TABLE II
GROUP WITH THE OPTIMAL STATISTICAL PERFORMANCE FOR THE FIRST-ORDER MARKOV PROCESS

	n															
ρ	6	8	12	16	18	24	30	32	36	40	42	48	54	56	60	64
.9	S_3	Q_2	$C_2 \times S_3$	C_{16}	C_{18}	C_{24}	C_{30}	C_{32}	C_{36}	C_{40}	C_{42}	C_{48}	C_{54}	C_{56}	C_{60}	C_{64}
.99	S_3	Q_2	$C_2 \times S_3$	$C_2 \times Q_2$	$C_3 \times S_3$	$C_3 \times Q_2$	$C_5 \times S_3$	$C_4 \times Q_2$	$S_3 \times S_3$	$Q_2 \times C_5$	$S_3 \times C_7$	$S_3 \times Q_2$	$S_3 \times C_9$	$Q_2 \times C_7$	$S_3 \times C_{10}$	$Q_2 \times Q_2$

TABLE III
DISPERSION FOR THE GROUP FILTER WITH $G = S_3 \times C_2$

ρ	0.5	0.6	0.7	0.8	0.9	0.92	0.94	0.96	0.99
$D^{S_3 \times C_2}$	0.4545	0.4294	0.3944	0.3437	0.2628	0.2398	0.2128	0.1800	0.111

TABLE IV
GROUP WITH THE OPTIMAL STATISTICAL PERFORMANCE FOR THE RANDOM SINE WAVE

	n															
λ	6	8	12	16	18	24	30	32	36	40	42	48	54	56	60	64
.01	C_6	C_8	C_{12}	C_{16}	$S_3 \times C_3$	$S_3 \times C_4$	$S_3 \times C_5$	$Q_2 \times C_4$	$S_3 \times C_6$	$Q_2 \times C_5$	$S_3 \times C_7$	$S_3 \times C_8$	$S_3 \times C_9$	$Q_2 \times C_7$	$S_3 \times C_{10}$	$Q_2 \times C_3$
	S_3	Q_2	$S_3 \times C_2$	$Q_2 \times C_2$												$Q_2 \times Q_2$
.05	C_6	Q_2	$C_2 \times S_3$	$C_2 \times Q_2$	$S_3 \times C_3$	$S_3 \times C_4$	$S_3 \times C_5$	$Q_2 \times C_4$	$S_3 \times S_3$	$Q_2 \times C_5$	$S_3 \times C_7$	$S_3 \times Q_2$	$S_3 \times C_9$	$Q_2 \times C_7$	$S_3 \times C_{10}$	$Q_2 \times Q_2$
	S_3															
.1	S_3	Q_2	$C_2 \times S_3$	$C_2 \times Q_2$	$S_3 \times C_3$	$S_3 \times C_4$	$S_3 \times C_5$	$C_4 \times Q_2$	$C_6 \times S_3$	$C_5 \times Q_2$	$C_7 \times S_3$	C_{48}	C_{54}	C_{56}	C_{60}	C_{64}

ACKNOWLEDGMENT

The authors wish to thank Mr. O. Zimmerman for conducting the computer experiments of Section III.

REFERENCES

[1] D. F. Elliott and K. R. Rao, *Fast Transforms: Algorithms, Analyses and Applications*. New York: Academic, 1983.

[2] H. C. Andrews, *Picture Processing and Digital Filtering*. New York: Springer-Verlag, 1975.

[3] N. Ahmed and K. R. Rao, *Orthogonal Transform for Digital Signal Processing*. New York: Springer-Verlag, 1975.

[4] B. J. Fino and V. R. Algazi, "A unified treatment of discrete fast unitary transforms," *SIAM J. Computing*, vol. 6, pp. 700–717, Dec. 1977.

[5] H. C. Andrews and K. L. Caspari, "Degrees of freedom and modular structure in matrix multiplication," *IEEE Trans. Comput.*, vol. C-19, pp. 16–25, 1970.

[6] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, pp. 90–93, 1974.

