A Modified Baum–Welch Algorithm for Hidden Markov Models with Multiple Observation Spaces

Paul M. Baggenstoss, *Member, IEEE*

***Abstract*—In this paper, we derive an algorithm similar to the well-known Baum–Welch algorithm for estimating the parameters of a hidden Markov model (HMM). The new algorithm allows the observation PDF of each state to be defined and estimated using a different feature set. We show that estimating parameters in this manner is equivalent to maximizing the likelihood function for the standard parameterization of the HMM defined on the input data space. The processor becomes optimal if the state-dependent feature sets are sufficient statistics to distinguish each state individually from a common state.**

***Index Terms*—Baum–Welch algorithm, class-specific, EM algorithm, expectation-maximization, Gaussian mixtures, hidden Markov model (HMM), parameter estimation, sufficient statistics.**

I. Introduction

THE class-specific method was recently developed as a method of dimensionality reduction in classification [1], [2]. Unlike other methods of dimension reduction, it is based on sufficient statistics and results in no *theoretical* loss of performance. Performance is always lost going from theory to practice due to (1) loss of information when reducing data to features, and (2) approximation of the theoretical feature PDFs. There is always a tradeoff between the desire to retain as much information as possible (by increasing the feature dimension) and the desire to obtain better PDF estimates (by decreasing the dimension). The class-specific method obtains a better compromise by allowing more information to be kept for a given maximum feature dimension. It does this by assigning a separate feature set to each class. Now we extend the idea further to the problem of HMM modeling when each state of the HMM may have its own approximate sufficient statistic.

II. Mathematical Results

We show in this section that the class-specific HMM is merely a different way to parameterize the likelihood function of the conventional HMM. Let $L(\mathbf{X};\lambda)$ be the likelihood function defined for the input data $\mathbf{X}$. A special class-specific likelihood function, $L^z(\mathbf{Z};\lambda^z)$ is defined using the class-specific (state-specific) statistics $\mathbf{Z}$. It is shown below that maximizing $L^z(\mathbf{Z};\lambda^z)$ over λ^z is equivalent to maximizing $L(\mathbf{X};\lambda)$ over λ with special constraints. While it is not necessary for $\mathbf{Z}$ to be sufficient for this to be true, the processor constructed from class-specific sufficient statistics will be optimal, provided there is no PDF estimation error.

A. Standard Parameterization and Notation

We consider a set of state occurrences $\boldsymbol{\theta} \triangleq \{q_1 \ldots q_T\}$ where $1 \le q_t \le N$. The sequence $\boldsymbol{\theta}$ is a realization of the Markov chain with state priors $\{\pi_j, j = 1,2 \ldots N\}$ and $N \times N$ state transition matrix $A = \{a_{ij}\}$. Rather than observing the states $\boldsymbol{\theta}$ directly, we observe the stochastic outputs $\mathbf{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_T\}$ which are realizations from a set of state PDFs

$$p_j(\mathbf{x}) \triangleq p(\mathbf{x}|H_j), \quad j = 1,2 \ldots N$$

where H_j is the condition that state j is true. We assume the observations are independent, thus

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p_{q_t}(\mathbf{x}_t).$$

The complete set of parameters defining the HMM are

$$\lambda \triangleq [\{\pi_j\}, \{a_{ij}\}, \{p_j(\,)\}]$$

where $\sum_{j=1}^{N} \pi_j = 1$, $\sum_{j=1}^{N} a_{ij} = 1$. The likelihood function is the joint density of the observation sequence given the model parameters and is written (see [3, Eq. 17])

$$\begin{aligned} L(\mathbf{X};\lambda) \triangleq p(\mathbf{X};\lambda) &= \sum_{\boldsymbol{\theta}} p(\mathbf{x},\boldsymbol{\theta};\lambda) \\ &= \sum_{\boldsymbol{\theta}} \pi_{q_1} p_{q_1}(\mathbf{x}_1;\lambda) \prod_{t=2}^{T} a_{q_{t-1}q_t} p_{q_t}(\mathbf{x}_t;\lambda) \end{aligned} \tag{1}$$

where $\sum_{\boldsymbol{\theta}}$ is a summation over all possible state sequences of length T. The maximum likelihood (ML) estimate of λ is defined as

$$\hat{\lambda} \triangleq \arg\max_{\lambda} L(\mathbf{X};\lambda). \tag{2}$$

We use notation similar to Rabiner [3] with the exception that we represent state PDFs as $p_j(\cdot)$, and observations as $\mathbf{x}_t$. In the paper, functions beginning with the letters "b" and "p," always denote PDFs. The letter "b" is reserved for components of mixture PDFs and "p" is used for all other PDFs. The exception is any function carrying the superscript "$*$" which is *not* a PDF.

Manuscript received March 3, 2000; revised August 29, 2000. This work was supported by the Office of Naval Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Rahim.

The author is with Naval Undersea Warfare Center, Newport, RI 02841 USA (e-mail: p.m.baggenstoss@ieee.org).

Publisher Item Identifier S 1063-6676(01)02991-1.

