Hidden Markov Models for Automatic Speech Recognition

Mbarki Aymen, Ammari Abdelaziz, Sghaier Halim, Hassen Maaref
Laboratory of Micro-Optoélectronique and Nanostructure, Faculty of sciences Monastir, Tunisia
mbarki.aymen@yahoo.fr

***Abstract*: In this paper we look into the problem of Hidden Markov Models (HMM): the evaluation, the decoding and the learning problem. We have explored an approach to increase the effectiveness of HMM in the speech recognition field. Although hidden Markov modeling has significantly improved the performance of current speech-recognition systems, the general problem of completely fluent speaker-independent speech recognition is still far from being solved. For example, there is no system which is capable of reliably recognizing unconstrained conversational speech. Also, there does not exist a good way to infer the language structure from a limited corpus of spoken sentences statistically. Therefore, we want to provide an overview of the theory of HMM, discuss the role of statistical methods, and point out a range of theoretical and practical issues that deserve attention and are necessary to understand so as to further advance research in the field of speech recognition.**

Keywords: Hidden Markov Models (HMMs), Speech recognition, HMM problems, Viterbi algorithm.

I. Introduction:

Speech recognition field is one of the most challenging fields that have faced the scientists from long time. The complete solution is beyond reach. The efforts have been concentrated with huge funds from the companies to different related and supportive approaches to reach the final goal.

The term HMM is now quite familiar in the speech signal processing community and is gaining acceptance for communication systems.

HMMs are the most successful techniques used in automatic speech-recognition (ASR) systems [1]. At the hidden level, however, the most commonly ASR systems represent only phonetic information about the underlying speech signal. Although there has been much success using this methodology, the approach does not explicitly incorporate knowledge of certain important aspects of human speech production [2].

We begin to present an overview of the field of speech recognition. Then, we mention the benefits of HMMs and their uses. Afterwards, we give the problems of these models as well as proposing solutions to them.

II. Speech recognition:

Speech recognition is a very important aspect. Particularly, with the computer widely used, it appears to be even more significant like voice driven service portals, speech interfaces in automotive navigation and guidance systems or speech driven applications in modern offices. Speech recognition will play an important role in future human–computer interfaces. In general, the field of speech recognition is a part of the ongoing research effort in developing computers that can hear and understand spoken information [3].

Speech recognition is a broad term which means that it can recognize almost anybody's speech - such as a call-centre system designed to recognize many voices. Voice recognition is a system trained to a particular user, where it recognizes their speech based on their unique vocal sound. Speech recognition is the process of converting an acoustic signal, captured by a microphone, to a set of words. The first speech recognizer appeared in 1952 and consisted in a device for the recognition of single spoken digits [4]. Another

early device was the IBM Shoebox, exhibited at the 1964 New York World's Fair. The research was subsequently considerably increased during the 70s with the work of Jelinek at IBM (1972-1993). Today, speech recognition is an area of strong growth thanks to the flood of embedded systems.

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with a word error rate (WER), whereas speed is measured with the real time factor. The dominant recognition paradigm in the past fifteen years has been known as HMMs.

In general, speech recognition systems have stored reference templates of phonemes or words whose input speech is compared and the closest word or phoneme is given out. Since it is the frequencies (at which energy is high) that are to be compared, the spectra of the input and reference template are compared rather than the actual waveform. Acoustic and phonetic models allow taking into account the constraints and acoustic phonetic level of a sound, while the language models define the constraints of higher levels of the language used. Finally, the decoding process is to find the optimal path in the graph of all possible sentences. The sentence is constructed as a sequence of words.

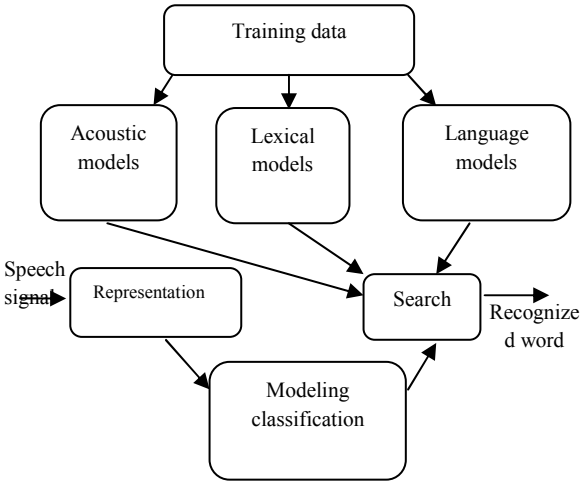


Figure 1: Architecture of speech recognition system

III. The Hidden Markov Models:

Markov chains have increasingly become a useful way of capturing stochastic nature of many economic and financial variables [5]. Although the hidden Markov processes have been widely employed for some time in many engineering applications like speech recognition, its effectiveness has now been recognized in areas of social science research as well.