[7] J. Pearl, "Basis restricted transformations and performance measures for spectral representations," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 751–752, 1971.

[8] H. S. Kwak, R. Srinivasan, and K. R. Rao, "C-matrix transforms," submitted to *IEEE Trans. Acoust. Speech, Signal Processing*.

[9] W. M. Chen, C. M. Smith, and S. C. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Commun.*, vol. COM-25, pp. 1004–1009, 1977.

[10] K. R. Rao, J. G. K. Kuo, and M. A. Narasimhan, "Slant-Haar transform," *Int. J. Computer Math.*, sec. B, 7, pp. 73–83, 1979.

[11] E. A. Trachtenberg, "Fast Wiener filtering computation technique," in *Proc. 1979 Int. Symp. on the Mathematical Theory of Networks and Systems*, vol. 3, Delft, The Netherlands pp. 174–177.

[12] E. A. Trachtenberg, "Construction of fast unitary transforms which are equivalent to Karhunen–Loeve spectral representations, in *Proc. 1980 IEEE Int. Symp. on EMC*, Baltimore, MD, USA, pp. 376–379.

[13] M. G. Karpovsky and E. A. Trachtenberg, "Some optimization problems for convolution systems over finite groups," *Inform. Contr.*, vol. 34, pp. 227–247, 1977.

[14] J. Pearl, "Optimal dyadic models of time-invariant systems," *IEEE Trans. Comput.*, vol. C-24, 1975, pp. 598–602.

[15] W. K. Pratt, "Generalized Wiener filtering computation technique," *IEEE Trans. Comput.*, vol. C-21, pp. 636–641, 1972.

[16] J. Pearl, "Walsh processing of random signals," *IEEE Trans. Electromagn. Compat.*, vol. EMC-13, pp. 137–141, 1971.

[17] ——, "On coding and filtering stationery signals by discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 229–232, 1973.

[18] ——, "Asymptotic equivalence of spectral representations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 547–551, 1975.

[19] Y. Yemini and J. Pearl, "Asymptotic properties of discrete unitary transforms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, pp. 366–371, 1979.

[20] M. G. Karpovsky, "Fast Fourier transforms on finite non-Abelian groups," *IEEE Trans. Comput.*, vol. C-26, pp. 1028–1030, 1977.

[21] P. J. Nicholson, "Algebraic theory of finite Fourier transforms," *J. Comput. Syst. Sci.*, pp. 524–547, 1971.

[22] L. Dornhoff, *Group Representation Theory*. New York: Marcel Dekker, 1971.

[23] M. G. Karpovsky and E. A. Trachtenberg, "Fourier transform over finite groups for error detection and error correction in computation channels," *Inform. Contr.*, vol. 40, pp. 335–358, 1979.

[24] M. D. Flickner and N. Ahmed, "A derivation for the discrete cosine transform," *Proc. IEEE*, vol. 70, pp. 1132–1134, 1982.

[25] M. Hamidi and J. Pearl, "Comparison of cosine and Fourier transforms of Markov-1 signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 428–429, 1976.

[26] E. A. Trachtenberg, "Systems over finite groups as suboptimal Wicner filters: A comparative study," presented at *1983 Int. Symp. on Mathematical Theory of Systems and Networks*, Beer-Sheva, Israel, 1983.

[27] M. G. Karpovsky, *Finite Orthogonal Series in the Design of Digital Devices*. New York: Wiley, 1976.

[28] E. A. Trachtenberg, "Construction of Group Transforms Subject to Several Performance Criteria," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-33, pp. 1521–1531, 1985.

Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains

B. H. JUANG, STEPHEN E. LEVINSON, SENIOR MEMBER, IEEE, AND M. M. SONDHI

***Abstract*—To use probabilistic functions of a Markov chain to model certain parameterizations of the speech signal, we extend an estimation technique of Liporace to the cases of multivariate mixtures, such as Gaussian sums, and products of mixtures. We also show how these problems relate to Liporace's original framework.**

Manuscript received September 10, 1984; revised July 12, 1985. This work was presented at the 1985 IEEE International Symposium on Information Theory, Brighton, England, June 24-28.