B. Class-Specific Parameterization

Define

$$\mathbf{Z} \triangleq \{[\mathbf{z}_{1,0} \dots \mathbf{z}_{N,0}], [\mathbf{z}_{1,1} \dots \mathbf{z}_{N,1}], \cdots [\mathbf{z}_{1,T} \dots \mathbf{z}_{N,T}]\}$$

where

$$\mathbf{z}_{j,t} \triangleq T_j(\mathbf{x}_t), \quad j = 1 \dots N, \ t = 1 \dots T.$$

The complete class-specific parameterization is written

$$\lambda^z \triangleq \left[\{\pi_j\}, \{a_{ij}\}, \{\boldsymbol{\mu}_{jk}^z\}, \{\mathbf{U}_{jk}^z\}, \{c_{jk}^z\}\right]$$

where $\{\pi_j\}, \{a_{ij}\}$ are identical to the corresponding components of λ, and have the same constraints. The state-dependent PDFs are modeled as Gaussian mixture densities

$$p_j^z(\mathbf{z}_j; \lambda^z) \triangleq \sum_{k=1}^{M} c_{jk}^z \mathcal{N}(\mathbf{z}_j, \boldsymbol{\mu}_{jk}^z, \mathbf{U}_{jk}^z) \tag{3}$$

where $\sum_k c_{jk}^z = 1$ and $\mathcal{N}(\mathbf{z}_j, \boldsymbol{\mu}^z, \mathbf{U}^z)$ are the joint Gaussian densities

$$\mathcal{N}(\mathbf{z}_j, \boldsymbol{\mu}^z, \mathbf{U}^z) \triangleq (2\pi)^{-P_j/2} |\mathbf{U}^z|^{-1/2} \times \exp\left\{-\frac{1}{2}(\mathbf{z}_j - \boldsymbol{\mu}^z)'(\mathbf{U}^z)^{-1}(\mathbf{z}_j - \boldsymbol{\mu}^z)\right\}$$

and P_j is the dimension of $\mathbf{z}_j$. The relationship between λ^z and λ will be established shortly.

1) Noise-Only Condition H_0: To apply the class-specific method in its simplest form [see note following (11)], we need to define a condition H_0 that is *common* to all states. One way to do this is to let H_0 represent the "noise-only" condition. For example, assume the PDF of $\mathbf{x}$ in each state is dependent on a "signal amplitude" parameter ρ_j. Under H_j, the PDF of $\mathbf{x}$ is marginalized over the distribution of ρ_j, thus

$$p(\mathbf{x}|H_j) = \int_{\rho_j} p(\mathbf{x}|\rho_j, H_j) p(\rho_j|H_j), \quad 1 \le j \le N.$$

Let there exist a common noise-only condition H_0 defined by

$$\begin{aligned} p(\mathbf{x}|H_0) &= p(\mathbf{x}|\rho_1 = 0, H_1) = p(\mathbf{x}|\rho_2 = 0, H_2) \\ &= \cdots = p(\mathbf{x}|\rho_N = 0, H_N). \end{aligned}$$

We assume $\mathbf{x}_t$ are independent under H_0. Thus,

$$p(\mathbf{X}|H_0) = \prod_{t=1}^{T} p(\mathbf{x}_t|H_0). \tag{4}$$

One further requirement is that

$$p(\mathbf{x}|H_0) > 0, \quad \text{for all} \quad \mathbf{x} \in \mathcal{X} \tag{5}$$

where $\mathcal{X}$ is the allowable range of $\mathbf{x}_t$. Note that this requirement is met if $p(\mathbf{x}|H_0)$ is Gaussian.

While this structure does not seem to fit many problems of interest, any problem can be modified to include an amplitude parameter even if the amplitude parameter is never zero in practice. Furthermore, the noise-only condition H_0 can have an arbitrarily small assumed variance because H_0 is only a theoretical tool and does not need to approximate any realistic situation. We will explain how the choice of H_0, affects the choice of the state-dependent statistics.

2) Sufficiency of $\mathbf{Z}$ and Relationship to H_0: We will show that if $\mathbf{Z}$ meets a special sufficiency requirement, the class-specific method becomes optimum. To understand the implications of the sufficiency of $\mathbf{Z}$, we must consider a conventional feature-based approach in which a common feature set replaces the raw data. Let $\{\mathbf{z}_t = T(\mathbf{x}_t), \ 1 \le t \le T\}$ and define the HMM based on the state-dependent distributions $\{p_j(\mathbf{z}), \ 1 \le j \le N\}$. This is the conventional HMM approach which has been very successful [3]. An example of $\mathbf{z}$ is a set of cepstrum-derived features. For optimality of the resulting processor, $\mathbf{z}$ must be a sufficient statistic for the classification of the N states. One way to express the sufficiency requirement is through the likelihood ratios, which are invariant when written as a function of a sufficient statistic [4], [5]. More precicely

$$\frac{p(\mathbf{x}|H_j)}{p(\mathbf{x}|H_k)} = \frac{p(\mathbf{z}|H_j)}{p(\mathbf{z}|H_k)}, \quad 1 \le j, \ k \le N, \quad j \ne k. \tag{6}$$

Clearly, $\mathbf{z}$ must contain all information necessary to distinguish any two states. This can be a very difficult requirement to meet in practice because a significant amount of information can be lost when reducing the data to features. In practice, the tradeoff consists of the contradictory goals of making $\mathbf{z}$ as close to *sufficient* as possible (by making $\mathbf{z}$ larger), while at the same time making the PDF estimation problem as tractable as possible (by making $\mathbf{z}$ smaller).

