Fast communication

Bernoulli versus Markov: Investigation of state transition regime in switching-state acoustic models

Jahanshah Kabudian\textsuperscript{a,}* , Mohammad Mehdi Homayounpour\textsuperscript{a}, Seyed Mohammad Ahadi\textsuperscript{b}

\textsuperscript{a} Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Hafez Avenue, Tehran 15914, Iran
\textsuperscript{b} Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Hafez Avenue, Tehran 15914, Iran

\textbf{ABSTRACT}

In this paper, a new acoustic model called time-inhomogeneous hidden Bernoulli model (TI-HBM) is introduced as an alternative to hidden Markov model (HMM) in continuous speech recognition. Contrary to HMM, the state transition process in TI-HBM is not a Markov process, rather it is an independent (generalized Bernoulli) process. This difference leads to elimination of dynamic programming at state-level in TI-HBM decoding process. Thus, the computational complexity of TI-HBM for probability evaluation and state estimation is $O(NL)$ (instead of $O(N^2L)$ in the HMM case, where $N$ and $L$ are number of states and sequence length respectively). As a new framework for phone duration modeling, TI-HBM is able to model acoustic-unit duration (e.g. phone duration) by using a built-in parameter named survival probability. Similar to the HMM case, three essential problems in TI-HBM have been solved. An EM-algorithm-based method has been proposed for training TI-HBM parameters. Experiments in phone recognition for Persian (Farsi) spoken language show that the TI-HBM has some advantages over HMM (e.g. more simplicity and increased speed in recognition phase), and also outperforms HMM in terms of phone recognition accuracy.

\section{Introduction}

Hidden Markov model (HMM) is the most popular and the most successful tool for analyzing and modeling stochastic sequences in speech processing [1]. The usual assumption in HMM is that the state transition process is a Markov process, and the generated state sequence is governed by Markov regime. It is experimentally approved that the state transition probabilities have less important roles compared to observation density functions (emission probabilities) in automatic speech recognition [2, Section 8.5.1]. There is no attempt on relaxing the Markov dependency in acoustic models like HMM.

In this paper, a new acoustic model named time-inhomogeneous hidden Bernoulli model (TI-HBM) has been proposed in which the Markov regime in state transition process is relaxed. There are many attempts on phone duration modeling [3–5]. The TI-HBM can model acoustic-unit duration (e.g. phone duration) by using a built-in parameter named survival probability, which is derived from joint state–time distribution parameters. Employing TI-HBM in speaker-independent phoneme recognition, leads to more simplicity of the acoustic model, increased speed in recognition, and also higher performance in terms of phoneme recognition accuracy.

The paper is organized as follows. In Section 2, the main elements of TI-HBM and some useful propositions and corollaries are presented. In Section 3, an algorithm for simulating TI-HBM is proposed. In Section 4, three essential problems (probability evaluation, state estimation, and training) of TI-HBM will be addressed. Also, an EM algorithm-based training method will be presented for
estimating TI-HBM parameters. In Section 5, we employ TI-HBM model for speaker-independent phoneme recognition on standard Persian continuous speech corpus FarsDat. Finally, in Section 6, we discuss about TI-HBM advantages over HMM.

2. Time-inhomogeneous hidden Bernoulli model

TI-HBM model is a new acoustic model which is able to simultaneously model both state transition and acoustic-unit (e.g. phone) duration by using a new parameter called joint state–time distribution $P_{ST}(i,t)$. The parameter $P(i,t)$ is probability of being in state $i$ at time $t$. Therefore, parameters of TI-HBM are:

1. Joint state–time distribution $P(i,t)$
2. Parameters of state-conditioned Gaussian mixtures \( \{w_{im}, \mu_{im}, \Sigma_{im}\} \), \( b_i(x_t) = \sum_{m=1}^{M} w_{im} \cdot \text{exp}^{-1} \left( x_t; \mu_{im}, \Sigma_{im} \right) \).