The authors are with Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 8406633.

Introduction

In a recently published paper, Liporace [6] derived a method for estimating the parameters of a broad class of elliptically symmetric probabilistic functions of a Markov chain. The corollary to that work presented here was motivated by the desire to use this general technique to model the speech signal for which it is known [4], [8] that, unfortunately, certain of its most useful parameterizations do not possess the prescribed symmetry. Since any continuous probability density function can be approximated arbitrarily closely by a normal mixture [9], it is reasonable to use such constructs to avoid the restrictions imposed by the requirement of elliptical symmetry. In this correspondence we adapt the method and proof of [6] to two types of mixture densities.

Nomenclature

Throughout this presentation we shall, where possible, adopt the notation used in [6]. Consider an unobservable n-state Markov chain with state transition matrix $\boldsymbol{A} = [a_{ij}]_{n\times n}$. Associated with each state j of the hidden Markov chain is a probability density function $b_j(\boldsymbol{x})$, of the observed d-dimensional random vector $\boldsymbol{x}$. Here we shall consider densities of the form

$$b_j(x) = \sum_{k=1}^{m} c_{jk}\mathcal{N}(\boldsymbol{x}, \mu_{jk}, \boldsymbol{U}_{jk}) \tag{1}$$

where m is known; $c_{jk} \geq 0$ for $1 \leq j \leq n$, $1 \leq k \leq m$; $\sum_{k=1}^{m} c_{jk} = 1$ for $1 \leq j \leq n$; and $\mathcal{N}(\boldsymbol{x}, \mu, \boldsymbol{U})$ denotes the d-dimensional normal density function of mean vector μ and covariance matrix $\boldsymbol{U}$.

It is convenient then to think of our hidden Markov chains as being defined over a parameter manifold $\Lambda = \{\mathcal{A}^n \times \mathcal{C}^m \times \mathcal{R}^d \times \mathcal{U}^d\}$, where $\mathcal{A}^n$ is the set of all $n \times n$ row-wise stochastic matrices; $\mathcal{C}^m$ is the set of all $m \times n$ row-wise stochastic matrices; $\mathcal{R}^d$ is the usual d-dimensional Euclidean space; and $\mathcal{U}^d$ is the set of all $d \times d$ real symmetric positive definite matrices. Then for a given sequence of observations, $\boldsymbol{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_T$, of the vector $\boldsymbol{x}$ and a particular choice of parameter values $\lambda \in \Lambda$, we can efficiently evaluate the likelihood function, $L_\lambda(\boldsymbol{O})$, of the hidden Markov chain by the forward–backward method of Baum [1].

The forward and backward partial likelihoods, $\alpha_t(j)$ and $\beta_t(i)$, are computed recursively from

$$\alpha_t(j) = \left[\sum_{i=1}^{n} \alpha_{t-1}(i)\, a_{ij}\right] b_j(\boldsymbol{O}_t) \tag{2}$$

and

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} b_j(\boldsymbol{O}_{t+1})\beta_{t+1}(j) \tag{3}$$

respectively. The recursion is initialized by setting $\alpha_0(1) = 1$, $\alpha_0(j) = 0$ for $2 \leq j \leq n$ and $\beta_T(i) = 1$ for $1 \leq i \leq n$, whereupon we may write

$$\mathcal{L}_\lambda(\boldsymbol{O}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_t(i)\, a_{ij} b_j(\boldsymbol{O}_{t+1})\beta_{t+1}(j) \tag{4}$$

for any t between 1 and $T-1$.

