Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer

Continuous Speech Recognition Group
Department of Computer Sciences
IBM Thomas J. Watson Research Center
P O Box 218, Yorktown Heights, NY 10598

Abstract

A method for estimating the parameters of hidden Markov models of speech is described. Parameter values are chosen to maximize the mutual information between an acoustic observation sequence and the corresponding word sequence. Recognition results are presented comparing this method with maximum likelihood estimation.

I. Introduction

Traditionally, a speech recognition system consists of an *acoustic processor* and a *linguistic decoder*. The acoustic processor receives as input a speech waveform which is the acoustic realization of a sequence $\tilde{w}$ of words spoken by the user, and it produces as output a sequence y of salient features. These features may be symbols from a discrete alphabet, such as phonetic labels, or may be vectors of continuous parameters, such as Fourier coefficients. The task of the linguistic decoder is to decode $\tilde{w}$ from y. It translates the feature sequence y produced by the acoustic processor into an estimate $\hat{\tilde{w}}$ of the speaker's original word string $\tilde{w}$.

To minimize the probability of error, $\hat{\tilde{w}}$ must be chosen so that

$$P(\boldsymbol{W} = \hat{\tilde{w}} \mid \boldsymbol{Y} = y) = \max_{w} P(\boldsymbol{W} = w \mid \boldsymbol{Y} = y). \tag{1}$$

By Bayes' rule

$$P(\boldsymbol{W} = w \mid \boldsymbol{Y} = y) = \frac{P(\boldsymbol{W} = w)P(\boldsymbol{Y} = y \mid \boldsymbol{W} = w)}{P(\boldsymbol{Y} = y)}. \tag{2}$$

Here $P(\boldsymbol{W} = w)$ is the *prior* probability that the word sequence w will be spoken, and $P(\boldsymbol{Y} = y \mid \boldsymbol{W} = w)$ is the probability that y will be produced by the acoustic processor given that the speaker uttered the word sequence w. $P(\boldsymbol{Y} = y)$ is not a function of w, and therefore need not concern us during recognition.

To calculate $P(\boldsymbol{Y} = y \mid \boldsymbol{W} = w)$ the linguistic decoder requires a probabilistic model of the speaker's phonological and acoustic-phonetic variations, and of the performance of the acoustic processor. In Section II we describe one such class of models, hidden Markov models. In Section III we review an algorithm for computing maximum likelihood estimates of their parameters. In Section IV we present an alternative method of parameter estimation, MMI, which seeks to maximize the mutual information between y and $\tilde{w}$. In Section V we mention a number of ways that the simple models in the previous sections may be extended. In Section VI we discuss the application of hidden Markov models to speech recognition. Finally, in Section VII we compare the performance of MMI estimates with maximum likelihood estimates.

II. Hidden Markov Models

A sequence of random variables $\boldsymbol{X} = X_1 X_2 \ldots$ is an n-*state Markov chain* provided each of the random variables X_t ranges over the integers 1 to n, and further,

$$P(X_t = x_t \mid X_1^{t-1} = x_1^{t-1}) = P(X_t = x_t \mid X_{t-1} = x_{t-1}). \tag{3}$$

A sequence of random variables $\boldsymbol{Y} = Y_1 Y_2 \ldots$ is a probabilistic function of a Markov chain $\boldsymbol{X}$, or, equivalently, $\boldsymbol{X}$ is a *hidden Markov model* for $\boldsymbol{Y}$, if for each $1 \leq t < T$,

$$P(Y_t = y_t \mid Y_1^{t-1} = y_1^{t-1}, X_1^T = x_1^T) = P(Y_t = y_t \mid \boldsymbol{X}_t = x_t, X_{t+1} = x_{t+1}). \tag{4}$$

The sequence of values $y = y_1 y_2 \ldots$ is the *output sequence* of the hidden Markov model. We define the *transition probability*, a_{ij}, the *output probability*, b_{ijy}, and the *initial state probability*, c_i, by

$$a_{ij} = P(X_t = j \mid X_{t-1} = i), \tag{5}$$

$$b_{ijy} = P(Y_t = y \mid X_t = i,\ X_{t+1} = j), \tag{6}$$

$$c_i = P(X_1 = i). \tag{7}$$

We can consider these probabilities as the components of a vector, θ, which we refer to as the parameter vector of the hidden Markov model. Because the components of θ are probabilities, they must all be non-negative. They must also satisfy the constraints

$$\sum_j a_{ij} = 1, \quad \sum_y b_{ijy} = 1, \quad \sum_i c_i = 1. \tag{8}$$

Sometimes, we may require θ to satisfy constraints in addition to (8). We may, for example, require that certain of the transition probabilities be zero. One particularly important constraint is *tying*. Two states, i and j, are *tied* if there exists a permutation of the states, $\pi: k \to \pi(k)$, such that $a_{ik} = a_{j\pi(k)}$, for all k. Similarly, two transitions, $i \to j$, and $k \to l$, are tied if there exists a permutation of the outputs, π, such that $b_{ijy} = b_{kl\pi(y)}$, for all y. Tying induces an equivalence relation on states and on transitions in an obvious way.

