International Journal of Management, IT & Engineering
Vol. 8 Issue 6, June 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

SPEECH RECOGNITION AND HIDDEN MARKOV MODEL

Dr. Anuradha Kanade*

Abstract

Vision, Speech, and Natural Language are three core areas of Artificial Intelligence. Research in these areas has motivated and influenced research in other areas including multi-processing, parallel processing, robotics, and learning. Speech understanding and speech recognition are two related tasks. Understanding speech means getting the meaning of an utterance such that one can respond properly. Speech recognition is transcribing the speech without necessarily knowing the meaning of the utterance. Automatic speech recognition system has a long history of being difficult problem. Speech recognition is the field of artificial intelligence through which an acoustic waveform is converted into text. Hidden markov model (HMM) is very popular and widely used technique for speech recognition. This paper reviews briefly the HMM technique and highlights the benefits and issues related to it.

Keywords:

Automatic Speech Recognition; Hidden Markov Model; Transcribing,; Natural Language Processing; Dynamic Time Wrapping; Cepstral Vectors.

* **Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)**

1. Introduction

Automated Speech Recognition (ASR) is a difficult task. For a complete system, one needs several layers of processing: the signal pre-processing, feature extraction, phone recognition and syntax analysis. Currently, the most successful technique being used in ASR is the Hidden Markov Model (HMM). HMMs have shown good performance with tractable computation. There are other techniques tools that help in the process of speech recognition such as DTW, SAPI, NLP, Finite State transducers, neural network etc. Each of them has its own beauty, and it is the choice of the researcher to follow any one or combination of them. Thus they are the means and hence one can contribute to extend the existing technique or can find out innovation in the process for better results. [13]

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation.

2. Evolution of Speech Recognition

Single-speaker digit recognition system was first developed in Bell Labs in 1952. The system could locate the formants in the power spectrum of each word. [5]

The source-filter model of speech production was developed by Gunnar Fant and then published in 1960. In the late 1960s, Raj Reddy was the first person who designed a continuous speech recognition system that could issue spoken commands for the chess game. During the late 1960s, Soviet researchers invented the dynamic time wrapping algorithm (DTW). It was capable to recognize 200-word vocabulary [1]. The algorithm processed the speech signal by dividing it into short frames. This technique was useful for further research and improvement. Yet speaker independence was not achieved. IEEE Acoustics, Speech and Signal Processing group held a conference in Massachuetts in 1972 [11]. In late 1960s Leonard Baum developed the mathematics of Markov chain at the Institute for Defense Analysis. In 1970 James Baker and Janet Baker started using the Hidden Markov Model (HMM) for speech recognition [3]. This helped researchers to use in combination acoustic, language and syntax, in unified probabilistic model.

In 1980 IBM created a voice activated typewriter called Tangora that could handle the word vocabulary of 20,000 under the lead of Fred Jelinek [9]. Although HMM was too simple to account for many common features of human languages, it proved to be become the dominant speech recognition algorithm in the 1980s [j]. The progress in the field of speech recognition is due to the rapidly increasing capabilities of computers. In 1976 at the end of DARPA program, the best available computer for researcher was having PDA-10 and 4 MB RAM [4]. It took up to 100 minutes to decode only 30 second speech. [7]

With the advanced technology and computers researchers started to tackle more hard problems such as larger vocabulary, speaker independence, noisy environment and conversational speech. In particular, this shifting to more difficult tasks has characterized DARPA funding of speech recognition since the 1980s. For example, progress was made on speaker independence first by training on a larger variety of speakers and then later by doing explicit speaker adaptation during decoding. Further reductions in word error rate came as researchers shifted acoustic models to be discriminative instead of using maximum likelihood models. [8]

New speech recognition microprocessors were released in the mid-Eighties: for example RIPAC, independent-speaker recognition (for continuous speech) chip tailored for telephone services, was presented in the Netherlands in 1986. It was designed by CSELT/Elsag and manufactured by SGS [2].

3. Speech Recognition and HMM

The Automatic Speech Recognition (ASR) can be done with the help of Hidden Markov Model (HMM).

The ASR problem can be attacked from two sides; namely

1. From the side of speech generation

2. From the side of speech perception

The Hidden Markov Model (HMM) is a result of the attempt to model the speech generation statistically, and thus belongs to the first category above. During the past several years it has become the most successful speech model used in ASR. The main reason for this success is it's wonderful ability to characterize the speech signal in a mathematically tractable way. HMM is a statistical time series model in that is widely used in various fields especially in speech recognition system. It is used to recognize the time series sequence of speech parameters as digit, character, word, or sentence by using several refined algorithms of HMM. Text-to-speech synthesis systems to generate speech from input text information has also made substantial progress by using the excellent framework of HMM. [6] The speech input is separated using feature extraction system and with the help of HMM decoder the extracted words are understood by the application as shown in the following Figure 1.

