A NEW SIGNAL MODEL AND IDENTIFICATION ALGORITHM FOR HIDDEN SEMI-MARKOV SIGNALS

Mehran Azimi, Panos Nasiopoulos, Rabab K. Ward

Electrical and Computer Engineering department, University of British Columbia

ABSTRACT

Markovian models form a powerful tool for modelling physical signals. In this approach, a signal generation model is employed, and its parameters are estimated from signal samples. In this paper, we present a novel signal generation model for Hidden Semi-Markov Models, HSMMs. Our model results in a significantly easier and more efficient parameters identification method. Instead of the constant probabilities presently used for modelling state transitions, we use state transition probabilities that are *state-duration dependant*. We then develop a parameter identification algorithm based on the maximum likelihood criterion.

Our numerical results show that our parameter identification algorithm can successfully and more efficiently estimate the actual values of the model parameters of an HSMM signal.

1. INTRODUCTION

Hidden Markov Models (HMMs) have proved to be a powerful tool in signal modelling, and have been widely used in many engineering applications including speech processing, signal estimation, queuing networks, etc., [1]. However, HMMs have a limitation in modelling the 'state durations' of physical signals. The state duration for each state of an HMM is defined as the time spent in that state before making a transition to another state. In an HMM, the probability of leaving a state is constant. Hence, it can be easily shown that the density of state durations have the form of a Geometrical probability mass function (pmf). However, this Geometrical probability mass function is inappropriate for modelling the state duration of a large class of physical signals. Therefore a more sophisticated class of Markov models, called Hidden Semi-Markov Models (HSMMs), are used, where the duration densities are modelled in some non-Geometrical form. Generally speaking, HSMMs are more powerful than HMMs, however, they are more complex. Specifically, the parameter identification methods for HSMMs are more complicated than HMMs.

In this paper, we present a novel signal generation model for HSMMs, which leads to easier and more efficient parameter identification algorithms. We introduce a new vector variable to the traditional HMMs, named 'state duration' variable, $\boldsymbol{d}_t$. Given that the state at time t is i, the i^{th} entry of $\boldsymbol{d}_t$, $d_t(i)$, denotes the time that the signal has spent in state i until time t. We use *state duration dependent transition probabilities* in our model. This assumes the probability of transition from state i to state j is not constant and depends on $d_t(i)$. We model the state durations using parameterized probability density functions. We will employ a special mathematical representation of the signal state space as in [1]. This facilitates the derivation of $\boldsymbol{d}_t$ from $\boldsymbol{d}_{t-1}$.

Then, we present an algorithm for parameter identification of HSMMs based on our signal generation model from a given a set of observations from an HSMM signal. These parameters are the state transition probabilities, the parameters of the state duration model and the parameters of the observation density of each state. The problem of identification of HSMMs is conceptually similar to the identification of HMMs. There is a powerful method available in the HMMs case, known as Baum-Welsh algorithm, which finds the maximum likelihood estimate of model parameters using the Expectation-Maximization (EM) algorithm [2, 3]. This algorithm has been extended to the context of HSMMs using either 'explicit state duration modelling' [4, 5, 3, 6] or 'parametric state duration modelling' [7]. Current methods based on these two approaches have the major drawback of greatly increased computational load compared to the HMM case. More precisely, if we let the maximum state duration in an HSMM for all states be D time units, then it can be shown that current approaches increase the memory usage by a factor of D and the computation load by a factor of $D^2/2$, when compared to the EM algorithm for HMMs. Since D is usually large in many applications (e.g., $D = 25$ in most speech processing applications), the computational load of these algorithms become extensively high.

Our algorithm for identifying the model parameters of an HSMM is a variant of the EM algorithm [2]. Our algorithm is based on our new signal model, and finds the local maximum likelihood estimate of the model parameters. Our algorithm has the advantage of requiring significantly much less computational effort compared to available methods. Therefore, our identification algorithm is useful in a larger set of practical applications. Also, our method does

not result in over parameterizations of the model and employs only $N^2 + 3N$ parameters, which is very close to the $N^2 + 2N$ parameters used in an HMM.