The parameters $P(i,t)$ play roles similar to $\pi_i$ and $a_{ij}$ in standard HMM. The following constraint must be satisfied:

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} P(i,t) = 1
\]

(1)

\[
P(i,t) = 0 \quad \text{for} \quad t > T_{\text{max}}
\]

(2)

where $T_{\text{max}}$ is the maximum length of observation sequence $X$. We derive some useful parameters from $P(i,t)$, which are needed for employing TI-HBM in the real world:

1. **Time distribution function** $P_t(t)$ or $P(t)$: The $P_t(t)$ is probability of being at time $t$ which is computed as follows:

\[
P(t) = \sum_{i=1}^{N} P(i,t)
\]

(3)

If we have $K$ observation sequences with length $L_k$ for $k$-th observation sequence, the time distribution function will be computed by relative frequency of observation vectors with time-index $t$ (frame number $t$). Therefore, the time distribution function $P_t(t)$ is empirically computed by the following formula:

\[
P(t) = \sum_{k=1}^{K} f_k(t) / \sum_{k=1}^{K} L_k
\]

where $f_k(t) = 1$ if cond is TRUE

\[
f_k(t) = 0 \quad \text{if cond is FALSE}
\]

(5)

2. **Survival probability** $P_{\text{next}}/t_{\text{run}}(t+1|t)$ or $P(t+1|t)$: Given that the process is at time $t$, the $P(t+1|t)$ is probability of process survival to time $t+1$. In other words, at time $t$, the process continues to time $t+1$ with probability $P(t+1|t)$, otherwise it is terminated at time $t$ with probability $1-P(t+1|t)$. The $P_{\text{next}}/t_{\text{run}}(t+1|t)$ is computed using Bayes formulation as follows:

\[
P_{\text{next}}/t_{\text{run}}(t+1|t) = P_{\text{next}}/t_{\text{run}}(t+1,t) / P_{t_{\text{run}}}(t)
\]

(6)

Since sequence length $L_k$ is always greater than zero, therefore:

\[
P_{\text{next}}/t_{\text{run}}(1|0) = 1
\]

(7)

The TI-HBM will be able to model acoustic-unit duration using survival probabilities. As we shall see in Corollary 1, survival probability is derived from $P(i,t)$ as follows:

\[
P_{\text{next}}/t_{\text{run}}(t+1|t) = P_T(t+1) / P_T(t) = \sum_{k=1}^{K} P_{\text{next}}/t_{\text{run}}(j,t+1) / \sum_{j=1}^{K} P_{\text{next}}/t_{\text{run}}(j,t)
\]

(8)

3. State selection probability given time $P_{\text{select}}(i|t)$ or $P(i|t)$: $P_{\text{select}}(i|t)$ is probability of selecting state $i$ at time $t$, and is computed using the following formula:

\[
P_{\text{select}}(i|t) = P_{\text{next}}(i,t) = P_{\text{next}}(i,t) / \sum_{j=1}^{K} P_{\text{next}}(j,t)
\]

(9)

It can be seen that the state selection and transition process is a generalized Bernoulli process with probabilities $P_{\text{select}}(i|t)$. Contrary to standard Bernoulli process which is a binary process (like coin tossing), the generalized Bernoulli process is a multi-valued one with $N$ outcomes (like die tossing in which $N = 6$) [6]. Since the probabilities $P(i|t)$ changes with respect to time, thus it is a time-inhomogeneous process.

Now, we present some useful propositions and corollaries relating to TI-HBM. The proofs for propositions are simple and straightforward.

**Proposition 1.** $P_{\text{next}}/t_{\text{run}}(t+1,t) = P_{\text{next}}(t+1,t)$.

**Proof.** If the next time-index, i.e. $T_{\text{next}}$ is $t+1$, then the current time-index, i.e. $T_{\text{curr}}$ is surely $t$. In other words

\[
P(T_{\text{curr}} = t; T_{\text{next}} = t+1) = 1
\]

(10)

\[
P(T_{\text{next}} = t+1; T_{\text{curr}} = t) = P(T_{\text{next}} = t+1) = P_{\text{next}}(t+1,t)
\]

(11)

\[
P(T_{\text{next}} = t+1; T_{\text{curr}} = t) = P(T_{\text{next}} = t+1)
\]

(12)

**Proposition 2.** The time-distribution $P_t(t)$ is a decreasing function with respect to time, i.e. $P_t(t+1) \leq P_t(t)$.