The Estimation Algorithm

The parameter estimation problem is then one of maximizing $\mathcal{L}_\lambda(\boldsymbol{O})$ with respect to λ for a given $\boldsymbol{O}$. One way to maximize $\mathcal{L}_\lambda$ is to use conventional methods of constrained optimization. Liporace, on the other hand, advocates a reestimation technique analogous to that of Baum *et al.* [1], [2]. It is essentially a mapping $\mathcal{T}: \Lambda \to \Lambda$ with the property that

$$\mathcal{L}_{\mathcal{T}(\lambda)}(\boldsymbol{O}) \geq \mathcal{L}_\lambda(\boldsymbol{O}) \tag{5}$$

with equality iff λ is a critical point of $\mathcal{L}_\lambda(\boldsymbol{O})$, that is, $\nabla\mathcal{L}_\lambda(\boldsymbol{O}) = 0$. Thus a recursive application of $\mathcal{T}$ to some initial value of λ converges to a local maximum (or possibly an inflection point) of the likelihood functions. Liporace's result [6] relaxed the original requirement of Baum *et al.* [2] that $b_j(\boldsymbol{x})$ be strictly log concave to the requirement that it be strictly log concave and/or elliptically symmetric. We will further extend the class of admissible pdf's to mixtures and products of mixtures of strictly log concave and/or elliptically symmetric densities.

For the present problem we will show that a suitable mapping $\mathcal{T}$ is given by the following equations:

$$\bar{a}_{ij} = \mathcal{T}(a_{ij}) = \frac{\sum_{t=1}^{T-1} \alpha_t(i)\, a_{ij} b_j(\boldsymbol{O}_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)}, \tag{6}$$

$$\bar{c}_{jk} = \mathcal{T}(c_{jk}) = \frac{\sum_{t=1}^{T} \rho_t(j,k)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}, \tag{7}$$

$$\bar{\mu}_{jk} = \mathcal{T}(\mu_{jk}) = \frac{\sum_{t=1}^{T} \rho_t(j,k)\beta_t(j)\boldsymbol{O}_t}{\sum_{t=1}^{T} \rho_t(j,k)\beta_t(j)}, \tag{8}$$

and

$$\bar{\boldsymbol{U}}_{jk} = \mathcal{T}(\boldsymbol{U}_{jk}) = \frac{\sum_{t=1}^{T} \rho_t(j,k)\beta_t(j)(\boldsymbol{O}_t - \mu_{jk})(\boldsymbol{O}_t - \mu_{jk})'}{\sum_{t=1}^{T} \rho_t(j,k)\beta_t(j)} \tag{9}$$

for $1 \leq i, j \leq n, 1 \leq k \leq m$ and $1 \leq r, s \leq d$.

In (7)–(9)

$$\rho_t(j,k) = \begin{cases} c_{jk}\dfrac{\partial b_j}{\partial c_{jk}}\bigg|_{\boldsymbol{O}_1}, & \text{for } t = 1, \\ \displaystyle\sum_{i=1}^{n} \alpha_{t-1}(i)\, a_{ij} c_{jk}\frac{\partial b_j}{\partial c_{jk}}\bigg|_{\boldsymbol{O}_t}, & \text{for } 1 < t \leq T. \end{cases} \tag{10}$$

(For fixed k, $\rho_t(j,k)$ is formally identical to $\rho_t(j)$, as defined by Liporace.)

Proof of the Formulas: Equations (6) and (7) for the reestimation of a_{ij} and c_{jk} follow directly from a theorem of Baum and Sell [3] because the likelihood function $\mathcal{L}_\lambda(\boldsymbol{O})$ given in (4) is a polynomial with nonnegative coefficients in the variables $a_{ij}, c_{jk}, 1 \leq i, j \leq n, 1 \leq k \leq m$.

To prove (8) and (9) our strategy, following Liporace, is to define an appropriate auxiliary function $Q(\lambda, \bar{\lambda})$. This function will have the property that $Q(\lambda, \bar{\lambda}) > Q(\lambda, \lambda)$ implies $\mathcal{L}_{\bar{\lambda}}(\boldsymbol{O}) > \mathcal{L}_\lambda(\boldsymbol{O})$. Further, as a function of $\bar{\lambda}$ for any fixed λ, $Q(\lambda, \bar{\lambda})$ will have a unique global maximum given by (6)–(9).