HMMs are the dominant technology used for speech recognition. HMMs provide a very useful paradigm to model the dynamics of speech signals [6] [7].

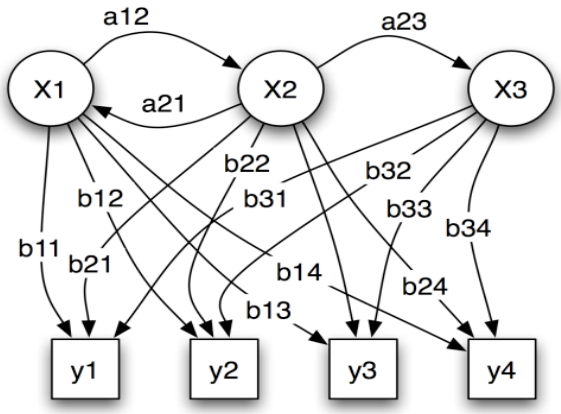


Figure 2: Probabilistic parameters of a hidden Markov model

They provide a solid mathematical formulation for the problem of learning HMM parameters from speech observations. Furthermore, efficient and fast algorithms exist for the problem of computing the most likely model given a sequence of observations. The HMM is basically a stochastic finite set of states, where each one is associated with a probability distribution (generally multidimensional) (figure 2). Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state, an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state, which is visible to an external observer and therefore states are hidden to the outside, hence the name HMMs. One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piecewise stationary signal

or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds that speech could be approximated as a stationary process. Speech could thus be thought of as a Markov model for many stochastic processes. Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use [8].

In order to define an HMM completely, the following elements are needed:

- The number of states of the model *N*.
- The number of observation symbols in the alphabet *M*. If the observations are continuous then *M* is infinite.
- A set of state transition probabilities A= $\{a_{ij}\}$.

$$a_{ij} = p\{q_{t+1} = j \setminus q_t = i\}, \quad 1 \le i, j \le N, \qquad (1)$$

where q_t denotes the current state.

- Probability distribution in each state: B= $\{b_j(k)\}$.

$$b_j(k) = p\{o_t = v_k \setminus q_t = j\},$$
$$1 \le j \le N,\ 1 \le k \le M \qquad (2)$$

where v_k denotes the k^{th} observation symbol in the alphabet and o_t the current parameter vector.

- The initial state distribution $\pi = \{\pi_i\}$.

$$\pi_i = p\{q_1 = i\}, \quad 1 \le i \le N \qquad (3)$$

- Therefore we can use the compact notation $\lambda = (A,B,\pi)$.

IV. The HMM problems:

Most applications of HMMs are finally reduced to solving three main problems. First, the evaluation problem given an HMM λ and a sequence of observations $O=o_1,o_2,\ldots,o_t$ where there is a probability that the observations are generated by the model $p\{O|\lambda\}$. Second, the decoding problem given a model λ and a sequence of observations $O=o_1,o_2,\ldots,o_t$ where the most likely state sequence in the model produces the observations. The third problem is the learning problem given a model λ and a sequence of observations $O=o_1,o_2,\ldots,o_t$ where we should adjust the model parameters $\lambda=(A,B,\pi)$ in order to maximize $p\{O|\lambda\}$. Thereafter, we describe several conventional solutions to these three standard problems

1. Evaluation problem:

We have a sequence of observations $O=o_1,o_2,\ldots,o_t$, a model $\lambda=(A,B,\pi)$ and $p\{O|\lambda\}$. We can calculate this quantity using simple probabilistic arguments. But this calculation involves a number of operations in the order of N^T. This is very large even if the length of the sequence *T* is moderate. Therefore we have to look for another method for this calculation. Fortunately there exists one which has a considerably low complexity and makes use of an auxiliary variable $\alpha_t(i)$, called forward variable.