**Proof.** Firstly, we define a set of functions $f_k(t) = 1(t \leq L_k)$.

It is obvious that $f_k(t+1) \leq f_k(t)$ for all $t$

Summing the above equations over different $k$’s and then dividing by $\sum L_k$, we have

\[
\sum_{k=1}^{K} f_k(t+1) \leq \sum_{k=1}^{K} f_k(t)
\]

(13)

\[
\left( \sum_{k=1}^{K} f_k(t+1) / \sum_{k=1}^{K} L_k \right) \leq \left( \sum_{k=1}^{K} f_k(t) / \sum_{k=1}^{K} L_k \right)
\]

(15)
1. \( t = 1 \), \( P_{r_t \rightarrow r_m}(1 \mid 0) = 1 \).

2. The Bernoulli process continues with survival probability \( P_{r_m \rightarrow r_m}(t \mid t - 1) \) (otherwise, it is terminated with probability \( 1 - P_{r_m \rightarrow r_m}(t \mid t - 1) \)).

3. At time \( t \), state \( q_i \) is selected with probability \( P(q_i \mid t) \).

4. In state \( q_i \), a vector \( x_t \) is generated using a Gaussian mixture probability density function \( p(x_t \mid q_i, t) \) which is usually assumed to be time-independent, i.e., \( p(x_t \mid q_i, t) = p(x_t \mid q_i) \).

5. \( t = t + 1 \).

6. Go to step (2).

**Fig. 1.** Algorithm for simulating TI-HBM.

\[ \Rightarrow P_t(t + 1) \leq P_t(t) \quad (16) \]

**Corollary 1.** \( P_{r_m \rightarrow r_n}(t + 1) = P_t(t + 1)/P_t(t) \).

**Proof.** Using Proposition 1, we have

\[ P_{r_m \rightarrow r_n}(t + 1) = \frac{P_{r_m \rightarrow r_n}(t + 1, t)}{P_t(t)} = \frac{P_t(t + 1)}{P_t(t)} \]

Two events \( T = t \) and \( T_{t+1} = t \), and also events \( T = t+1 \) and \( T_{t+1} = t+1 \) are equivalent. \( \square \)

**Corollary 2.** Probability of generating a sequence of minimum length \( d \) is \( P(D \geq d) = (P(d) - P(d + 1))/P_t(1). \)

**Proof.** If \( D \) is a variable for the sequence length, then

\[ P(D \geq d) = P(1) = \sum_{t=2}^{d} P(t - 1) \]

\[ = \frac{P_t(2)}{P_t(1)} \frac{P_t(3)}{P_t(2)} \ldots \frac{P_t(d)}{P_t(d - 1)} \]

\[ \Rightarrow P(D \geq d) = (P(d) - P(d + 1))/P_t(1) \quad (18) \]

**Corollary 3.** Probability of generating a sequence of exact length \( d \) is \( P_0(d) = (P(d) - P(d + 1))/P_t(1). \)

**Proof.** If the sequence length is exactly \( d \), then the process will be terminated before time \( d+1 \) with probability \( 1 - P(d + 1)/d \):

\[ P_0(d) = P_D(d = d) = P_D(1) = \sum_{t=2}^{d} P_t(t - 1) \]

\[ = \frac{P_t(2)}{P_t(1)} \frac{P_t(3)}{P_t(2)} \ldots \frac{P_t(d)}{P_t(d - 1)} \]

\[ \Rightarrow P_0(d) = (P(d) - P(d + 1))/P_t(1) \quad (19) \]

The Corollary 2 is a way for converting duration-distribution function \( P_0(\cdot) \) to time distribution function \( P_t(\cdot) \):

\[ P_t(d) = P_t(1)P(D \geq d) \quad (20) \]

\[ P_t(1) = \frac{K}{\sum_{k=1}^{L_k} K_k} = \left( \frac{\sum_{k=1}^{L_k} K_k}{K} \right) = \frac{1}{E[D]} = \frac{1}{\sum_{d=1}^{L} P_D(d)} \quad (21) \]

According to Corollary 1 and Eq. (20), we can derive survival probabilities using duration-distribution function as follows:

\[ P_{r_m \rightarrow r_n}(t + 1|t) = \frac{P_t(t + 1)}{P_t(t)} = \frac{P(D \geq t + 1)}{P(D \geq t)} \quad (22) \]

Eq. (22) is compatible with some result achieved in [7, p. 1115], and verifies the propositions and corollaries in another way.