As a first step to derive such a function we express the likelihood function as a sum over the set, $\mathcal{S}$, of all state sequences $\boldsymbol{S}$:

$$\mathcal{L}_\lambda(\boldsymbol{O}) = \sum_{\mathcal{S}} \mathcal{L}_\lambda(\boldsymbol{O}, \boldsymbol{S}) = \sum_{\mathcal{S}} \prod_{t=1}^{T} a_{s_{t-1}s_t} \sum_{k=1}^{m} c_{s_t k}\mathcal{N}(\boldsymbol{O}_t, \mu_{s_t k}, \boldsymbol{U}_{s_t k}). \tag{11}$$

Let us partition the likelihood function further by choosing a particular sequence, $\boldsymbol{K} = (k_1, k_2, \cdots, k_T)$, of mixture densities. As in the case of state sequences we denote the set of all mixture sequences as $\mathscr{K} = \{1, 2, \cdots, m\}^T$. Thus for some particular $\boldsymbol{K} \in \mathscr{K}$ we can write the joint likelihood of $\boldsymbol{O}$, $\boldsymbol{S}$, and $\boldsymbol{K}$ as

$$\mathscr{L}_\lambda(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K}) = \prod_{t=1}^{T} a_{s_{t-1}s_t} \mathscr{N}\left(\boldsymbol{O}_t, \mu_{s_t k_t}, \boldsymbol{U}_{s_t k_t}\right) c_{s_t k_t}. \quad (12)$$

We have now succeeded in partitioning the likelihood function as

$$\mathscr{L}_\lambda(\boldsymbol{O}) = \sum_{\boldsymbol{S} \in \mathscr{S}} \sum_{\boldsymbol{K} \in \mathscr{K}} \mathscr{L}_\lambda(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K}). \quad (13)$$

In view of the similarity of the representation (13) to that of $\mathscr{L}_\lambda$ in [6], we now define the auxiliary function

$$Q(\lambda, \bar{\lambda}) = \sum_{\boldsymbol{S}} \sum_{\boldsymbol{K}} \mathscr{L}_\lambda(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K}) \log \mathscr{L}_{\bar{\lambda}}(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{K}). \quad (14)$$

When the expressions for $\mathscr{L}_\lambda$ and $\mathscr{L}_{\bar{\lambda}}$ derived from (12) are substituted in (14), we get

$$Q(\lambda, \bar{\lambda}) = \sum_{\boldsymbol{S} \in \mathscr{S}} \sum_{\boldsymbol{K} \in \mathscr{K}} \sum_{t=1}^{T} \gamma_{s_t k_t t} \log \mathscr{N}\left(\boldsymbol{O}_t, \bar{\mu}_{s_t k_t}, \bar{\boldsymbol{U}}_{s_t k_t}\right), \quad (15)$$

where $\gamma_{s_t k_t t} \geq 0$. The innermost summation in (15) is formally identical to that used by Liporace in his proof; therefore, the properties which he demonstrated for his auxiliary function with respect to μ and $\boldsymbol{U}$ hold in our case as well, thus giving us (8) and (9). We may thus conclude that (5) is correct for $\mathscr{T}$ defined by (6)–(9). Furthermore, the parameter separation made explicit in (12)–(15) allows us to apply the same algorithm to mixtures of strictly log concave densities and/or elliptically symmetric densities as treated by Liporace in [6].

Discussion

In [6] Liporace notes that by setting $a_{ij} = p_j, 1 \leq j \leq n, \forall i$, the special case of a single mixture can be treated. It is natural then to think of using a model with n clusters of m states, each with a single associated Gaussian density function as a way of treating the Gaussian mixture problem considered here.