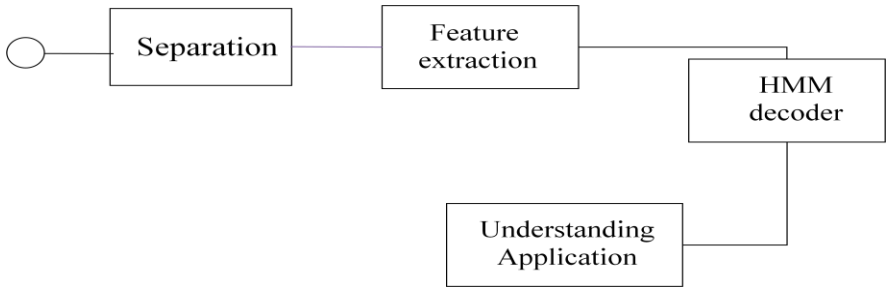

Figure 1. HMM in ASR

A hidden Markov model (HMM) is a finite state machine which generates a sequence of discrete time observations. At each time unit, the HMM changes states at Markov process in accordance with a state transition probability, and then generates observational data o in accordance with an output probability distribution of the current state. An N-state HMM is defined by the state transition probability

$A = \{a_{ij}\}_{i,j=1}^{N}$,

the output probability distribution

$B = \{b_i\ (o)\}_{i=1}^{N}$,

and initial state probability

$\prod = \{\prod_i\}_{i=1}^{N}$

For notational simplicity, we denote the model parameters of the HMM as follow:

$\lambda = (A, B, \Pi)$ ... (i)

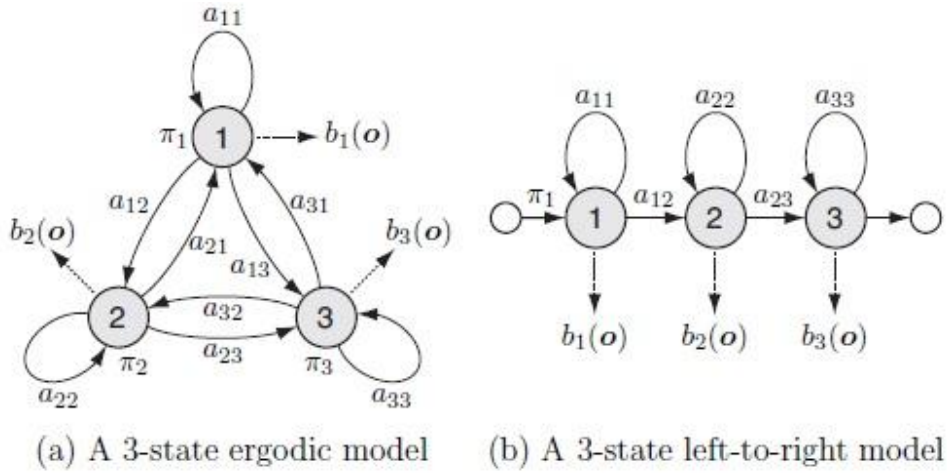

Figure 2. Examples of HMM structure

The above figure shows the HMM structure. The Figure 2 (a) shows the 3-state ergodic model, in which each state of the model can be reached from every other state of the model in a single transition, and Figure 2 (b) shows a 3-state left-to-right model, in which the state index simply increases or stays depending on time increment. The left-to-right models are often used as speech units to model speech parameter sequences since they can appropriately model signals whose properties successively change. The output probability distribution bi(o) of the observational data o of state i can be discrete or continuous depending on the observations. In continuous distribution HMM (CD-HMM) for the continuous observational data, the output probability distribution is usually modeled by a mixture of multivariate Gaussian distributions as follows.

$$bi(o) = \sum_{m=1}^{M} W_{im}\ N\left(o; \mu_{im}, \Sigma_{im}\right) \quad \text{... (ii)}$$

where M is the number of mixture components for the distribution, and wim, μim and Σim are a weight, a L-dimensional mean vector, and a L × L covariance matrix of mixture component m of state i, respectively. [6]

In a Hidden Markov model, the variables influenced by the state are visible. The state is not directly visible. Each state has a probability distribution over the possible output tokens. Hence the sequence of tokens generated by HMM gives information about the sequence of states. HMM can be considered a generalization of a mixture model where the hidden variables control the mixture component to be selected for each observation, are related through Markov process rather than independent of each other. HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. This uses the theory of statistics in order to arrange the feature vectors into a Markov matrix i.e. chains that stores probabilities of state transitions. HMM are simple and computationally feasible to use. Also they can be trained automatically. Hence they are more popular. HMM considers the speech signal as quasi-static for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another. HMM are simple networks that can generate speech , sequences of cepstral vectors, using a number of states for each model and modeling the short term spectra associated with each state with usually a mixture of multivariate Gaussian distributions. In other words, if each code word were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state. [12]

In the isolated mode one HMM for each of the speech unit is used. But in the continuous case this is not possible because a sequence of connected speech units, which is usually called a sentence, is to be recognized and hence the number of possible sentences may be prohibitively large even for a small vocabulary. In addition to this, there are two other fundamental problems associated with continuous recognition.

(1) End points of the speech units contained in the sentence are not known.