For optimality of the class-specific method, however, we require that $\mathbf{z}_j = T_j(\mathbf{x})$ be a sufficient statistic for the binary hypothesis test between H_j and H_0. Specifically, if $\mathbf{z}_j$ is sufficient, we have

$$\frac{p(\mathbf{x}|H_j)}{p(\mathbf{x}|H_0)} = \frac{p(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_0)}, \quad 1 \le j \le N. \tag{7}$$

Clearly, $\mathbf{z}_j$ must contain *all* information which helps distinguish H_j from H_0. In contrast to the conventional method which places all information in $\mathbf{z}$, the class-specific method distributes the information among the class-specific statistics. For a fixed feature set dimension, more information is allowed.

Clearly, the selection of H_0 affects the selection of the feature transformations $T_j(\ )$. For optimality to hold, no information which can help distinguish H_j from H_0 can be discarded. An example is background noise. If H_j contains background noise inconsistent with H_0, then z_j should contain information about the background noise. To reduce the complexity of the feature sets, it may be necessary to whiten and normalize the data in such a way that the background noise resembles H_0. On the other hand, it may be acceptable to discard the information and suffer a slight performance loss. This is especially true a high signal-to-noise ratio (SNR).

3) Class-Specific Likelihood Function: Define the class-specific likelihood function as

$$L^z(\mathbf{Z};\lambda^z) \triangleq \sum_{\boldsymbol{\theta}} \pi_{q_1}\left[\frac{p^z_{q_1}(\mathbf{z}_{q_1,1};\lambda^z)}{p(\mathbf{z}_{q_1,1}|H_0)}\right] \times \prod_{t=2}^{T}\left[a_{q_{t-1}q_t}\frac{p^z_{q_t}(\mathbf{z}_{q_t,t};\lambda^z)}{p(\mathbf{z}_{q_t,t}|H_0)}\right]. \quad (8)$$

The maximum likelihood (ML) estimate of λ^z is defined as

$$\hat{\lambda^z} \triangleq \arg\max_{\lambda^z} L^z(\mathbf{Z};\lambda^z). \quad (9)$$

The objective is to derive an algorithm to solve (2) by solving (9).

4) Relationship to the Standard Parameterization and Optimality: It is not clear yet that (8) is related to (1), however in fact we can solve (2) by solving (9). To demonstrate this, we need to convert any class-specific parameter set λ^z into a valid conventional parameter set λ. This requires that the PDF parameters $[\{\boldsymbol{\mu}^z_{jk}\},\{\mathbf{U}^z_{jk}\},\{c^z_{jk}\}]$ can be converted into PDFs defined on $\mathcal{X}$. For this, we need Theorems 1 and 2.

To introduce the theorems, we define a feature set $\mathbf{z}=T(\mathbf{x})$. Because $T(\mathbf{x})$ is many-to-one, there is no way to reconstruct the PDF of $\mathbf{x}$ unambiguously given the PDF of $\mathbf{z}$. However, Theorems 1 and 2 show that given an arbitrary PDF $f_z(\mathbf{z})$, and arbitrary feature transformation $T(\mathbf{x})$, it is possible to construct a PDF $f_x(\mathbf{x})$ such that when $\mathbf{x}$ is drawn from $f_x(\mathbf{x})$, the distribution of $\mathbf{z}$ will be $f_z(\mathbf{z})$. We will also mention other desirable properties that can be attributed to $f_x(\mathbf{x})$.

Theorem 1: Let the PDF $p_x(\mathbf{x}|H_0)$ be defined on $\mathcal{X}$ and let $p_x(\mathbf{x}|H_0)>0$ for all $x\in\mathcal{X}$. Let the r.v. $\mathbf{z}$ be related to $\mathbf{x}$ by the many-to-one feature transformation $\mathbf{z}=T(\mathbf{x})$ where $T(\mathbf{x})$ is any measurable function of $\mathbf{x}$. Let $\mathcal{Z}$ be the image of $\mathcal{X}$ under transformation $T(\mathbf{x})$. Let $p_z(\mathbf{z}|H_0)$ be the PDF of $\mathbf{z}$ when $\mathbf{x}$ is drawn from the PDF $p_x(\mathbf{x}|H_0)$. Thus, $p_z(\mathbf{z}|H_0)>0$ for all $z\in\mathcal{Z}$. Let $f_z(\mathbf{z})$ be any PDF defined on $\mathcal{Z}$. Then the function defined by

$$f_x(\mathbf{x}) = \frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)} f_z(T(\mathbf{x})) \quad (10)$$

is a PDF defined on $\mathcal{X}$.