The probability of a particular output sequence, $y_1, \ldots, y_T$, is a function of the parameter vector θ. In fact, we can write explicitly

$$P_\theta(y_1^T) \equiv P(Y_1^T = y_1^T) = \sum_{i_1} \cdots \sum_{i_{T+1}} c_{i_1} \prod_{t=1}^{T} a_{i_t i_{t+1}} b_{i_t i_{t+1} y_t}. \tag{9}$$

The computation of $P_\theta(y_1^T)$ can be efficiently organized as follows. Let

$$\alpha_i(t) = P(Y_1^t = y_1^t, X_{t+1} = i). \tag{10}$$

Then $\alpha_i(0) = c_i$ and, for $t > 0$, α_i obeys the recursion

ICASSP 86, TOKYO CH2243-4/86/0000-0049 $1.00 © 1986 IEEE

$$\alpha_i(t) = \sum_j \alpha_j(t-1) a_{ji} b_{jiy_t}. \tag{11}$$

Clearly, $P_\theta(y_1^T) = \sum_j \alpha_j(T)$.

Let $\boldsymbol{M} = (m_1, m_2, \ldots, m_r)$ be a family of hidden Markov models. The parameter vector of $\boldsymbol{M}$ is defined to be $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)$ where θ_i is the parameter vector of m_i. The concept of tying is extended to include states from different members of the family. Thus, we say the state i in m is tied to the state j in m' if there exists a mapping π from the states of m into the states of m' such that $a_{ik} = a'_{j\pi(k)}$ for all k. We extend the concept of tied transitions in the same manner. In addition, we say that the initial state probability of state i in model m is tied to the initial probability of state j in model m' if $c_i = c_j'$. Because of tying or other constraints on Θ, it may be that a change to the parameter vector of one of the members of a family, say m, will necessitate a change to the parameter vector of some other member of the family, say m'. When this is the case, we say that m entails m'. A member, m, is *representative* of a family if it entails each member of the family. A family which has a representative member is said to be *close-knit*.

III. Maximum Likelihood Estimation

Given an output sequence y_1^T from a hidden Markov model m, we wish to estimate the parameter vector Θ of the close-knit family $\boldsymbol{M}$ of which m is a representative member. In maximum likelihood estimation, we attempt to choose Θ so as to make $P_\Theta(y_1^T)$ as large as possible while satisfying any constraints on Θ. If we ignore constraints other than those of equation (8), then we must find a maximum of the associated function

$$F(\Theta, \lambda, \mu, \nu)$$

$$= P_\Theta(y_1^T) + \sum_i \lambda_i \Big(1 - \sum_j a_{ij}\Big) + \sum_{ij} \mu_{ij}\Big(1 - \sum_y b_{ijy}\Big) + \nu\Big(1 - \sum_i c_i\Big) \tag{12}$$

as a function of Θ and the Lagrange multipliers λ_i, μ_{ij}, ν. Thus, we must find a_{ij}, b_{ijy}, c_i such that

$$\frac{\partial F}{\partial a_{ij}} = 0, \quad \frac{\partial F}{\partial b_{ijy}} = 0, \quad \frac{\partial F}{\partial c_i} = 0, \tag{13}$$

or, from equation (12),

$$\frac{\partial P_\Theta}{\partial a_{ij}} - \lambda_i = 0, \quad \frac{\partial P_\Theta}{\partial b_{ijy}} - \mu_{ij} = 0, \quad \frac{\partial P_\Theta}{\partial c_i} - \nu = 0. \tag{14}$$