The rest of this paper is organized as follows: In section 2, we present our signal model for HSMM. Our algorithm for off-line identification of HSMMs is presented in section 3. In section 4, we present numerical results of implementing our algorithm for identification of HSMMs. In section 5, we present a conclusion of the presented methods.

2. SIGNAL MODEL

We consider a signal model where the state of the signal at time t, $\boldsymbol{s}_t$, $t \in \mathbb{N}$, is determined by a finite-state discrete-time semi-Markov chain. We assume the initial state $\boldsymbol{s}_1$ is given or its distribution is known. The state space has N distinct states. Without loss of generality, we assume $\boldsymbol{s}_t$ takes its values from the set $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_N\}$, where $\boldsymbol{e}_i$ is a $N \times 1$ vector with unity as the i^{th} element and zeros elsewhere. The semi-Markov property of the model implies that the probability of a transition from state $\boldsymbol{e}_j$ to $\boldsymbol{e}_i$ at time t depends on the duration spent in state $\boldsymbol{e}_j$ prior to time t. This can be written as

$$\begin{aligned}\mathbb{P}(\boldsymbol{s}_{t+1} = \boldsymbol{e}_i | \boldsymbol{s}_t = \boldsymbol{e}_j, \boldsymbol{s}_{t-1} = \boldsymbol{e}_k, \cdots, \boldsymbol{s}_1 = \boldsymbol{e}_l) \\ = \mathbb{P}(\boldsymbol{s}_{t+1} = \boldsymbol{e}_i | \boldsymbol{s}_t = \boldsymbol{e}_j, d_t(j)) \qquad (1)\end{aligned}$$

where $d_t(j)$ is defined as the duration spent in state j prior to time t. For each time t, we define the *'state duration'* vector $\boldsymbol{d}_t$ of size $N \times 1$ where

$$\boldsymbol{d}_t(j) = \begin{cases} d_t(j) & \text{if } \boldsymbol{s}_t = \boldsymbol{e}_j \\ 1 & \text{if } \boldsymbol{s}_t \neq \boldsymbol{e}_j \end{cases} \qquad (2)$$

$d_t(j)$ is easily constructed from $d_{t-1}(j)$ as $d_t(j) = s_t(j) \times d_{t-1}(j) + 1$, which can be written in vector format as $\boldsymbol{d}_t = \boldsymbol{s}_t \odot \boldsymbol{d}_{t-1} + 1$, where $\odot$ denotes element-by-element product.

We model the state duration densities (i.e. density of $d_t(i)$'s) with a parametric *probability mass function*, pmf, $\phi_i(d)$. That is, the probability that $\boldsymbol{s}_t$ stays exactly d time units in state i is given by $\phi_i(d)$. $\phi_i(d)$ should be selected such that it adequately captures the properties of the signal under study. Hence, the selection of $\phi_i(d)$ should be justified by some evidence from samples of the signal. Even though the state durations in a semi-Markov chain are inherently discrete, it is noted in many studies that continuous parametric density functions are also suitable for modelling state durations in many applications, including speech processing [6, 7]. In this approach, state durations are modelled with the best fitting parametric *probability density function*, pdf, and then the discrete counterpart of this density function is taken as the best pmf. That is, if $\phi_i(x)$ is the continuous pdf of the state duration of the i^{th} state, then the probability that the signal stays in state i for exactly d time units is given by $\int_{d-1}^{d} \phi_i(x)dx$. Since negative state durations are not physically meaningful, it is usually more appropriate to select $\phi_i(x)$ from the family of exponential distributions [6]. Specifically, the family of Gamma distributions are considered in [7] for speech processing applications. In this paper, we assume that $\phi_i(x)$ is a *Gamma* distribution function with *shape parameter* ν_i and *scale parameter* η_i, that is

$$\phi_i(x) = \frac{\eta_i^{\nu_i}}{\Gamma(\nu_i)} x^{\nu_i - 1} e^{-\eta_i x} \qquad (0 < x < \infty) \qquad (3)$$

where Γ is the gamma function. The mean and variance of ϕ_i are ν_i/η_i and ν_i/η_i^2 respectively. Note that the signal model we present here is applicable with minor changes to HSMM signals whose state duration densities are modelled with a pdf other than Gamma. Furthermore, let $\Phi_i(x)$ denote the cumulative distribution function of $\phi_i(x)$, i.e., $\Phi_i(d) = \int_0^d \phi_i(x)dx$.