### 3. Simulation of TI-HBM

Simulating TI-HBM means that how an observation sequence \( X = \{x_1, x_2, \ldots, x_t, \ldots, x_k\} \) is generated by TI-HBM.

For this purpose, an algorithm in Fig. 1 is followed.

Time-independency assumption for state-conditioned Gaussian mixture probability density functions is prevalent in acoustic models like HMM, and is also made in TI-HBM as follows:

\[ p(x_t \mid q_i, t) \simeq p(x_t \mid q_i) = \sum_{m=1}^{M} W_{im} \cdot \nu(x_t, \mu_{im}, \sigma_{im}) \quad (23) \]

If \( \mathcal{T} = \{1, 2, 3, \ldots, L\} \) is the time-index sequence, and \( Q \) is the state sequence of generalized Bernoulli process, then the joint probability of surviving up to time \( L \), traversing state sequence \( Q \), and generating observation sequence \( X \) by TI-HBM will be

\[ P(\mathcal{T}, X, Q) = (1 - P(L + 1|L)) \prod_{l=1}^{L} P(t(l - 1)|q_i)p(x_t|q_i, t) \]

\[ \simeq (1 - P(L + 1|L)) \prod_{l=1}^{L} P(t(l - 1)|q_i)p(x_t|q_i) \]

\[ = \left( \frac{P_t(L - 1)}{P_t(1)} \right) \prod_{l=1}^{L} P(q_l|t)p(x_t|q_l) \quad (24) \]

The above equation can be written in another form

\[ P(\mathcal{T}, X, Q) = P(\mathcal{T})P(Q|\mathcal{T}) \]

\[ \times P(X|Q, \mathcal{T}) \simeq P(\mathcal{T})P(Q|\mathcal{T})P(X|Q) \quad (25) \]

\[ P(\mathcal{T}) = P_0(L) = \frac{P_0(L) - P_{r_f}(L + 1)/P_t(1)}{P_t(1)} \quad (26) \]
are optimally and globally determined by Eq. (4) and are fixed (constant values). Therefore, $P(\mathcal{F})$ will be treated as constant value in log-likelihood function of TI-HBM.

3.1. Efficient evaluation of probability $P(\mathcal{F}, X)$

Probability of generating an observation sequence $X$ of length $L$ is computed as follows:

$$
P(\mathcal{F}, X) = (1 - P(L+1|L)) \prod_{t=1}^{L} P(x_t|q_t)$$

$$
= \left( \frac{P_T(L) - P_T(L+1)}{P_T(1)} \right) \sum_{t=1}^{L} p(x_t, i|t)
$$

$$
= \left( \frac{P_T(L) - P_T(L+1)}{P_T(1)} \right) \sum_{i=1}^{N} P(i|t)p(x_t, i|t)
$$

$$
\simeq \left( \frac{P_T(L) - P_T(L+1)}{P_T(1)} \right) \sum_{t=1}^{L} \sum_{i=1}^{N} p(i|t)p(x_t, i|t)
$$

$$
= P_T(L) \sum_{t=1}^{L} \sum_{i=1}^{N} p(i|t)p(x_t, i|t)
$$

(29)

In standard HMM, this probability is computed using dynamic programming (DP)-based methods (forward and backward procedures). The order of computations for evaluating $P(X)$ in HMM is $O(N^2L)$ [1], while in TI-HBM, the order for evaluating $P(\mathcal{F}, X)$ is $O(NL)$. Since the state transition process in TI-HBM is not Markov-dependent, therefore the DP-type search is not needed for computing $P(\mathcal{F}, X)$.

3.2. Optimal state sequence estimation

Since the term $P(\mathcal{F})$ has no effect on $Q^*$, i.e.

$$
Q^* = \arg \max_{Q} P(\mathcal{F}, X, Q) = \arg \max_{Q} P(X, Q|\mathcal{F})
$$

(30)

therefore, the $P(X, Q|\mathcal{F})$ is used instead of $P(\mathcal{F}, X, Q)$. If $Q^*$ is the optimal state sequence for generating $X$ by TI-HBM, then

$$
P(X, Q^*|\mathcal{F}) = \max_{Q} P(X, Q|\mathcal{F})
$$

$$
= \max_{Q} \left\{ \prod_{t=1}^{L} p(q_t|t)p(x_t|q_t) \right\}
$$

$$
= \prod_{t=1}^{L} \max_{q_t} p(q_t|t)p(x_t|q_t)
$$

$$
= \prod_{t=1}^{L} p(q_t^*|t)p(x_t|q_t^*)
$$

(31)

$$
q_t^* = \arg \max_{q_t} \left\{ p(q_t|t)p(x_t|q_t) \right\}
$$

(32)

It can be seen that the DP-search (Viterbi algorithm with order $O(N^2L)$) is eliminated from the state estimation problem in TI-HBM, and the order of computations is $O(NL)$.