If we carry out the indicated differentiations, we find, after some manipulation,

$$a_{ij} = \lambda_i^{-1} \sum_{\tau=1}^{T} P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j), \tag{15a}$$

$$b_{ijy} = \mu_{ij}^{-1} \sum_{\tau=1}^{T} P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j, Y_\tau = y), \tag{15b}$$

$$c_i = \nu^{-1} P(Y_1^T = y_1^T, X_1 = i). \tag{15c}$$

It is instructive to treat the derivation of equation (15a) in some detail. We have

$$\frac{\partial a_{kl}}{\partial a_{ij}} = \delta_{ik}\delta_{jl} = \frac{a_{kl}}{a_{ij}}\delta_{ik}\delta_{jl}, \tag{16}$$

and so, from equation (9),

$$\frac{\partial P_\Theta(y_1^T)}{\partial a_{ij}} = \frac{\displaystyle\sum_{\tau=1}^{T} \sum_{i_1} \cdots \sum_{i_{T+1}} c_{i_1}\delta_{ii_\tau}\delta_{ji_{\tau+1}} \prod_{t=1}^{T} a_{i_t i_{t+1}} b_{i_t i_{t+1} y_t}}{a_{ij}} \tag{17}$$

If we substitute this in the first of equations (14) and multiply through by $a_{ij}\lambda_i^{-1}$, and use the fact that

$$P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j) = \sum_{i_1} \cdots \sum_{i_{T+1}} c_{i_1}\delta_{ii_\tau}\delta_{ji_{\tau+1}} \prod_{t=1}^{T} a_{i_t i_{t+1}} b_{i_t i_{t+1} y_t}, \tag{18}$$

we obtain equation (15a).

While equations (15) do not constitute a solution to the maximization problem because the unknowns appear on both sides of the equal signs, they do suggest a recursive method for obtaining a solution. Given values for a_{ij}, b_{ijy}, and c_i, we can use them to evaluate the right-hand sides of equations (15) and then use the left-hand sides as new values for a_{ij}, b_{ijy}, and c_i. These recursive formulas were derived and shown to converge to a solution of equations (13) by Baum [3]. It is clear from the form of equations (15) that none of the parameters is ever negative.

The following equations together with equations (10) and (11) allow rapid computation of the probabilities appearing in equations (15). Let

$$\beta_i(t) = P(Y_{t+1}^T = y_{t+1}^T \mid X_{t+1} = i). \tag{19}$$

Then $\beta_i(T) = 1$ and, for $0 \le t < T$, β_i obeys the recursion

$$\beta_i(t) = \sum_j a_{ij} b_{ijy_{t+1}} \beta_j(t+1). \tag{20}$$

In terms of α_i and β_i, we can write

$$P(Y_1^T = y_1^T, X_t = i, X_{t+1} = j) = \alpha_i(t-1) a_{ij} b_{ijy_t} \beta_j(t), \tag{21}$$

$$P(Y_1^T = y_1^T, X_t = i, X_{t+1} = j, Y_t = y) = \alpha_i(t-1) a_{ij} b_{ijy_t} \beta_j(t) \delta_{yy_t}, \tag{22}$$

$$P(Y_1^T = y_1^T, X_t = i) = \beta_i(0) c_i. \tag{23}$$

IV. Maximum Mutual Information Estimation

Let M be a random variable ranging over the members of $\boldsymbol{M}$, and let $P(m) = P(M = m)$. Consider the joint distribution $P(M = m, Y_1^T = y_1^T)$. We have,

$$\begin{aligned} P(M = m, Y_1^T = y_1^T) &= P(Y_1^T = y_1^T \mid M = m) P(m) \\ &= P_\Theta(y_1^T \mid m) P(m). \end{aligned} \tag{24}$$

The mutual information between the event $M = m$ and the event $Y_1^T = y_1^T$ is a function of Θ given by

$$\begin{aligned} I_\Theta(m, y_1^T) &= \log \frac{P(Y_1^T = y_1^T, M = m)}{P(Y_1^T = y_1^T) P(M = m)} \\ &= \log P_\Theta(y_1^T \mid m) - \log \sum_{m'} P_\Theta(y_1^T \mid m') P(m'). \end{aligned} \tag{25}$$

The idea of maximum mutual information estimation is to choose Θ so as to make $I_\Theta(m, y_1^T)$ as large as possible.

We can proceed as we did in the last section. In addition to the constraints of equation (8), we allow tyings from one model to another. Because m is representative, each of the parameters of any model m' can be identified through tying with one of the parameters of m. Rather than introduce Lagrange multipliers to handle these tying constraints, we will simply rewrite $P_\Theta(y_1^T \mid m')$ in terms of parameters of m directly. Corresponding to equation (12), we have

$$F(\Theta, \lambda, \mu, \nu)$$

$$= I_\Theta(m, y_1^T) + \sum_i \lambda_i \Big(1 - \sum_j a_{ij}\Big) + \sum_{ij} \mu_{ij}\Big(1 - \sum_y b_{ijy}\Big) + \nu\Big(1 - \sum_i c_i\Big). \tag{26}$$