We construct our model for HSMMs using state duration dependant transition probabilities. We define the state transition matrix $\boldsymbol{A}_{\boldsymbol{d}_t}$, as $\boldsymbol{A}_{\boldsymbol{d}_t} = [a_{ij}(\boldsymbol{d}_t)]$ where $a_{ij}(\boldsymbol{d}_t) = \mathbb{P}(s_{t+1} = \boldsymbol{e}_j | s_t = \boldsymbol{e}_i, d_t(i))$. Clearly, $a_{ij}(\boldsymbol{d}_t)$'s are not constant and change in time; however, we will denote $a_{ij}(\boldsymbol{d}_t)$ with a_{ij} for notational simplicity. For the diagonal elements of $\boldsymbol{A}_{\boldsymbol{d}_t}$, a_{ii}'s, it is easily shown that

$$a_{ii} = \frac{1 - \Phi_i(d_t(i))}{1 - \Phi_i(d_t(i) - 1)} \qquad (4)$$

The probability that the state process $\boldsymbol{s}_t$ stays in the i^{th} state during its visit to this state for exactly d time units is given by $(1 - a_{ii}(d)) \cdot \prod_{k=1}^{d-1} a_{ii}(k)$. By substituting a_{ii} from (4), it is easily shown that the probability density function of the state space durations is actually equal to the selected model $\phi_i(d)$.

For $i \neq j$, a_{ij} is the probability of leaving state i and entering state j, and is given by

$$a_{ij} = (1 - a_{ii}) \cdot a_{ij}^o \qquad (5)$$

where $a_{ij}^o = \mathbb{P}(s_{t+1} = \boldsymbol{e}_j | s_t = \boldsymbol{e}_i, i \neq j)$ is defined as the probability of transition from state i to state j, knowing that the signal leaves state i. We write the matrix $\boldsymbol{A}_{\boldsymbol{d}_t}$ in terms of a diagonal matrix $\boldsymbol{P}(\boldsymbol{d}_t)$ representing the recurrent state transition probabilities, and a constant matrix $\boldsymbol{A}^o$ representing the non-recurrent state transition probabilities.

$$\boldsymbol{A}_{\boldsymbol{d}_t} = \boldsymbol{P}(d_t) + (\boldsymbol{I} - \boldsymbol{P}(d_t))\boldsymbol{A}^o \qquad (6)$$

$$p_{ij}(\boldsymbol{d}_t) := \begin{cases} 0 & ,i \neq j \\ \dfrac{1 - \Phi_i(d_t(i))}{1 - \Phi_i(d_t(i) - 1)} & ,i = j \end{cases} \qquad (7)$$

Note that a^o_{ij}'s are constrained to $\sum_{j=1}^N a^o_{ij} = 1$. Since $\boldsymbol{P}(\boldsymbol{d}_t)$ is a diagonal matrix, one can show that $\sum_{j=1}^N a_{ij}(\boldsymbol{d}_t) = 1$ for all t.

Hence, the hidden state process $\boldsymbol{s}_t$ evolves in time based on the following equations:

$$\begin{aligned} \boldsymbol{s}_{t+1} &= \boldsymbol{A}_{\boldsymbol{d}_t} \cdot \boldsymbol{s}_t + \boldsymbol{v}_{t+1} \\ \boldsymbol{A}_{\boldsymbol{d}_t} &= \boldsymbol{P}(\boldsymbol{d}_t) + (\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{d}_t)) \cdot \boldsymbol{A}^o \\ \boldsymbol{d}_{t+1} &= \boldsymbol{s}_{t+1} \odot \boldsymbol{d}_t + 1 \end{aligned} \tag{8}$$

where $\boldsymbol{v}_{t+1}$ is a Martingale increment; that is, $\mathbb{E}(\boldsymbol{v}_{t+1}|\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_t) = 0$.