a) Forward Procedure:

Initialization for $1 \le i \le N$

$$\alpha_1(i) = \pi_i f_i(O_1) \qquad (4)$$

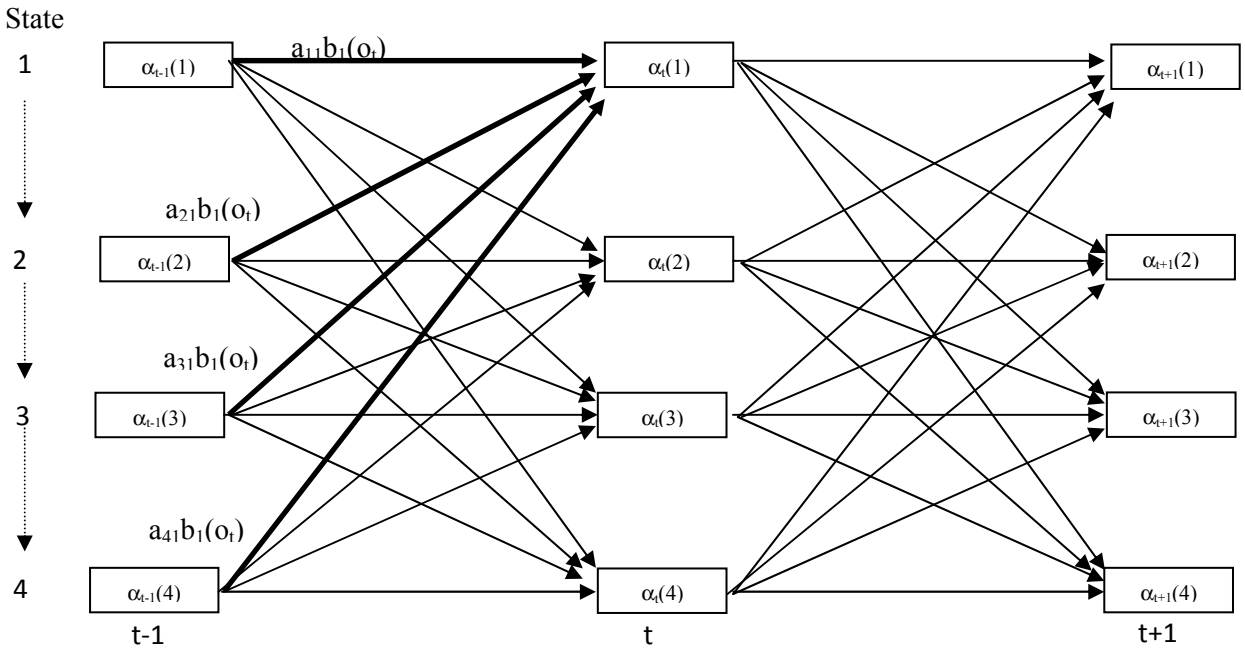
Forward Recurrence for $t = 1,2,\ldots,T-1; 1 \le j \le N$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] f_j(O_{t+1}) \qquad (5)$$

Probability computation

$$p(O \setminus \lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad (6)$$

A trellis diagram can be used to visualize probability computation of HMMs. Figure 3 shows such a diagram for a HMM with 4 states. Each column in the trellis shows the possible states at time t. Each state in one column is connected to each state in the adjacent columns by the transition probability given by the elements a_{ij} of the transition matrix A.



$$\alpha_t(1)= \alpha_{t-1}(1)\, a_{11}b_1(o_t) + \alpha_{t-1}(2)\, a_{21}b_1(o_t) + \alpha_{t-1}(3)\, a_{31}b_1(o_t) + \alpha_{t-1}(4)\, a_{41}b_1(o_t)$$

Figure 3: A Trellis Structure for the Calculation of the Forward Partial Probabilities $\alpha_t(i)$

b) Backward Procedure:

Initialization for $1 \le i \le N$

$$\beta_T(i) = 1 \qquad (7)$$

Forward Recurrence for
$t = T-1, T-2, ..., 1; 1 \le i \le N$

$$\beta_t(i) = \sum_{j=1}^{N} \alpha_{ij} f_j(O_{t+1}) \beta_{t+1}(j) \qquad (8)$$