Proof:

$$\begin{aligned}\int_{\mathbf{x}\in\mathcal{X}} \frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)} f_z(T(\mathbf{x})) &= \mathrm{E}_{x|H_0}\left\{\frac{f_z(T(\mathbf{x}))}{p_z(T(\mathbf{x})|H_0)}\right\}\\ &= \mathrm{E}_{z|H_0}\left\{\frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)}\right\}\\ &= \int_{\mathbf{z}\in\mathcal{Z}} \frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)} p_z(\mathbf{z}|H_0)d\mathbf{z}\\ &= \int_{\mathbf{z}\in\mathcal{Z}} f_z(\mathbf{z})d\mathbf{z}\\ &= 1.\end{aligned}$$

The equivalence of the expected values in lines one and two is an application of the change of variables theorem [6]. For example, let $h(\mathbf{z})$ be any function defined on $\mathcal{Z}$. If $\mathbf{z}=T(\mathbf{x})$, then $\mathrm{E}_z\{h(\mathbf{z})\}=\mathrm{E}_x\{h(T(\mathbf{x}))\}$. This can be seen when the expected values are written as the limiting form of the sample mean of a size-K sample set as $K\to\infty$, i.e., Theorem 2.

Theorem 2: Let $\mathbf{x}$ be drawn from the distribution $f_x(\mathbf{x})$ as defined in (10). Then if $\mathbf{z}=T(\mathbf{x})$, the PDF of $\mathbf{z}$ is $f_z(\mathbf{z})$.

Proof: Let $M_z(\mathbf{y})$ be the joint moment generating function (MGF) of $\mathbf{z}$. By definition,

$$\begin{aligned}M_z(\mathbf{y}) &= \mathrm{E}_z\{e^{\mathbf{y}'\mathbf{z}}\}\\ &= \mathrm{E}_x\{e^{\mathbf{y}'T(\mathbf{x})}\}\\ &= \int_{\mathbf{x}\in\mathcal{X}} e^{\mathbf{y}'T(\mathbf{x})}\frac{p_x(\mathbf{x}|H_0)}{p_z(T(\mathbf{x})|H_0)} f_z(T(\mathbf{x}))d\mathbf{x}\\ &= \mathrm{E}_{x|H_0}\left\{e^{\mathbf{y}'T(\mathbf{x})}\frac{f_z(T(\mathbf{x}))}{p_z(T(\mathbf{x})|H_0)}\right\}\\ &= \mathrm{E}_{z|H_0}\left\{e^{\mathbf{y}'\mathbf{z}}\frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)}\right\}\\ &= \int_{\mathbf{z}\in\mathcal{Z}} e^{\mathbf{y}'\mathbf{z}}\frac{f_z(\mathbf{z})}{p_z(\mathbf{z}|H_0)} p_z(\mathbf{z}|H_0)d\mathbf{z}\\ &= \int_{\mathbf{z}\in\mathcal{Z}} e^{\mathbf{y}'\mathbf{z}} f_z(\mathbf{z})d\mathbf{z}\end{aligned}$$

from which we may conclude that the PDF of $\mathbf{z}$ is $f_z(\mathbf{z})$.

The PDF $f_x(\mathbf{x})$ has the following properties.

1) Let H_1 be some arbitrary hypothesis with PDF defined on $\mathcal{X}$. Then, when $T(\mathbf{x})$ is a sufficient statistic for the binary test of H_1 versus H_0, then as $f_z(\mathbf{z})\to p(\mathbf{z}|H_1)$, we have $f_x(\mathbf{x})\to p(\mathbf{x}|H_1)$.
2) Let $\mathbf{z}^*$ be a point in $\mathcal{Z}$. Then
$$\frac{f_x(\mathbf{x})}{p_x(\mathbf{x}|H_0)} = \frac{f_z(\mathbf{z}^*)}{p_z(\mathbf{z}^*|H_0)} \quad \text{for all } \mathbf{x} \text{ such that } T(\mathbf{x})=\mathbf{z}^*.$$
Thus, $f_x(\mathbf{x})$ has the property that all points $\mathbf{x}$ such that $T(\mathbf{x})=\mathbf{z}^*$ are *equally distinguishable* from H_0.
3) Although Theorems 1 and 2 do not impose any sufficiency requirements on $\mathbf{z}$, it results that $\mathbf{z}$ are sufficient statistics for the constructed PDF. More precicely, $\mathbf{z}$ is an exact sufficient statistic for the binary hypothesis test of $f_x(\mathbf{x})$ versus $p_x(\mathbf{x}H_0)$.

