Review on Training Hidden Markov Models with Multiple Observations

Bohumir Jelinek

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762 USA
email: jelinek@isip.mstate.edu

ABSTRACT

Hidden Markov models (HMMs) are stochastic models able to represent sequential signals. They are widely and successfully used in speech recognition and recently also in handwriting recognition applications. As each modeling techniques also HMMs impose some constraint on the modeled signal. Feature independence assumption is considered to be one of the major drawbacks of the hidden Markov modeling approach. Lot of research effort is devoted to relax this assumption. Reviewed article presents a theoretical justification of the multiple observation HMM training method that does not impose the independence assumption. The derived training algorithm is proven to guarantee a convergence of the training process. Training equations constrained by the feature independence assumption are shown to be a special case. In this paper we will make a theoretical review of the suggested method and its usability for a large vocabulary continuous speech recognition.

1. Introduction

HMM techniques are becoming more and more popular in various research and application areas. They are used to model divers stochastic sequential signals. Current research effort is devoted primary to the area of speech and handwritten letters recognition.

We always use some kind of preprocessing to get the significant features of the signal. Since a characteristic of the speech signal is extremely heterogeneous, we need a lot of parameters to describe each specific feature distribution. To describe the static features in the speech use an observation probability densities modeled by Gaussian mixture distributions. Temporal nature of the speech signal is characterized by the transitional structure (left-right topology) of the HMM. Also some specific models allowing back-transition are designed. They can be used to model e.g. long lasting silence.

The hidden Markov model parameters estimation problem is not trivial and usually requires some underlying assumptions to be valid. If these assumption are not realistic, they can significantly decrease the system learning and recognition performance. The reviewed paper is concentrated on observation independence assumption during the training process. Theoretical justification of a new training method not imposing the observation independence constrain is provided.

The standard hidden Markov model elements and the forward-backward observation evaluation procedure are described in the first part of the article. The derivation of the standard Baum-Welch model training algorithm is reviewed consecutively.

Then a new technique (combinatorial method for multiple observation training) is suggested. The observation sequence probability is revealed in terms of conditional probabilities, thus avoiding the assumption of statistically uncorrelated observation sequences.

The training equations for two special cases are derived at the end of the article. One special case assumes observations independence. The authors proved that this special case leads to a standard Levinson training equation. The second special case assumes uniform dependence of training observation. Special training eqauation is derived on the base of this assumption.

2. Article Review

Short history of the hidden Markov model studies is reviewed in the Section 1 of the article. The main application areas are also outlined. The authors then introduce a new approach utilizing observation independence assumption. The main benefit of the approach is implied freedom in training equations derivation. At the end of the introductory section the paper organization is introduced.

2.1. First Order Hidden Markov Model - Standard approach

The traditional first order hidden Markov model elements are reviewed in the Subsection 2.1. Two standard topologies - ergodic and left-right model are introduced in the Subsection 2.2. Subsection 2.3 defines Forward-Backward Procedure for observation evaluation.

Baum-Welch model training algorithm derivation using Baum auxiliary function is briefly outlined in the Subsection 2.4. The Baum-Welch training algorithm is a useful learning algorithm based on the Expectation-Maximization theorem. Then a basic state and transition parameter updating equations are expressed in terms of joint events and state variables.

2.2. Multiple Observation Training - Combinatorial Method

Now we introduce a combinatorial HMM training method that does not impose the condition of statistical independence on training samples.

Lets O is an observation sequence of a particular pattern class

$$O = \{O^{(1)}, O^{(2)} \ldots, O^{(K)}\} \tag{1}$$

where

$$O^{(k)} = \left\{ O_1^{(k)}, O_2^{(k)} \ldots, O_{T_K}^{(k)} \right\} \text{ for } 1 \le k \le K \tag{2}$$

are the individual observation sequences of the pattern class.

If the observation sequences are independent, we can compute the joint observation density as:

$$P(O|\lambda) = \bigcap_{k=1}^{K} P(O^{(k)}|\lambda) \tag{3}$$

In case we assume that observation sequences depend on one another, authors suggest using the following expressions for $P(O|\lambda)$:

$$\begin{aligned}
&P(O^{(1)}|\lambda)P(O^{(2)}|O^{(1)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(K)}|O^{(K-1)}\ldots, O^{(1)}\lambda) \\
&P(O^{(2)}|\lambda)P(O^{(4)}|O^{(3)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(1)}|O^{(K)}\ldots, O^{(2)}\lambda) \\
&\ldots \\
&(P(O^{(K)}|\lambda)P(O^{(1)}|O^{(K)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(K-1)}|O^{(K-2)}\ldots, O^{(1)}, O^{(K)}\lambda)
\end{aligned} \tag{4}$$

These equations do not introduce any additional constraint on the task.