(2) Total numbers of speech units contained in the sentence are not known.

Because of the problems including those mentioned above, continuous recognition is more complicated than the isolated recognition.

However HMMs provide a good frame work for continuous mode of speech recognition.
Modern general-purpose speech recognition systems are generally based on hidden Markov models (HMMs). This is a statistical model which outputs a sequence of symbols or quantities.
One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piece-wise stationary signal or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Speech could thus be thought as a Markov model for many stochastic processes (known as states).

Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, to give the very simplest setup possible, the hidden Markov model would output a sequence of n-dimensional real-valued vectors with n around, say, 13, outputting one of these every 10 milliseconds. The vectors, again in the very simplest case, would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short -time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have, in each state, a statistical distribution called a mixture of diagonal covariance Gaussians which will give likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes. [10]

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phones (so phones with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker

normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semitied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques which dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE).

Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Marko model which includes both the acoustic and language model information, or combining it statically beforehand (the finite state transducer, or FST, approach). Like the HMM model there are some other models and algorithms developed which assist the speech recognition process mathematically and or statistically.

The most popular and widely used model is HMM. Its use has become more saturated these days.

Some researchers are trying to find alternating models for this which will definitely contribute to the accuracy in the results produced during the process.

HMM and Neural network-based speech recognition

Another approach in acoustic modeling is the use of neural networks. They are capable of solving much more complicated recognition tasks, but do not scale as well as HMMs when it comes to large vocabularies. Rather than being used in general -purpose speech recognition applications they can handle low quality, noisy data and speaker independence. Such systems

can achieve greater accuracy than HMM based systems, as long as there is training data and the vocabulary is limited. A more general approach using neural networks is phoneme recognition. This is an active field of research, but generally the results are better than for HMMs. There are also NN-HMM hybrid systems that use the neural network part for phoneme recognition and the hidden markov model part for language modeling. [14]

4. Conclusions

The above paper briefly takes an overview of various tools and techniques used in the process of Speech Recognition. The HMM model is most popular statistical model. It is called hidden because underlying walk between the states is hidden and only the symbols emitted by the system are observable. The HMM has strong statistical base and has efficient learning algorithms. It can handle variable length input. It can handle variations in record structure with optional fields and varying field ordering.

It has a wide variety of applications including multiple alignment, data mining and classification, structural analysis, and pattern discovery.

There are some limitations as well. HMM requires training using annotated data. It is not completely automatic and may require manual markup. The size of training data may be an issue. The work on Natural language processing and the neural networks which are the branches of AI also supports speech recognition. It is also clear that continuous speech recognition is more complicated than the isolated recognition. However HMMs provide a good frame work for continuous mode of speech recognition feature extraction, feature optimization using HMM and classification of extracted feature is done using ANN. The recognition rate for the isolated words is improved by using hybrid HMM/ANN technology for speech recognition.

5. References

[1]. Benesty, Jacob, Sondhi, M. M., Huang, Yiteng (2008), Springer Handbook of Speech Processing. Springer Science & Business Media. ISBN 3540491252.

[2]. Cecinati R, Ciaramella A, Venuti G, Vicenzi C (February 1987), "A Custom Integrated Circuit with Dynamic Time Warping for Speech Recognition", CSELT Technical Reports, 15 (1)

[3]. Huang Xuedong, Baker James, Reddy Raj, "A Historical Perspective of Speech Recognition", Communications of the ACM, Retrieved 20 January 2015

[4]. http://ethw.org/First-Hand:The_Hidden_Markov_Model, Retrieved 23 January 2018

[5]. Juang, B. H.; Rabiner, Lawrence R. "Automatic speech recognition–a brief history of the technology development" (PDF): 6. Retrieved 17 January 2015.

[6]. Junichi Yamagishi, October 2006, "An Introduction to HMM-Based Speech Synthesis"

[7]. McKean, Kevin (8 April 1980), "When Cole talks, computers listen", Sarasota Journal AP, Retrieved 23 November 2015.

[8]. Morgan Nelson, Cohen Jordan, Krishnan Sree Hari, Chang S, Wegmann S., (2013), "Final Report- OUCH Project (Outing Unfortunate Characteristics of HMMs)", CiteSeer X 10.1.1.395.7249

[9]. "Pioneering Speech Recognition", Retrieved 18 January 2015.

[10]. Pravin Yannawar, 2010, "A Review on Speech Recognition Technique", International Journal of Computer Applications, November 2010

[11]. Rabiner (1984),"The Acoustics, Speech, and Signal Processing Society- A Historical Perspective" (PDF), Retrieved 23 January 2018.

[12]. Rupali S Chavan, Dr. Ganesh. S Sable, "An Overview of Speech Recognition Using HMM", International Journal of Computer Science and Mobile Computing Vol.2 Issue. 6, June-2013, pg. 233-238

[13]. Shaheena Sultana , M. A. H. Akhand , Prodip Kumer Das , M. M. Hafizur Rahman, "Bangla Speech-to-Text conversion using SAPI", 3-5 July 2012

[14]. Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications, Volume 10– No.3, November 2010