The state process $\boldsymbol{s}_t$ is hidden and is not observed. We observe the *observation process* y_t, where the probabilistic distribution of y_t is determined by the state at t, $\boldsymbol{s}_t$. In this study, we assume that for each state i, y_t has a normal distribution. That is, if $\boldsymbol{s}_t = e_i$ then

$$b_i(y_t) := \mathbb{P}(y_t|\boldsymbol{s}_t = \boldsymbol{e}_i) = \mathcal{N}(y_t; \mu_i, \sigma_i^2) \tag{9}$$

where μ_i and σ_i^2 are the mean and standard deviation of the observation process y_t for state i. y_t may be written as $y_t = \langle \boldsymbol{\mu}, \boldsymbol{s}_t \rangle + \langle \sqrt{\boldsymbol{\sigma}^2}, \boldsymbol{s}_t \rangle w_t$ where $\boldsymbol{\mu} = \left[\mu_1, \mu_2, \cdots, \mu_N\right]$, $\boldsymbol{\sigma}^2 = \left[\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2\right]$, $\langle ., . \rangle$ denotes the inner product and w_t is Gaussian white noise with zero mean and variance 1.

2.1. Model Parameterizations

There are $N^2 + 3N$ parameters that define an HSMM signal in our model. These parameters are the $N^2 - N$ non-recurrent transition probabilities a^o_{ij}, the mean and variance of the observation process, μ_i and σ_i^2 for $1 \le i \le N$, and the parameters of the state-duration densities η_i and ν_i for $1 \le i \le N$. We define $\boldsymbol{\theta}$ as a vector containing all the model parameters; $\boldsymbol{\theta} = [\mu_1, \ldots, \mu_N, \sigma_1^2, \ldots, \sigma_N^2, a^o_{12}, \ldots, a^o_{N-1,N}, \eta 1, \ldots, \eta N, \nu_1, \ldots, \nu_N]'$

3. OFFLINE IDENTIFICATION OF HSMMS

Given a set of observations from an HSMM signal, $\mathcal{Y}_T = \{y_1, y_2, \ldots, y_T\}$, we like to estimate $\boldsymbol{\theta}$, the parameters of the HSMM model. The algorithm we use is a variant of the EM algorithm [2]. We first initialize $\boldsymbol{\theta}$ to an initial guess. Analogous to the EM algorithm for identifications of HMMs [3], in the **E** step of the algorithm we define a set of probabilistic measures, which describe the evolution of the hidden state variable $\boldsymbol{s}_t$. We define the 'forward variables' $\alpha_t(i)$ as $\alpha_t(i) := \mathbb{P}(\boldsymbol{s}_t = \boldsymbol{e}_i, y_1, y_2, \ldots, y_t|\boldsymbol{\theta})$. Let $\hat{\boldsymbol{d}}_t = \left[\hat{d}_t(i)\right]'$, where $\hat{d}_t(i) = \mathbb{E}(d_t(i)|\boldsymbol{s}_t = i, \boldsymbol{\theta}, y_1, y_2, \ldots, y_t)$ is our estimate of the state-duration variable for state i at time t. $\hat{\boldsymbol{d}}_t$ is initialized to $\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}'$ for $t = 1$. We construct $\hat{d}_{t+1}(i)$ iteratively as

$$\begin{aligned} \hat{d}_{t+1}(i) &= 1 + \mathbb{E}(\boldsymbol{s}_t(i)|y_1, y_2, \ldots, y_t, \boldsymbol{\theta}) \cdot \hat{d}_t(i) \\ &= 1 + \frac{\alpha_t(i)}{\sum_{i=1}^N \alpha_t(i)} \cdot \hat{d}_t(i), \quad 1 \le i \le N \end{aligned}$$

The state transition matrix $\boldsymbol{A}_{\boldsymbol{d}_t}$ is updated for each t as $\boldsymbol{A}_{\boldsymbol{d}_t} = \boldsymbol{P}(\hat{\boldsymbol{d}}_t) + (\boldsymbol{I} - \boldsymbol{P}(\hat{\boldsymbol{d}}_t))\boldsymbol{A}^o$.