We now show that we can solve (2) by solving (9). Suppose that given λ^z, one constructed a standard parameterization $\lambda^z\longrightarrow\lambda$, written $\lambda=G(\lambda^z)$, by constructing the PDFs

$$p_j(\mathbf{x};G(\lambda^z)) \triangleq \left[\frac{p(\mathbf{x}|H_0)}{p(T_j(\mathbf{x})|H_0)}\right] p^z_j(T_j(\mathbf{x});\lambda^z), \quad \text{for } 1\le j\le N. \quad (11)$$

Note that, in general, the reference hypothesis can be a function of j, written $H_{0,j}$. For simplicity, we have chosen to use a common reference $H_{0,j}=H_0$. That $p_j(\mathbf{x};G(\lambda^z))$ are indeed PDFs can be seen from Theorem 1. Furthermore, from Theorem 2, it may be seen that these densities are such that they induce the densities $p^z_j(\mathbf{z}_j;\lambda^z)$ on $\mathbf{z}_j$. Next, from (1), (4), (8), and (11), we see that

$$L^z(\mathbf{Z};\lambda^z) = \frac{L(\mathbf{X};G(\lambda^z))}{p(\mathbf{X}|H_0)}. \quad (12)$$

Therefore, if we define $\hat{\lambda}^g \triangleq G(\hat{\lambda}^z)$, we have

$$\hat{\lambda}^g = \arg\max_{\lambda^z} L(\mathbf{X};G(\lambda^z)).$$

Thus, we can claim that $\hat{\lambda}^g$ maximizes $L(\mathbf{X};\lambda)$ over all λ which satisfy $\lambda=G(\lambda^z)$ for some λ^z. Furthermore, when the class-

specific statistics are sufficient, (7) holds and it follows from (7), (11) that if $p_j^z(\mathbf{z}_j;\hat{\lambda}^z) \to p(\mathbf{z}_j|H_j)$, then $p_j(\mathbf{x};G(\hat{\lambda}^z)) \to p(\mathbf{x}|H_j)$. Thus, one is able to construct the true HMM parameters from the class-specifc parameter estimates. Furthermore,

$$L^z(\mathbf{Z};\hat{\lambda}^z) \to \frac{L(\mathbf{X};\lambda)}{p(\mathbf{X}|H_0)}$$

and the class-specific classifier becomes the optimal Neyman–Pearson classifier for comparing competing HMM hypotheses.

C. Class-Specific Baum–Welch Algorithm

An iterative algorithm for solving (2) based on the EM method, and due to Baum [7] is available. Formulas for updating the parameters λ at each iteration are called the reestimation formulas [3]. The derivation by Juang [8] is well known for the case when $p_j(\mathbf{x})$ are Gaussian mixtures. We need to modify the derivation of Juang to solve (9). We may write

$$L^z(\mathbf{Z};\lambda^z) = \sum_{\boldsymbol{\theta}} \pi_{q_1} \left[\sum_{k=1}^{M} c_{q_1 k}^z b_{q_1,k}^*(\mathbf{z}_{q_1,1};\lambda^z)\right] \times \prod_{t=2}^{T}\left[a_{q_{t-1}q_t}\sum_{k=1}^{M} c_{q_t k}^z b_{q_t,k}^*(\mathbf{z}_{q_t,t};\lambda^z)\right] \quad (13)$$

where

$$b_{jk}^*(\mathbf{z}_j;\lambda^z) \triangleq \frac{\mathcal{N}(\mathbf{z}_j,\boldsymbol{\mu}_{jk}^z,\mathbf{U}_{jk}^z)}{p(\mathbf{z}_j|H_0)}. \quad (14)$$

This may then be rewritten as (see Juang [8, Eqs. 8–11])

$$L^z(\mathbf{Z};\lambda^z) = \sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}} p^*(\mathbf{Z},\boldsymbol{\theta},K;\lambda^z) \quad (15)$$

where $\sum_{\mathbf{K}} \triangleq \sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\cdots\sum_{k_T=1}^{M}$ and

$$p^*(\mathbf{Z},\boldsymbol{\theta},K;\lambda^z) \triangleq \pi_{q_1} b_{q_1,k_1}^*(\mathbf{z}_{q_1,1};\lambda^z)\, c_{q_1,k_1}^z \times \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t,k_t}^*(\mathbf{z}_{q_t,t};\lambda^z) c_{q_t,k_t}^z. \quad (16)$$

We wish to maximize the function $L^z(\mathbf{Z};\lambda^z)$ over λ^z. To this end, we seek an algorithm that given a parameter value λ^z, we can always find a new $\lambda^{z\prime}$ such that $L^z(\mathbf{Z};\lambda^{z\prime}) \geq L^z(\mathbf{Z};\lambda^z)$.

1) Auxiliary Function:

Theorem 3: Define

$$Q(\lambda^z,\lambda^{z\prime}) \triangleq \sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}} p^*(\mathbf{Z},\boldsymbol{\theta},K;\lambda^z)\log p^*(\mathbf{Z},\boldsymbol{\theta},K;\lambda^{z\prime}). \quad (17)$$

If $Q(\lambda^z,\lambda^{z\prime}) \geq Q(\lambda^z,\lambda^z)$, *then* $L^z(\mathbf{Z};\lambda^{z\prime}) \geq L^z(\mathbf{Z};\lambda^z)$.

Proof: $\log x$ is strictly concave for $x > 0$. Hence, see (18), shown at the bottom of the page.

The inequality is strict unless $p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^{z\prime}) = p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)$. Note that this proof differs in no meaningful way from Baum's [7] or Juang's [8]. One important difference is that $p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)$ is not a PDF. But the proof relies on Jenssen's inequality which is based on expected values using the probability measure $p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)/L^z(\mathbf{Z};\lambda^z)$, which is a discrete PDF due to (15).