Probability calculation

$$p(O \setminus \lambda) = \sum_{i=1}^{N} \pi_i f_i(O_1) \beta_1(i) \qquad (9)$$

2. Decoding problem:

In this case we want to find the most likely state sequence for a given sequence of observations $O=o_1,o_2,\ldots,o_t$ and a model $\lambda=(A,B,\pi)$. The solution to this problem depends on the way the most likely state sequence is defined. One approach is to find the most likely state q_t at t=t and to concatenate all such q_t. But sometimes this method does not give a physically meaningful state sequence. Therefore we would go for another method which does not have such problems. In this method, commonly known as Viterbi algorithm [9] [10], the whole state sequence with the maximum likelihood is found. In order to facilitate the computation we define an auxiliary variable. It is an inductive algorithm which keeps at each instant the best possible state sequence for each of the N states as the intermediate state for the desired observation sequence $O=o_1,o_2,\ldots,o_t$. In this way we finally have the best path for each of the N states as the last state for the desired observation sequence. Out of these, we select the one which has the highest probability.

1. Initially

$$\delta_1(i) = \pi_i b_i(o_1) \quad for \quad 1 \le i \le N \qquad (10)$$
$$\psi_1(i) = 0$$

2. $For \quad t = 2, ..., T$

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t)$$
$$\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i) a_{ij}\right]$$
$$for \quad 1 \le j \le N \qquad (11)$$

3. Finally

$$\Delta^* = \max_i [\delta_T(i)] \qquad (12)$$

$$x_T^* = \arg\max_i [\delta_T(i)] \qquad (13)$$

4. Trace back $for \quad t = T-1, T-2, ..., 1$

$$x_t^* = \psi_{t+1}(x_{t+1}^*), \qquad \text{and } X^* = \{x_1^*, x_2^*, ..., x_T^*\} \qquad (14)$$

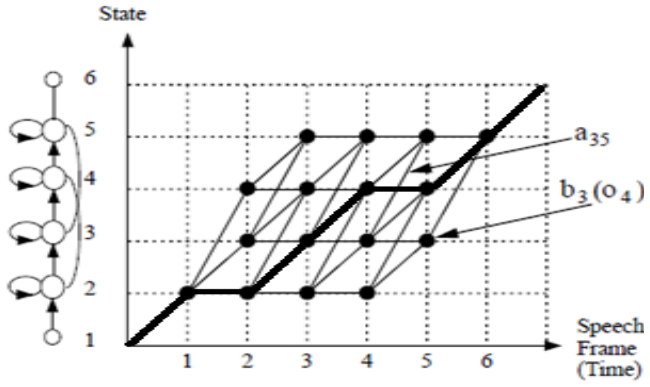


Figure 4: Recursive determination of the shortest path via the Viterbi algorithm

We can see in figure 4 through this trellis the path made by the Viterbi algorithm to try to find the sequence that has been issued. We take the path that has the smallest Hamming distance.

3. Learning problem:

Generally, the learning problem is how to adjust the HMM parameters, given an observation sequence O, where the estimation problem involves finding the right model parameter values that specify the most likely model to produce the given sequence. Thus it would be clear that the quantity we wish to optimize during the learning process can be different from one application to another. In other words, there may be several optimization criteria for learning, out of which a suitable one is selected depending on the application. In speech recognition, this is often called training, and the given sequence, on the basis of which we obtain the model parameters, is called the training sequence, even though the formulation here is statistical. There are two main optimization criteria found in ASR literature. Firstly, maximum likelihood (ML) [11] when we try to maximize the probability of a given sequence of observations this probability is the total likelihood of the observations and can be expressed mathematically as: $L_{tot} = \{O|\lambda\}$. However there is no known way to analytically solve the model $\lambda = (A,B,\pi)$ which maximizes the quantity L_{tot}. But we can choose model parameters locally maximized, using an iterative procedure, like Baum-Welch method or a gradient based method [12]. In ML we optimize an HMM of only one class at a time, and do not touch the HMMs for other classes at that time. This procedure does not involve the concept discrimination which is of great interest in Pattern Recognition and of course in ASR. Thus the ML learning procedure gives a poor discrimination ability to the HMM system, especially when the estimated parameters (in the training phase) of the HMM system do not match with the speech inputs used in the recognition phase. This type of mismatches can arise due to two reasons. One is that the training and recognition data have considerably different statistical properties, and the other is the difficulties of obtaining reliable parameter estimates in the training. Secondly, maximum mutual information (MMI) considers HMMs of all the classes simultaneous, during training. Parameters of the correct model are updated to enhance its contribution to the observations, while parameters of the alternative models are updated to reduce their contributions. This procedure gives a high discriminative ability to the system and thus MMI belongs to the so called discriminative training category.