4. Training TI-HBM parameters

Suppose that we have a set $X$ of $K$ observation sequences for training TI-HBM parameters. If $X^{(k)}$ is k-th
If the TI-HBM parameter set is \( \Theta \), we want to find some parameter estimator \( \hat{\Theta} \) (with maximum likelihood criterion):

\[
\hat{\Theta} = \arg \max_{\Theta} \log P(X; \Theta)
\]

\[
P(X; \Theta) = \prod_{k=1}^{K} P(X^{(k)}; \Theta)
\]

(34)

4.3.1. Estimating \( P_{T}(t) \) parameters of TI-HBM

The estimate for \( P_{T}(t) \) parameters is the number of observation vectors with time-index \( t \) divided by the total number of observation vectors (as in Eq. (4)). This parameter estimator for \( P_{T}(t) \) depends only upon \( L_{k} \) parameters and is independent of \( X^{(k)} \)'s. Therefore, it yields the final estimate of \( P_{T}(t) \) parameters, it is kept fixed, and will be treated as constant value in next stages of training. In the EM algorithm, we only estimate \( P_{O}(t) \) parameters. After EM, parameters \( P_{S,T}(i,t) \) are simply derived by using \( P_{O}(t) = P_{T}(t)P_{O}(t) \).

In practice, the estimator in Eq. (4) must be smoothed. One way is to parameterize \( P_{O}(\cdot) \) with a suitable distribution (e.g. Gamma distribution), and convert \( P_{O}(\cdot) \) to \( P_{T}(\cdot) \) by Eqs. (20) and (21).

4.3.2. Training TI-HBM by EM algorithm

We have used EM algorithm [8] for training TI-HBM parameters. The details of mathematical manipulations can be found in [9]

\[
\hat{P}(i|t,x^{(k)}_{t};(\theta^{(n-1)})) = \frac{\sum_{l=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))}{\sum_{l=1}^{L_{k}} (t \leq L_{k})}
\]

(35)

\[
\hat{w}_{im} = \frac{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))}{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))}
\]

(36)

\[
\hat{\mu}_{im} = \frac{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))x^{(k)}_{l}}{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))}
\]

(37)

\[
\hat{\Sigma}_{im} = \frac{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))(x^{(k)}_{l} - \mu_{im})(x^{(k)}_{l} - \mu_{im})^T}{\sum_{l=1}^{L_{k}} \sum_{t=1}^{L_{k}} \hat{P}(i|l,x^{(k)}_{l};(\theta^{(n-1)}))}
\]

(38)

\[
P(i|t,x^{(k)}_{t};(\theta^{(n-1)})) = \frac{P(i|t,x^{(k)}_{t};(\theta^{(n-1)}))}{P(t,x^{(k)}_{t};(\theta^{(n-1)}))} = \frac{P(i|t)p(x^{(k)}_{t}|i)}{\sum_{j=1}^{N} P(j|t)p(x^{(k)}_{t}|j)}
\]

(39)

\[
p(x^{(k)}_{t}|i) = \sum_{m=1}^{M} w_{im}N(x^{(k)}_{t}; \mu_{im}, C_{im})
\]

(40)

\[
P(m|i,t,x^{(k)}_{t};(\theta^{(n-1)})) = \frac{P(i|t,x^{(k)}_{t};(\theta^{(n-1)}))}{P(m|i,t,x^{(k)}_{t};(\theta^{(n-1)}))} = \frac{w_{im}N(x^{(k)}_{t}; \mu_{im}, C_{im})}{\sum_{m=1}^{M} w_{im}N(x^{(k)}_{t}; \mu_{im}, C_{im})}
\]

(41)

\[
P(m|i,t,x^{(k)}_{t};(\theta^{(n-1)})) = \frac{w_{im}N(x^{(k)}_{t}; \mu_{im}, C_{im})}{\sum_{m=1}^{M} w_{im}N(x^{(k)}_{t}; \mu_{im}, C_{im})}
\]

(42)

The estimated values \( \hat{\Theta} \) will be stored in \( \Theta^{(n)} \) for next iteration.