2) Reestimation Algorithm: The problem now is to solve for

$$\max_{\lambda^{z\prime}} Q(\lambda^z,\lambda^{z\prime}). \quad (19)$$

We have

$$\begin{aligned} Q(\lambda^z,\lambda^{z\prime}) &= \sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}} p^*(\mathbf{Z},\boldsymbol{\theta},K|\lambda^z)\ \log p^*(\mathbf{Z},\boldsymbol{\theta},K;\lambda^{z\prime}) \\ &= \sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}}\Big[\pi_{q_1} b_{q_1,k_1}^*(\mathbf{z}_{q_1,1}) c_{q_1,k_1}^z \\ &\qquad \times \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t,k_t}^*(\mathbf{z}_{q_t,t}) c_{q_t,k_t}^z\Big] \\ &\quad \cdot \Big\{\log u_{q_1}' + \sum_{t=2}^{T}\log a_{q_{t-1}q_t}' \\ &\qquad + \sum_{t=1}^{T}\log b_{q_t,k_t}^{*\prime}(\mathbf{z}_{q_t,t}) + \sum_{t=1}^{T}\log c_{q_t,k_t}^{z\prime}\Big\}. \end{aligned} \quad (20)$$

We then may follow the proof of Juang, provided the necessary requirements of $b_{jk}^*(\mathbf{x};\lambda^z)$ are met. Notice that $b_{jk}^*(\mathbf{z}_j;\lambda^z)$ depends on λ^z only through the multivariate Gaussian density, the data $\mathbf{z}_j$ may be considered fixed. Thus, $b_{jk}^*(\mathbf{z}_j;\lambda^z)$ meets the necessary log-concavity and elliptical symmetry requirements necessary for the reestimation formulas that follow. We can proceed in the proof of Juang, until it is necessary to differentiate $b_{jk}^*(\mathbf{z}_j;\lambda^z)$ with respect to λ^z. At that point, the additional terms in (14), not dependent on λ^z, do not affect the solution of the

$$\begin{aligned} \log\left(\frac{L^z(\mathbf{Z};\lambda^{z\prime})}{L^z(\mathbf{Z};\lambda^z)}\right) &= \log\left(\sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}}\frac{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^{z\prime})}{L^z(\mathbf{Z};\lambda^z)}\right) \\ &= \log\left(\sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}}\frac{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)}{L^z(\mathbf{Z};\lambda^z)}\frac{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^{z\prime})}{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)}\right) \\ &\geq \sum_{\boldsymbol{\theta}}\sum_{\mathbf{K}}\frac{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)}{L^z(\mathbf{Z};\lambda^z)}\log\left(\frac{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^{z\prime})}{p^*(\mathbf{Z},\boldsymbol{\Theta},\mathbf{K};\lambda^z)}\right) \\ &= [L(\mathbf{Z};\lambda^z)]^{-1}[Q(\lambda^z,\lambda^{z\prime}) - Q(\lambda^z,\lambda^z)] \geq 0 \end{aligned} \quad (18)$$

maximization; we only need to let $\mathbf{z}_{j,t}$ take the place of $\mathbf{x}_t$. The resulting algorithm is provided below.

3) Class-Specific Forward Procedure: The joint probability (1) may be calculated with the *forward procedure* [3], [8] by recursively computing the quantities

$$\alpha_t(i) \triangleq p(\mathbf{x}_1 \ldots \mathbf{x}_t, q_t = i; \lambda).$$

Similarly, the class-specific likelihood function (8) may be calculated recursively by recursively computing the quantities

$$\alpha_t^c(i) \triangleq \frac{p(\mathbf{x}_1 \ldots \mathbf{x}_t, q_t = i; \lambda^z)}{p(\mathbf{x}_1 \ldots \mathbf{x}_t | H_0)}.$$

1) Initialization:

$$\alpha_1^c(i) = \pi_i \frac{p_i^z(\mathbf{z}_{i,1}; \lambda^z)}{p(\mathbf{z}_{i,1}|H_0)}, \quad 1 \le i \le N. \tag{21}$$

2) Induction:

$$\alpha_{t+1}^c(j) = \left[\sum_{i=1}^{N} \alpha_t^c(i) a_{ij}\right] \frac{p_j^z(\mathbf{z}_{j,t+1}; \lambda^z)}{p(\mathbf{z}_{j,t+1}|H_0)}, \quad 1 \le t \le T-1, \quad 1 \le j \le N. \tag{22}$$

3) Termination:

$$L^z(\mathbf{Z}; \lambda^z) = \frac{L(\mathbf{X}; G(\lambda^z))}{p(\mathbf{X}|H_0)} = \sum_{i=1}^{N} \alpha_T^c(i). \tag{23}$$

4) Class-Specific Backward Procedure: The backward parameters $\beta_t^c(i)$ are similarly defined.