The transformation can be accomplished in the following way. First we expand the state space of our n-state model as shown in Fig. 1, in which we have added states j_0 through j_m for each state j in the original Markov chain. Associated with states $j_1, j_2, \cdots, j_m$ are distinct Gaussian densities corresponding to the m terms of the jth Gaussian mixture in our initial formulation. The transitions exiting state j have probabilities equal to the corresponding mixture weights. State j_0 is a distinguished state that is entered with probability 1 from the other new states, exits to state j with probability a_{jj}, and generates no observation in so doing. The transition matrix for this configuration can be written down by inspection. A large number of the entries in it will be zero or unity. As these are unaltered by (6) and (7), they need not be reestimated. Using this reconfiguration of the state diagram, Liporace's formulas can be used in case $b_j(\boldsymbol{x})$ is any mixture of elliptically symmetric densities.

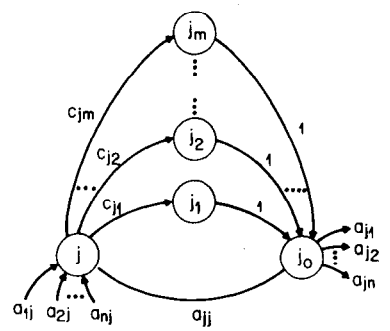

Fig. 1. Equivalent $m + 2$ state configuration for each state with m-term Gaussian mixture.

A variant on the Gaussian mixture theme results from using $b_j(\boldsymbol{x})$ of the form of a product of mixtures,

$$b_j(x) = \prod_{r=1}^{D} \sum_{k=1}^{m} c_{jkr} \mathscr{N}\left(\boldsymbol{x}_r, \mu_{jkr}, U_{jkr}\right). \quad (16)$$

What we have considered so far is the special case of (16) for $D = 1$.

From the structure of our derivation it is clear that for hidden Markov chains having densities of the form (16), reestimation formulas can be derived as before by solving $\nabla_{\bar{\lambda}} Q(\lambda, \bar{\lambda}) = 0$. Such solutions will yield results quite analogous to (6)–(10). Note that this case too can be represented as a reconfiguration of the state diagram.

One numerical difficulty which may be manifest in the methods described is the phenomenon noted by Nadas [7] in which one or more of the mean vectors converge to a particular observation while the corresponding covariance matrix approaches a singular matrix. Under these conditions, $\mathscr{L}_\lambda(\boldsymbol{O}) \to \infty$ but the value of λ is meaningless. A practical, if unedifying, remedy for this difficulty is to try a different initial λ. Alternatively, one can drop the offending term from the mixture since it is only contributing at one point of Λ.

Finally, we call attention to two minor facets of these algorithms. First, for flexibility in modeling, the number of terms in each mixture may vary with state, so that m in (1) could as well be m_j. A similar dependence on dimension results if m in (16) is replaced by m_{jr}. In either case, the constraints on the mixture weights must be satisfied.

Second, for realistic numbers of observations, for example, $T \geq 5000$, the reestimation formulas will underflow on any existing computer. The basic scaling mechanism described in [5] can be used to alleviate the problem but must be modified to account for the fact that the $\rho_t \beta_t$ product will be missing the tth scale factor. To divide out the product of scale factors, the tth summand in both numerator and denominator of (7), (9), and (10) must be multiplied by the missing coefficient.

At this writing numerical experiments based on Monte Carlo simulations and classification experiments using real speech signals are being conducted. We hope to report the results of these studies upon their completion.

References

[1] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process," in *Inequalities*, vol. III, O. Shisha, Ed. New York: Academic, 1972, pp. 1–8.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains" *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[3] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds," *Pac. J. Math.*, vol. 27, pp. 211–227, 1968.

[4] A. H. Gray and J. D. Markel, "Quantization and bit allocation in speech processing," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-24, pp. 459–473, Dec. 1976.

[5] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell. Syst. Tech. J.*, vol. 62, pp. 1035–1074, Apr. 1983.

[6] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, Sept. 1982.

[7] A. Nadas, "Hidden Markov chains, the forward-backward algorithm and initial statistics," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-31, pp. 504–506, Apr. 1983.

[8] L. R. Rabiner, J. G. Wilpon, and J. G. Ackenhusen, "On the effects of varying analysis parameters on an LPC-based isolated word recognizer," *Bell Syst. Tech. J.*, vol. 60, pp. 893–911, 1981.

[9] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.