After EM re-estimation of \( \hat{P}(i|t) \), the final value of \( \hat{P}(i,t) \) is computed as follows:

\[
\hat{P}_{S,T}(i,t) = \hat{P}^{\text{sm}}_{T}(i,t) \cdot \hat{P}_{S,T}(i,t)
\]

(43)

where \( \hat{P}^{\text{sm}}_{T}(i,t) \) is smoothed \( \hat{P}_{T}(i,t) \) derived from parameterized \( P_{O}(\cdot) \) (as described in section 4.3.1).

5. Experiments

We have employed TI-HBM in speaker-independent phone recognition for Persian (Farsi) spoken language. For training HMM and TI-HBM phone models, the standard Farsi phonetically-balanced continuous speech database FarsDat [10] was used (available via ELDA website [11]). The FarsDat contains utterances of 304 speakers from 10 dialect regions inside Iran. Each speaker has uttered 20 sentences (of which 2 sentences are common among speakers). The utterances of first 250 speakers was used for training phone models (5000 sentences), and utterances of remaining 54 speakers was used for test (1080 sentences). Thirty-two phone models were trained. Feature vectors are 13 cepstral coefficients \( (c_{0} - c_{12}) \) derived from perceptual linear prediction analysis, plus first-, second-, and third-order derivatives (52-dimensional). The HMM and TI-HBM models have 3 states, and 2, 4, 8, 16, 24 and 32 diagonal-covariance Gaussian PDFs per state. For improving the results, a phone-bigram language model was used, and trained using phone labels of the training set. The final value of \( L_{\text{max}} \) was \( 2L_{\text{train}} \). The \( P_{O}(\cdot) \) was parameterized (smoothed) with a Gamma distribution and truncated outside the interval \( L_{\text{min}} \leq t \leq L_{\text{max}} \). Then converted to \( P_{T}(\cdot) \) using Eqs. (20) and (21) in interval \( 1 \leq t \leq L_{\text{max}} \). The survival probabilities were then computed using smoothed \( P_{O}(\cdot) \). Both HMM and TI-HBM models were trained by EM algorithm. As it is seen from re-estimation formulas, TI-HBM training is completely independent of HMM training. But for making the comparison fair, we have to start the HMM and TI-HBM training from the same initial points. The initial values of Gaussian mixture parameters for HMM and TI-HBM were the same. Since state transition-related parameters in HMM and TI-HBM are intrinsically different, we have to start the training from equivalent initial points. One way is to initialize HMM parameters with \( \Theta^{(0)}_{\text{HMM}} \), and then convert \( \Theta^{(0)}_{\text{HMM}} \) to equivalent values in TI-HBM, i.e. \( \Theta^{(0)}_{\text{HBM}} \). The algorithm is as follows: the HMM parameters were initialized with \( \Theta^{(0)}_{\text{HMM}} \). By using \( \Theta^{(0)}_{\text{HMM}} \), the optimal state sequence for all observation sequences were determined, and the ratio of number of observation vectors with time-index \( t \) which assigned to state \( i \), to the number of observation vectors with time-index \( t \), was used as initial value of \( P(i|t) \) in TI-HBM. The initial values
of Gaussian mixture parameters for HMM and TI-HBM were the same. Therefore, both models have been trained using EM algorithm with starting from equivalent initial points. In decoding process, survival probabilities \( P(t|t-1) \) are used instead of \( P_d(d) \), because phone durations \( (d's) \) are not known before the end of the search. In practice, \( [P(t|t-1)]^{\text{DP}} \) and \( [1-P(t|t-1)]^{\text{DP}} \) was used. This is equivalent to putting a weight on duration-distribution function, i.e. using \( P_d(L)^{\text{DP}} \) instead of \( P_d(L) \). The DSF parameter was optimally set to 3. After estimating \( P(t|t) \) by EM algorithm, these parameters were extended to interval \( L_{\text{max}} \leq t \leq L_{\text{max}} \) like as follows:

\[
P(t|t) = P(t|t_{\text{train}}) \quad \text{for all } i \text{ and } L_{\text{max}} < t \leq L_{\text{max}}
\]

In recognition phase, DP-type search is eliminated at the state-level, but the DP-search is needed for finding the best phoneme segmentation (phoneme boundaries) up to time \( t \) by performing a DP-search on \( P^{\text{DP}}(\cdot|t) \) and \( P^{\text{DP}}(\cdot'|t'-1) \) probabilities.