The forward variable $\alpha_t(i)$ for $t = 1$ is initialized to the given initial state, i.e., $\alpha_1(i) = \boldsymbol{s}_1(i)$ for $1 \le i \le N$. The other forward variables are constructed iteratively as $\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) \cdot a_{ij}\right] b_j(y_{t+1})$.

Similarly, the backward variables $\beta_t(i)$ are defined as $\beta_t(i) := \mathbb{P}(y_{t+1} y_{t+2} \ldots y_T|\boldsymbol{s}_t = e_i, \boldsymbol{\theta})$. β_t's are computed by initializing $\beta_T(i) = 1$ for $1 \le i \le N$, and constructing the other variables iteratively as $\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) \cdot a_{ij} \cdot b_j(y_{t+1})$.

In the **M** step of the algorithm, the model parameters are updated to the maximum likelihood estimate of the model parameters computed from the forward-backward variables in the **E** step. There are different approaches to obtaining the update equations, which all result in the same update equations [3]. We use the concept of counting the event occurrence to find the update equations. It can be easily shown that the update equation for $a^o_{i,j}$'s, μ_i's and σ_i^2's are identical to the formulae presented in [3] for identification of HMMs. We estimate the mean and variance of state-duration pmf's for state i, $\mu_{i,s}$ and $\sigma^2_{s,i}$ as

$$\mu_{s,i} = \frac{\sum_{t=1}^{T-1} \left(\alpha_t(i) \sum_{j=1, j\neq i}^N a_{ij} b_j(y_{t+1}) \beta_{t+1}(j)\right) \hat{d}_t(i)}{\sum_{t=1}^{T-1} \left(\alpha_t(i) \sum_{j=1, j\neq i}^N a_{ij} b_j(y_{t+1}) \beta_{t+1}(j)\right)}$$

$$\sigma^2_{s,i} = \frac{\sum_{t=1}^{T-1} \left(\alpha_t(i) \sum_{j=1, j\neq i}^N a_{ij} b_j(y_{t+1}) \beta_{t+1}(j)\right) (\hat{d}_t(i) - \mu_{s,i})^2}{\sum_{t=1}^{T-1} \alpha_t(i) \sum_{j=1, j\neq i}^N a_{ij} b_j(y_{t+1}) \beta_{t+1}(j)}$$

Then, η_i and ν_i are computed as

$$\nu_i = \mu^2_{i,s}/\sigma^2_{i,s} \qquad \eta_i = \mu_{i,s}/\sigma^2_{i,s} \tag{10}$$

The algorithm repeats the **E** and **M** steps, until $\boldsymbol{\theta}$ converges to a constant vector. Our forward-backward algorithm has the computational complexity of $\mathcal{O}(N^2 T)$ per pass and requires a memory of $3NT$ because all the forward-backward variables and estimate of the state duration variables need to be stored.

It is noted that as t increases (decreases), α_t's (β_t's) decrease very fast, and can quickly exceed the numerical range of any computer. To avoid this, we suggest to use a scaling scheme similar to the scheme used in [3, 8] for the HMM