Equation (13) is unchanged, and in place of equation (14), we have,

$$\frac{\partial I_\Theta}{\partial a_{ij}} - \lambda_i = 0, \quad \frac{\partial I_\Theta}{\partial b_{ijy}} - \mu_{ij} = 0, \quad \frac{\partial I_\Theta}{\partial c_i} - \nu = 0. \tag{27}$$

If we carry out the indicated differentiations, we find, after some manipulation,

$$\frac{\sum_{\tau=1}^{T} P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j \mid M = m)}{a_{ij} P_\Theta(y_1^T \mid m)} - \frac{\sum_{m'} \sum_{\tau=1}^{T} P(Y_1^T = y_1^T, A_{ij}(m', \tau) \mid M = m') P(m')}{a_{ij} \sum_{m'} P_\Theta(y_1^T \mid m') P(m')} = \lambda_i, \tag{28a}$$

$$\frac{\sum_{\tau=1}^{T} P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j, Y_\tau = y \mid M = m)}{b_{ijy} P_\Theta(y_1^T \mid m)} - \frac{\sum_{m'} \sum_{\tau=1}^{T} P(Y_1^T = y_1^T, B_{ijy}(m', \tau) \mid M = m') P(m')}{b_{ijy} \sum_{m'} P_\Theta(y_1^T \mid m') P(m')} = \mu_{ij}, \tag{28b}$$

$$\frac{P(Y_1^T = y_1^T, X_1 = i \mid M = m)}{c_i P_\Theta(y_1^T \mid m)} - \frac{\sum_{m'} P(Y_1^T = y_1^T, C_i(m') \mid M = m') P(m')}{c_i \sum_{m'} P_\Theta(y_1^T \mid m') P(m')} = \nu, \tag{28c}$$

Here, $A_{ij}(m', \tau)$ is the event that, for some k and l, $X_\tau = k$ and $X_{\tau+1} = l$ and $a_{kl}^{m'}$ is identified with a_{ij} through tying; $B_{ijy}(m', \tau)$ is the event that, for some k, l, and z, $X_\tau = k$, $X_{\tau+1} = l$, and $Y_\tau = z$ and $b_{klz}^{m'}$ is identified with b_{ijy} through tying; and $C_i(m')$ is the event that for some j, $X_1 = j$ and $c_j^{m'}$ is identified with c_i through tying. These equations lead directly to equations similar to equations (15). For example, if we multiply equation (28a) by $a_{ij}\lambda_i^{-1}$ and rearrange terms, we find

$$a_{ij} = \frac{\sum_{\tau=1}^{T} P(Y_1^T = y_1^T, X_\tau = i, X_{\tau+1} = j \mid M = m)}{\lambda_i P_\Theta(y_1^T \mid m)} - \frac{\sum_{m'} \sum_{\tau=1}^{T} P(Y_1^T = y_1^T, A_{ij}(m', \tau) \mid M = m') p(m')}{\lambda_i \sum_{m'} P_\Theta(y_1^T \mid m') p(m')} \tag{29}$$

Although equation (29) suggests a recursion for finding the maximum of I_Θ, it differs in two important respects from the recursion suggested by equations (15). First, we have no proof that the recursion converges. Second, there is no guarantee that the right-hand side of equation (29) is positive.

The recursion suggested by equation (29) warrants further investigation but for the current paper, we have simply used gradient descent to maximize $F(\Theta, \lambda, \mu, \nu)$. The left-hand sides of equations (27) are the components of the gradient of F with respect to Θ. The components of the gradient with respect to λ_i, μ_{ij}, and ν are, of course, $1 - \sum_j a_{ij}$, $1 - \sum_y b_{ijy}$, and $1 - \sum_i c_i$.

V. Extensions

In order to streamline the discussion in sections III and IV, we have described the algorithms in a simple setting. In practice, it is useful to extend these algorithms in a number of ways. Because the modifications to equations (15) and (28) are straightforward, we simply list the extensions briefly.

We may require that certain of the parameters have specified values. In particular, we may require that certain of the transition probabilities be zero, thereby prohibiting the associated transition. Similarly, by fixing certain initial state probabilities at zero, we can ensure that all state sequences with non-zero probability begin in some set of *initial states.* We can likewise ensure that state sequences with non-zero probability terminate in some set of *final states.*

The number of transitions required to adequately model a process can often be reduced by using *null transitions,* which allow the model to change state without producing any output.

Finally, we have discussed models in which the outputs are chosen from a finite alphabet. It is often desirable to consider outputs which are continuous values or vectors of continuous values. In such cases, in place of the output probability, b_{ijy} we would have an output probability density characterized by a vector of parameters, ϕ_{ij}.