On the base of the above equations the multiple observation probability can be expressed as:

$$P(O|\lambda) = \sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \tag{5}$$

where weights can be determined from the conditional probabilities:

$$\begin{aligned}
w_1 &= \frac{1}{K} P(O^{(2)}|O^{(1)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(K)}|O^{(K-1)}\ldots, O^{(1)}\lambda) \\
w_2 &= \frac{1}{K} P(O^{(4)}|O^{(3)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(1)}|O^{(K)}\ldots, O^{(2)}\lambda) \\
&\ldots \\
w_K &= \frac{1}{K} (P(O^{(1)}|O^{(K)}, \lambda)\ldots \\
&\ldots\ldots P(O^{(K-1)}|O^{(K-2)}\ldots, O^{(1)}, O^{(K)}\lambda)
\end{aligned} \tag{6}$$

3. Training Equation Derivation

Baum auxiliary function is introduced like a summation of Baum auxiliary functions for individual observations:

$$Q(\lambda|\bar{\lambda}) = \sum_{k=1}^{K} w_k Q_k(\lambda|\bar{\lambda}) \tag{7}$$

Then an unconstrained training equation is derived by the Lagrange multiplier method. The following training equations are obtained (very similar to the standard - observation independence assuming) equations.

1.) state transition probability:

$$\bar{a}_{mn} = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_K-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_K-1} \gamma_t^{(k)}(m,n)} \tag{8}$$

for $1 \le m \le N$, $1 \le n \le N$

2.) symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1,\, o_t^k = v_m}^{T_K} \gamma_t^{(k)}(n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \sum_{t=1}^{T_K} \gamma_t^{(k)}(n)} \tag{9}$$

for $1 \le m \le M$, $1 \le n \le N$

3.) initial state probability:

$$\bar{\pi}_n = \frac{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda) \gamma_1^{(k)}(n)}{\sum_{k=1}^{K} w_k P(O^{(k)}|\lambda)} \tag{10}$$

for $1 \le n \le N$.

3.1. Special Cases

For the independent observation case we obtain Levinson training equation. For the uniform dependence assumption we obtain:

1.) state transition probability:

$$\bar{a}_{mn} = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_K-1} \xi_t^{(k)}(m,n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_K-1} \gamma_t^{(k)}(m,n)} \tag{11}$$

for $1 \le m \le N$, $1 \le n \le N$

2.) symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1,\, o_t^k = v_m}^{T_K-1} \gamma_t^{(k)}(n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \sum_{t=1}^{T_K} \gamma_t^{(k)}(n)} \tag{12}$$

for $1 \le m \le M$, $1 \le n \le N$

3.) initial state probability:

$$\bar{\pi}_n = \frac{\sum_{k=1}^{K} P(O^{(k)}|\lambda) \gamma_1^{(k)}(n)}{\sum_{k=1}^{K} P(O^{(k)}|\lambda)} \tag{13}$$

for $1 \le n \le N$.

4. Critical Objections

The problem how to get the conditional observation probabilities is not solved in the article. No one is able to get the modified training equations without knowing weights values. This is the major disadvantage of the suggested approach.

Another drawback of the article is that there is no experimental verification experiment provided. The experimental verification should be provided at least for the case of the uniform observation sequence

dependence assumption. In this case there is only a minor difference between the resulting training equation and the commonly used training equation.

5. Contributions of Suggested Technique

The main contribution of the suggested technique is modified parameter estimation equation for hidden Markov model training that does not impose constraining observation independence assumption.

The HMM training theory is enriched by the derivation of training equation for the case of a given conditional dependence of class observation sequence.

Practical application is not obvious. The major obstacle is determination of conditional observation sequence probabilities, i.e. weights in derived training equation.

6. Summary

Model topology primary used in the area of speech recognition is left-to-right (also called Bakis) hidden Markov model. We can not train such a model with only one observation sequence available, because the model has many parameters and we can not estimate them reliably from just one observation sequence per model. We are always using multiple observation sequences in speech recognition applications.

The application of the suggested technique in the area of large vocabulary continuous speech recognition system is not manageable, because we are not able to separate observation sequences for particular classes. Speech recognition is a hierarchical problem. It is necessary to identify data classes on different levels - phones, words and sentences.

In fact the dependence between multiple training observation for a particular class in speech recognition applications is not observable. The training data are supposed to be collected randomly. It is necessary to make this assumption valid, because otherwise speech training database is corrupted. The models resulting from the training can not be used for the recognition of the independent test set.

7. Acknowledgments

I would like to thank to my wife Veronika for encouraging me to work and for bearing all of my problems together with me.

I give many thanks to John for coordination of my effort against the procrastination and also to my supervisor Dr. Joseph Picone for forcing me to do painful steps to increase the quality of my personality and life.

References

[1] L.Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[2] L.Rabiner, B.H.Juang, Fundamental of Speech Recognition, Englewood Cliffs, N.J.:Prentice Hall, 1993.

[3] S.M. Ross, Introduction to Probability Models, Seventh Edition, Academic Press, 2000.

[4] L.R. Bahl, F. Jelinek, and R.L. Mercer, ªA Maximum Likelihood Approach to Continuous Speech Recognition,º IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 5, pp. 179-190, 1983.

[5] F.Jelinek, Statistical Methods for Speech Recognition, MITT Press, 1997.

[6] R.E.Duda, P.E.Hart, D.G.Stork, Pattern Classification, Second Edition, John Wiley & Sons, 2001.

[7] L. Xiaolin, M. Parizeau, R. Plamondon, “Training hidden Markov models with multiple observations - a combinatorial method,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 4, pp. 371-377, April 2000.

[8] J. Hu, M.K., Brown, W. Turin, ªHMM Based On-Line Handwriting Recognition,º IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 10, pp. 1,039-1,045, Nov. 1996.