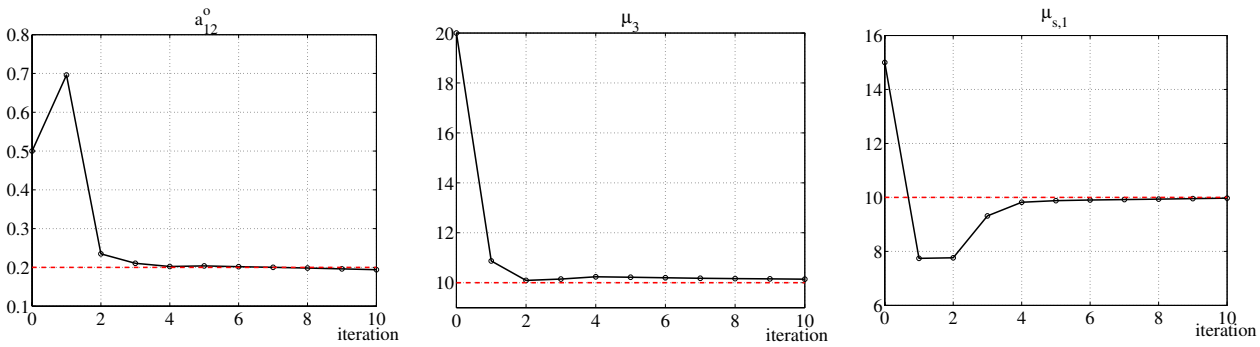

Fig. 1. Parameter estimates versus the iteration number. The dotted lines show the actual values of the parameters.

Parameter	Actual parameter values	Initial parameter values
$\boldsymbol{A}^o$	$\begin{bmatrix} 0 & 0.20 & 0.80 \\ 0.50 & 0 & 0.50 \\ 0.30 & 0.70 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.50 & 0.50 \\ 0.10 & 0.00 & 0.90 \\ 0.50 & 0.50 & 0.00 \end{bmatrix}$
$\boldsymbol{\mu}$	$\begin{bmatrix} -10 & 0 & 10 \end{bmatrix}'$	$\begin{bmatrix} -20 & 4 & 20 \end{bmatrix}'$
$\boldsymbol{\sigma}^2$	$\begin{bmatrix} 8 & 8 & 8 \end{bmatrix}'$	$\begin{bmatrix} 10 & 10 & 10 \end{bmatrix}'$
$\boldsymbol{\mu}_s$	$\begin{bmatrix} 10 & 20 & 30 \end{bmatrix}'$	$\begin{bmatrix} 15 & 15 & 15 \end{bmatrix}'$
$\boldsymbol{\sigma}_s^2$	$\begin{bmatrix} 4 & 4 & 4 \end{bmatrix}'$	$\begin{bmatrix} 10 & 10 & 10 \end{bmatrix}'$

Table 1. Actual and initial values of the model parameters used in our experiment.

case, where α_t's are scaled to sum up to one for all t. It can be shown that this scaling has no effect on the final parameter estimates.

4. NUMERICAL RESULTS

In this section, we present the numerical results of implementing our algorithm for identifications of HSMMs. In our experiment, the parameters of an HSMM signal with $N = 3$ distinct states were estimated using the algorithm presented in section 3. The total number of observations was $T = 10000$. The actual and initial values of the model parameters are given in table 1. Simulation showed that the log-likelihood of the total observations $\mathcal{Y}_T$ given the parameters estimate $\boldsymbol{\theta}$, (i.e., $\log(\mathbb{P}(\mathcal{Y}_T|\boldsymbol{\theta}))$, increased in each iteration. This verifies that our algorithm finds the maximum-likelihood estimate of the model parameters. The simulation results show that all the model parameters converge to their actual value after only a few iterations. Figure 1 illustrates the case for three of the model parameters (a^o_{12}, μ_3 and $\mu_{s,1}$), where our estimates of these parameters are plotted versus the iteration number.

5. CONCLUSION

We presented a novel signal generation model for hidden semi-Markov signals. Our model captures the state-duration densities in an HSMM using state duration dependent transition probabilities. We also presented a variant of the EM algorithm for the identification of our model parameters. Our identification algorithm finds the local maximum likelihood estimate of the model parameters. We also numerically showed that our algorithm can successfully estimate the actual values of the model parameters with significantly less computational effort.

6. REFERENCES

[1] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.

[2] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, November 1996.

[3] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 4, pp. 257–286, 1989.

[4] M. J. Russel and R. K. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," *ICASSP*, pp. 5–8, March 1985.

[5] B. Sin and J. H. Kim, "Nonstationary hidden markov model," *Signal Processing*, vol. 46, pp. 31–46, 1995.

[6] L. H. Jamieson and C. D. Mitchell, "Modelling duration in a hidden markov model with the exponential family," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 331 –334, 1993.

[7] S. E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.

[8] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.