VI. Speech Recognition

Automatic speech recognition can be performed by using a hidden Markov model for each word in the recognizer's vocabulary. For simplicity, we shall assume that each such model has one initial state and one final state. A sequence of N models may be concatenated into a combined model by making the initial state of the first model the initial state of the combined model, by making the final state of the Nth model the final state of the combined model, and by creating null transitions from the final state of the ith model to the initial state of the $(i + 1)$st model for $1 \leq i < N$. In this way, we can create a hidden Markov model m from any word sequence w. Note that there will be tied parameters in m if any word appears more than once in w.

The parameter values for the individual word models are estimated from a sequence of words $\tilde{w}$, the training script, and a corresponding sequence of acoustic observations y, the training data. We can use either maximum-likelihood estimation or MMI estimation to estimate the parameter vector θ for the model created for $\tilde{w}$.

When using MMI estimation, each possible word sequence w gives rise to a model m, and the training script $\tilde{w}$ gives rise to $\tilde{m}$. The objective function in MMI estimation becomes $I_\Theta(\tilde{w}, y)$. By maximizing $I_\Theta(\tilde{w}, y)$, we are maximizing the probability of the correct word sequence given the training data, or equivalently, maximizing the information provided by y about $\tilde{w}$. Maximum likelihood estimation, on the other hand, is a method of choosing parameter values to maximize the probability of the acoustic observation sequence y given the model $\tilde{m}$ for the word sequence $\tilde{w}$. Unlike MMI estimation, it is not explicitly designed to maximize the ability of the resultant statistical models to discriminate between the correct word sequence and any other word sequence.

The computation in MMI estimation can be reduced by making a few simplifying assumptions.

The first of these assumptions is that

$$P(W = w_1^N) = \prod_{i=1}^{N} P(W_i = w_i). \tag{30}$$

Each $P(W_i = w_i)$ can easily be estimated from word frequency counts.

It is also convenient to assume that the training data y can be segmented into N subsequences $y_1, \ldots, y_N$, one y_i for each word w_i in w_1^N, and to assume that

$$P(Y = y \mid W = w) = \prod_{i=1}^{N} P(Y_i = y_i \mid W_i = w_i), \tag{31}$$

if the length of w is N, and that $P(Y_i = y_i \mid W_i = w_i) = 0$, otherwise.

In practice, for each acoustic segment y_i there is a set Λ_i of only a few words which contribute significantly to $\sum_w P(W = w)\ P(Y = y_i \mid W = w)$, and it reduces computation to assume that

$$\sum_w P(W = w)P(Y = y_i \mid W = w)$$

$$= \sum_{w \in \Lambda_i} P(W = w)P(Y = y_i \mid W = w). \tag{32}$$

Each set Λ_i can be precomputed from a list of words confusable with w_i.

VII. Results

We tested MMI estimation in a speaker-dependent isolated word recognition task which is described in more detail in [1]. The vocabulary size was 2000 words. The training data consisted of 99 natural sentences containing a total of 1453 words. The test data consisted of 100 natural sentences containing a total of 1297 words. The test data was recognized using the one-gram language model of equation (30).

First, we estimated the hidden Markov model parameters with maximum likelihood estimation. The log probability of the correct script given the acoustic observations in the training data, computed with the resultant parameters was -190.11. There were 78 recognition errors made on the test data with these parameters.

We then recomputed the output probabilities using the method of maximizing the mutual information between the acoustic observations in the training data and the training script. The Λ_i's appearing in (32) were determined by including in each Λ_i all words w such that $P(Y_i = y_i \mid W = w) \geq 10^{-3} P(Y_i = y_i \mid W = \tilde{w}_i)$. With these parameter estimates, the log probability of the correct script computed on the training data, was -1.16, and there were 64 recognition errors made on the test data.

The above figures show that in this example estimating parameters by maximizing mutual information resulted in the training script having a probability 10^{189} times greater than when parameters were estimated by maximum likelihood estimation. More importantly, training by maximizing mutual information resulted in 18 percent fewer recognition errors.

VIII. Acknowledgements

We would like to thank the other members of the IBM Continuous Speech Recognition Group for their useful comments and suggestions.

IX. References

1. L. R. Bahl, S. K. Das, P. V. de Souza, F. Jelinek, S. Katz, R. L. Mercer, M. A. Picheny, "Some experiments with large-vocabulary isolated-word sentence recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* IEEE Catalog No. 84CH1945-5, Vol. 2, pp. 26.5.1 - 26.5.2, San Diego (1984).
2. L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-5, no. 2, March 1983, pp. 179-190.
3. L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," in *Inequalities,* vol. 3, 1972, pp. 1-8.