V. Results:

A sequence of words M is enunciated by a speaker, then it is processed by the acoustic processor to produce a sequence of acoustic vectors. These will be used to generate a sequence of acoustic observation O calculated by standard techniques of signal processing. Finally, a decoding module (recognition) converts this sequence of acoustic observations O with a sequence of word results. Figure 5 shows the results of applying an

HMM model to determine the best road recognition of word sequence. The log likelihood value increases from one state to another sta.

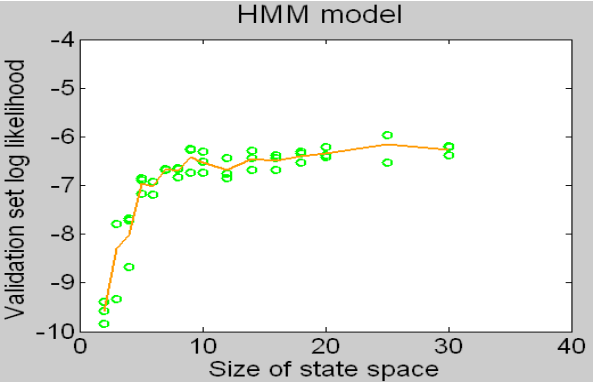


Figure 5: Road recognition with HMM model

VI. Conclusions:

In this paper, a brief description of the HMMs is presented. We have proposed theoretically analyzed algorithms that solve three problems of HMMs. The first problem needs the forward backward algorithm, and the second problem requests the Viterbi algorithm for selecting the highest probability of observations. Finally, the Baum Welch algorithm adjusts the HMM parameters and involves finding the right model parameter values that specify the most likely model to produce the given sequence.

VII. References:

[1] Rabiner, L.R., "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proceedings of the IEEE, 77, No.2, 1989, pp.257-286.

[2] Young, S., "A review of large-vocabulary continuous-speech recognition," IEEE Signal Processing Magazine, 13, No.5, 1996, pp.45-57.

[3] Ganapathiraju A, Hamaker J, Picone J, Ordowski M, Doddington G R. "Syllable-based large vocabulary continuous speech recognition". IEEE Transactions on Speech and Audio Processing, 2001, 9(4): 358–366

[4] Davies, K.H., Biddulph, R. and Balashek, S. (1952) *Automatic Speech Recognition of Spoken Digits*, J. Acoust. Soc. Am. 24(6) pp.637 – 642.

[5] Rabiner L, Juang B. An introduction to hidden Markov Models. *IEEE ASSP Magazine,* Jan., 1986, pp.4-16.

[6] K. Lee and H. Hon Speaker-independent phone recognition using hidden Markov models, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, No. 11, Nov. 1989.

[7] L.R. Rabiner, Mathematical foundations of hidden Markov models, in H. Niemann. M. Lang and G. Sagerer (eds.), Recent Advances in Speech Understanding and Dialog Systems, Vol. F46 of NATO ASI Series, Springer, Berlin. 1988, pp. 183-205.

[8] Sameh M. Awaidah, Sabri A. Mahmoud, "A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models", International Journal of Signal Processing 89 (2009) 1176–1184

[9] G. D. Forney Jr. "The Viterbi Algorithm," *IEEE Proceedings,* IT-61(3):268-278, March 1973.

[10] N. Seshadri and C-E. W. Sundberg. "List Viterbi Algorithms with applications." IEEE Transactions on Communications. Vol. 42, No. 2/3/4, Feb/Mar/Apr 1994.

[11] Juang B H, Levinson S, Sondhi M."Maximum likelihood estimation for multivariate mixture observations of Markov chains". IEEE Transactions on Information Theory, 1986, 32(2): 307–309

[12] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1235–1249, 1985.