1) Initialization:

$$\beta_T^c(i) = 1. \tag{24}$$

2) Induction:

$$\beta_t^c(i) = \sum_{j=1}^{N} a_{ij} \frac{p_j^z(\mathbf{z}_{j,t+1}; \lambda^z)}{p(\mathbf{z}_{j,t+1}|H_0)} \beta_{t+1}^c(j), \quad t = T-1, T-2, \ldots 1, \quad 1 \le i \le N. \tag{25}$$

5) HMM Reestimation Formulas: Define $\gamma_t(j)$ as $p(q_t = j|\mathbf{X})$. We have

$$\gamma_t(j) = \frac{\alpha_t^c(j)\beta_t^c(j)}{\sum_{i=1}^{N} \alpha_t^c(i)\beta_t^c(i)}. \tag{26}$$

Let [see (27), shown at the bottom of the page]. The updated state priors are

$$\hat{u}_i = \gamma_1(i). \tag{28}$$

The updated state transition matrix is

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}. \tag{29}$$

Keep in mind that $\gamma_t(j)$, $\xi_t(i,j)$, $\hat{u}_i$, and $\hat{a}_{ij}$ will be identical to those estimated by the conventional approach if (7) is true or if $\lambda = G(\lambda^z)$.

6) Gaussian Mixture Reestimation Formulas: Let

$$\gamma_t^c(j,m) \triangleq \gamma_t(j) \left[\frac{c_{jm}^z \mathcal{N}(\mathbf{z}_{j,t}, \boldsymbol{\mu}_{jm}^z, \mathbf{U}_{jm}^z)}{p_j^z(\mathbf{z}_{j,t}; \lambda^z)}\right], \tag{30}$$

$$\hat{c}_{jm}^z = \frac{\sum_{t=1}^{T} \gamma_t^c(j,m)}{\sum_{t=1}^{T} \sum_{l=1}^{M} \gamma_t^c(j,l)}, \tag{31}$$

$$\hat{\boldsymbol{\mu}}_{jm}^z = \frac{\sum_{t=1}^{T} \gamma_t^c(j,m) \mathbf{z}_{j,t}}{\sum_{t=1}^{T} \gamma_t^c(j,m)} \tag{32}$$

and

$$\hat{\mathbf{U}}_{jm}^z = \frac{\sum_{t=1}^{T} \gamma_t^c(j,m)(\mathbf{z}_{j,t} - \boldsymbol{\mu}_{jm}^z)(\mathbf{z}_{j,t} - \boldsymbol{\mu}_{jm}^z)'}{\sum_{t=1}^{T} \gamma_t^c(j,m)}. \tag{33}$$

III. Applying the Method

Since truly sufficient statistics can never be found in practice, the practitioner must be satisfied with approximate sufficiency. Partially sufficiency of the features poses no theoretical problems because the class-specific Baum–Welch algorithm maximizes the true likelihood function without requiring sufficiency, albeit subject to the constraint that the state PDFs are of the form (11). As theory guides the practice, the sufficiency of the form (6) which is required for the optimality of the standard approach tells practitioners to look for features which discriminate between the states. In contrast, the sufficiency (7) required for the optimality of the class-specific approach tells practitioners to look for features which discriminate states from H_0 and whose exact PDF can be derived under H_0.

Approximations of $\{p_j^z(\mathbf{z}_j|H_0)\}$ may also be used as long as these approximations are valid in the tails. Tail behavior is important because as samples diverge from H_0, the denominators in (8) approach zero. Approximations with accurate tail behavior are available for a wide range of important feature sets in signal processing including autocorrelation and cepstrum estimates [9].

A. Example

The following conceptual example illustrates the selection of class-specific features. Consider a Markov process in which there are three states characterized by

- H_1: a low-order autoregressive process (such as a whistle) of unknown variance;
- H_2: a pure tone of unknown frequency, amplitude, and phase in additive Gaussian noise of unknown variance;
- H_3: a positive-valued impulse of duration 1 sample in additive Gaussian noise of unknown variance.

$$\xi_t(i,j) = \frac{\alpha_t^c(i) a_{ij} (p_j^z(\mathbf{z}_{j,t+1}; \lambda^z)/p(\mathbf{z}_{j,t+1}|H_0)) \beta_{t+1}^c(j)}{\sum_{k=1}^{N} \sum_{m=1}^{N} \alpha_t^c(k) a_{km} (p_m^z(\mathbf{z}_{m,t+1}; \lambda^z)/p(\mathbf{z}_{m,t+1}|H_0)) \beta_{t+1}^c(m)} \tag{27}$$

Let H_0 be independent zero-mean Gaussian noise of variance 1. At each time step t, a length-K time-series $\mathbf{x}_t = [x_{t,1} \ldots x_{t,K}]'$ is generated according to the state in effect.

1) Feature Selection: Desirable features are those that are approximately sufficient to distinguish the given hypothesis from H_0 and have a distribution known under H_0. Consider the following feature sets.