The phone recognition results are shown in Table 1. It can be seen that the TI-HBM improves the phone recognition accuracy compared to standard HMM.

In another experiment, we compared recognition time for both HMM and TI-HBM models. Table 2 shows the elapsed time for decoding 200 s of speech signal (on an Intel Pentium IV, 3.2 GHz processor). We can see that the TI-HBM is always faster than HMM, and speed-up factor is greater for low number of Gaussians per state. This is because of the fact that the main computational complexity of HMM and TI-HBM is due to the computation of emission probabilities (Gaussian mixtures). The TI-HBM will be quite faster than HMM for applications with low number of Gaussians per state, or feature vectors with low number of dimensions, or for those discrete HMM cases in which the index of observation vector is known a priori without any computations (like amino acids in bioinformatics applications where discrete HMM is widely used). Also, the TI-HBM was always faster than HMM in training phase in our experiments (not reported here).

Another issue is number of free parameters of the model. The number of HMM parameters \( \{\pi_i, a_{ij}, W_m, \mu_{mk}, C_{imk}\} \) with diagonal-covariance Gaussian mixtures is \( N+2N+NM+2NM \), while for TI-HBM parameters \( \{P(t), W_m, \mu_{mk}, C_{imk}\} \), it is \( N_{\text{avg}}+NM+2NM \) (where \( L_{\text{max}} \) is average \( L_{\text{max}} \) for Persian phoneme set in FarsDat database). Assuming that \( 2NM \gg L_{\text{max}} \) (for example \( M = 32, Dim = 52, L_{\text{max}} = 35 \)), there is no considerable increase in the number of TI-HBM parameters. Furthermore, \( P(t) \) for each state \( i \) can be parameterized by a Gamma distribution to decrease term \( N_{\text{avg}} \) to \( 2N \) (each Gamma distribution has two parameters).

### 6. Conclusion

In this paper, a new acoustic model named TI-HBM was introduced as an alternative to HMM for speech recognition. In TI-HBM, state transition process is a generalized Bernoulli process instead of a Markov one. In terms of phoneme recognition accuracy, the TI-HBM outperforms the HMM. Also, TI-HBM has some simplicities and advantages over HMM, including:

1. TI-HBM is a new theoretical framework for processing time series data, especially for automatic speech recognition, by defining a set of new parameters called joint state–time distribution.
2. DP-search is eliminated at the state-level in TI-HBM which makes it simpler compared to HMM.
3. TI-HBM is faster than HMM in recognition and training phase.
4. TI-HBM is capable of modeling acoustic-unit duration (e.g. phone duration) by employing a parameter named survival probability.
5. Computation of probability in TI-HBM is performed in a non-recursive manner. Therefore, differentiation of
TI-HBM likelihood function with respect to its parameters is simpler and faster compared to that of HMM, and does not need calculation of recursive forward and backward variables \( \alpha_t(i) \) and \( \beta_t(i) \).

According to the obtained results on comparison between HMM and TI-HBM, it is approved that the state transition structure in acoustic models like HMM or TI-HBM is less important compared to the observation density structure. Therefore, the TI-HBM can be an alternative to the HMM with easier use for applications like speech recognition, in which the state and the time-index (frame number) have strong relationship. Using uniform segmentation (equally segmenting speech signal which corresponds to an acoustic-unit, and assigning each segment to a state in HMM) in speech recognition for initializing HMM parameters is an evidence for this relationship [1]. Furthermore, TI-HBM can be used for modeling other speech acoustic-units like word, syllable, etc. Employing TI-HBM in other applications like bioinformatics, time series and pattern recognition can further reveal other advantages of this model.

Acknowledgement

This research was supported by the Iran Telecommunication Research Center (ITRC) under Contract T-500-9269.

References