities >0.5). A simple

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

dialog system may in fact use a diagram like this to define probable responses. Simply stated, the probability of the occurrence of any word in the chain is based solely on the word which precedes it. Unfortunately, this level of simplicity is not possible in large vocabulary speech recognition tasks.

6) Neural Network Language Model

Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of word. Feature extraction, acoustic modeling and language modeling are important parts of modern speech recognition systems. Feature extraction involves analysis of speech signal. An acoustic model, created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for words.A statistical language model assigns a probability to a sequence of m words $P(W_1,W_2,W_3,\ldots..W_n)$ by means of a probability distribution. Clustering is one of the most useful language modeling techniques. Words can be grouped together into clusters then the probability of a cluster can be predicted instead of the probability of the word. The types of language models are unigram and n gram language models. The simplest form of language model simply throws away allconditioning context, and estimates each term independently. Such a model is called a unigram language model is given in (4):

$$P_{uni}(T_1,T_2,T_3,T_4) = P(T_1)P(T_2)P(T_3)P(T_4) \quad \ldots. (4)$$

There are many more complex kinds of language models, such as n gram language models, which condition on the previous term,

$$P_{bi}(T_1,T_2,T_3,T_4) = P(T_1).P(T_2\,|\,T_1).P(T_3\,|\,T_2).\,P(T_4\,|\,T \quad \ldots\ldots(5)$$

Neural network language model is given in (5) a language model based on Neural Networks , exploiting their ability to learn distributed representations to reduce the impact of the curse of dimensionality.

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.

Recurrent neural network-based language model (RNNLM) with applications to speech recognition is presented. Results indicate that it is possible to obtain around 50% reduction of perplexity by using mixture of several RNN LMs, compared to other language models. Back propagation through time (BPTT) is an efficient algorithm for training recurrent neural networks. Some of the language modeling toolkits are The CMU-Cambridge Statistical Language Modeling Toolkit, Sphinx Knowledge Base Tool, Random Forest ***Language Model Toolkit***, RNNLM - Recurrent Neural Network ***Language Modeling Toolkit***, SRILM - An Extensible Language Modeling Toolkit .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.

B. Comparison of Language Models

1) Number of distinct tokens: Number of distinct tokens is an important concept in the determination of the vocabulary size. It gives the minimum vocabulary size needed to cover 100% of the training data. In word-based model, the tokens are the words in the training text, in morpheme-based model the tokens are themorphemes, in stem-ending-based model the tokens are the stems and endings and in syllable-based model the tokens are the syllables. Firstly, the training speech corpus is divided into nine overlapping data sets. Each data set is generated by adding approximately one million words to the previous group. Then, token statistics are calculated on each data set. Figure9 shows the comparison of models in terms of number of distinct tokens. The curves for morpheme-based and stem-ending-based cases resemble each other. The slope difference between the curves corresponds to the new endings introduced

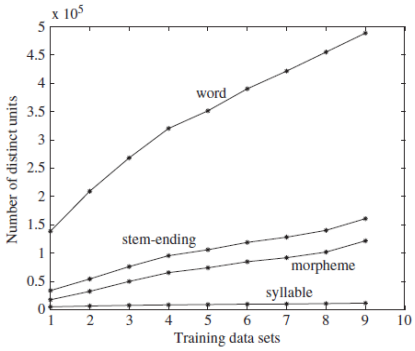

Figure9 Comparison of models in terms of number of distinct tokens.

III. CONCLUSION

January - February 2020
ISSN: 0193 - 4120 Page No. 16635 - 16640

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. we have also encountered a number of practical limitations which hinder a widespread deployment of application and services. In most speech recognition tasks; human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited.

Although these areas of investigations are important the significant advances will come from studies in acoustic phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving knowledge required for natural conversation. Although significant progress has been made in the last two decades, there is still work to be done, and we believe that a robust speech recognition system should be effective under full variation in: environmental conditions, speaker variability s etc. Speech Recognition is a challenging and interesting problem in and of itself. We have attempted in this paper to provide a comprehensive cursory, look and review of some language models used in speech recognition technology progressed. Speech recognition is one of the most integrating areas of machine intelligence, since, humans do a daily activity of speech recognition. Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction. We hope this paper brings about understanding and inspiration amongst the research communities of ASR.

REFERENCES

1. M.A.Anusuya, S.K.Katti, Speech Recognition by Machine: A Review,IJCSIS Vol. 6, No. 3,
2. S.Saraswathi,T.V.Geetha Design of language models at various phases of Tamil speech recognition system Vol. 2, No. 5, 2010, pp. 244-257
3. Ebru Arısoy_, Helin Dutag˘ acı, Levent M. Arslan, A unified language model for large vocabulary continuous speech recognition
4. Dat Tat Tran, Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.
5. Kanevsky, 1998, Statistical language model for inflectedanguages, US Patent No. 5,835,888,1998.
6. E. Mengusoglu, O. Deroo, Turkish LVCSR: Database preparation and language modeling for an agglutinativelanguage, in: Proceedings of the IEEE International Conferenceon Acoustics, Speech and Signal Processing, ICASSP 2001, Student Forum, Salt-Lake City, UT,