- H_1: We use a second-order set of autocorrelation estimates

$$\mathbf{z}_1 = [\hat{r}_0,\ \hat{r}_1,\ \hat{r}_2].$$

- H_2: Let $\{X_i\},\ 1 \le i \le K$ be the length-K FFT of $\mathbf{x}_t$. We use the index and squared value of the largest FFT bin ($a^2_{\max} = \max_i |X_i|^2$), and the average power

$$\mathbf{z}_2 = [i_{\max}, a^2_{\max}, \hat{r}_0].$$

- H_3: We use the time index and value of the largest input sample ($x_{\max} = \max_k x_{t,k}$), and the average power

$$\mathbf{z}_3 = [k_{\max}, x_{\max}, \hat{r}_0].$$

These feature sets are approximately sufficient to discriminate the corresponding state from H_0, while being low in dimension.

2) Feature Dependence on H_0: Notice that $\hat{r}_0$ is included in each feature set because the variance for each state is unknown, while it is fixed under H_0. Thus, the variance estimate has information that can discriminate against H_0. If H_0 had an unknown variance, $\hat{r}_0$ itself would be irrelevant in distinguishing the input data from H_0 and could be discarded, however, it would be necessary to first normalize the other features.

3) Obtaining Exact PDF under H_0: For each of feature sets shown above, the exact joint PDF of the statistics can be derived under the H_0 assumption.

- H_1: For $\mathbf{z}_1$, it is necessary to use a specific autocorrelation function (ACF) estimator whose distribution is known. The PDF of the FFT-method ACF estimates is known exactly [10], [11] and approximations are available with accurate tail behavior for other ACF estimators [9].
- H_2: The FFT bins are Gaussian, independent, and identically distributed under H_0. However, notice that $\hat{r}_0$ is not statistically independent of $a^2_{\max}$. The statistic

$$\hat{r}'_0 \triangleq \frac{1}{K-1} \sum_{i \in \{1 \ldots K\}, i \ne i_{\max}} |X_i|^2$$

however, is independent of $a^2_{\max}$ when $i_{\max}$ is specified. Thus, if

$$\mathbf{z}'_2 = [a^2_{\max}, i_{\max}, \hat{r}'_0]$$

we have

$$p(\mathbf{z}'_2|H_0) = p(a^2_{\max}|i_{\max}, H_0)\, p(\hat{r}'_0|i_{\max}, H_0)\, p(i_{\max}|H_0). \tag{34}$$

Each term in (34) has a known PDF. Notice that except for a scale factor, the first term in (34) is distributed $\chi^2(2)$ (exponential), and the second is $\chi^2(2K-2)$, and the last term is a uniform distribution. It is then possible to obtain $\hat{r}_0$ from the pair $(\hat{r}'_0, a^2_{\max})$. The PDF of $\mathbf{z}_2$ is then easily found by a change of variables.

- H_3: We obtain $p(\mathbf{z}_3|H_0)$ in essentially the same manner as $p(\mathbf{z}_2|H_0)$, however the time dimension takes the place of the frequency index and $x_{\max}$ is Gaussian.

IV. Conclusion

In this paper, we have demonstrated that it is possible to parameterize a HMM using different features for each state. This parameterization requires that the exact densities of the state-dependent feature sets be known for some fixed "common" hypothesis H_0 and that these densities are nonzero for the allowable range of the random variables. The method can lead to an optimal classifier if these feature sets are sufficient statistics for discrimination of the corresponding state from the common state H_0. In practice, this means that more information can be extracted from the raw data for a given maximum PDF dimension. In principle, the reference hypothesis does not need to be common and can be a function of the state; however, we have not explored this possibility in this paper.

References

[1] P. M. Baggenstoss, "Class-specific features in classification," *IEEE Trans. Signal Processing*, vol. 47, pp. 3428–3432, Dec. 1999.

[2] S. Kay, "Sufficiency, classification, and the class-specific feature theorem," *IEEE Trans. Inform. Theory*, to be published.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[4] E. H. Lehmann, *Testing Statistical Hypotheses*. New York: Wiley, 1959.

[5] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. London, U.K.: Charles Griffin, 1979, vol. 2.

[6] H. L. Royden, *Real Analysis*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[7] L. E. Baum, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[8] 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.

[9] S. Kay, A. Nuttall, and P. Baggenstoss, "Accurate evaluation of multidimensional probability density functions (pdfs) for signal processing," *IEEE Trans. Signal Processing*, to be published.

[10] M. H. Quenouille, "The joint distribution of serial correlation coefficients," *Ann. Math. Statist.*, vol. 20, pp. 561–571, 1949.

[11] E. J. Hannan, *Multiple Time Series*. New York: Wiley, 1970.

Paul M. Baggenstoss (S'82–M'82) was born in 1957 in Gastonia, NC. He received the B.S.E.E. degree in 1979 and M.S.E.E. degree in 1982 from Rensselear Polytechnic Institute (RPI), Troy, NY. He received the Ph.D. in statistical signal processing from the University of Rhode Island, Kingston, in 1990, under the supervision of Prof. S. Kay.

From 1979 to 1996, he was with Raytheon Company and joined the Naval Undersea Warfare Center in August 1996. Since then, he has worked in classification and pattern recognition. He was an Adjunct Professor with the University of Connecticut, Storrs, where he taught detection theory and DSP. In February 2000, he began a joint research effort with the Pattern Recognition Chair at the University of Erlangen, Erlangen, Germany.