TIME-INHOMOGENEOUS HIDDEN BERNOULLI MODEL: AN ALTERNATIVE TO HIDDEN MARKOV MODEL FOR AUTOMATIC SPEECH RECOGNITION

Jahanshah Kabudian[1], M. Mehdi Homayounpour[1], S. Mohammad Ahadi[2]

[1] Department of Computer Engineering,
[2] Department of Electrical Engineering,
AmirKabir University of Technology (Tehran Polytechnic), Tehran, IRAN.
{kabudian, homayoun, sma} *at* aut.ac.ir

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 automatic 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 $\mathcal{O}(NL)$ (instead of $\mathcal{O}(N^2L)$ in the HMM case). 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.

Index Terms— Time-Inhomogeneous Hidden Bernoulli Model, Hidden Markov Model, Speech Recognition, Acoustic Modeling, Phone Recognition, Phone Duration Modeling, Persian (Farsi) Spoken Language.

1. 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 obeys a Markov regime. It is experimentally approved that the state transition probabilities have less important roles compared to observation density functions. There is no attempt on relaxing the Markov dependency in acoustic models like HMM. In this paper, a new acoustic model named TI-HBM has been proposed in which the Markov regime in state transition process is relaxed. There are many attempts on phone duration modeling [2,3,4]. The TI-HBM models 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. In the next sections, we introduce TI-HBM and its basic definitions.

2. TI-HBM

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 $P_{S,T}(i,t)$. 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 Gaussian mixtures, *i.e.* w_{im}, μ_{im} and C_{im}.

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

$$\sum_{i=1}^{N}\sum_{t=1}^{L_{\max}} P(i,t) = 1 \quad (2.1)$$

$$P(i,t) = 0 \quad \text{for} \quad t > L_{\max} \quad (2.2)$$

where $L_{\max}$ 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 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:

$$P(t) = \sum_{i=1}^{N} P(i,t) \quad (2.3)$$

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:

$$\hat{P}(t) = \frac{\sum_{k=1}^{K} \mathbb{1}(t \le L_k)}{\sum_{k=1}^{K} L_k} \quad (2.4)$$

$$\mathbb{1}(cond) = \begin{cases} 1 & \text{if } cond \text{ is TRUE} \\ 0 & \text{if } cond \text{ is FALSE} \end{cases} \quad (2.5)$$

2. Survival probability $P_{T_{next}|T_{curr}}(t+1 \mid t)$ or $P(t+1 \mid t)$:

Given that the process is at time t, the $P(t+1 \mid t)$ is probability of process survival to time $t+1$. In other words, at time t, the process continues to time $t+1$ with probability $P(t+1 \mid t)$, otherwise it is terminated at time t with probability $1 - P(t+1 \mid t)$. The $P_{T_{next}|T_{curr}}(t+1 \mid t)$ is computed using Bayes formulation as follows:

$$P_{T_{next}|T_{curr}}(t+1 \mid t) = \frac{P_{T_{next},T_{curr}}(t+1,t)}{P_{T_{curr}}(t)} \quad (2.6)$$

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

$$P_{T_{next}|T_{curr}}(1 \mid 0) = 1 \quad (2.7)$$

The TI-HBM will be able to model acoustic-unit duration using survival probabilities.

3. State selection probability given time $P_{S|T}(i \mid t)$ or $P(i \mid t)$:

$P_{S|T}(i \mid t)$ is probability of selecting state i at time t, and is computed using the following formula:

$$P_{S|T}(i \mid t) = \frac{P_{S,T}(i,t)}{P_T(t)} = \frac{P_{S,T}(i,t)}{\sum_{j=1}^{N} P_{S,T}(j,t)} \quad (2.8)$$

It can be seen that the state selection and transition process is a generalized Bernoulli process with probabilities $P_{S|T}(i \mid t)$. 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 dice tossing in which $N = 6$) [5]. Since the probabilities $P(i \mid t)$ 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 2.1. $P_{T_{next},T_{curr}}(t+1,t) = P_{T_{next}}(t+1)$.

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

$$P(T_{curr} = t \mid T_{next} = t+1) = 1 \quad (2.9)$$

$$\frac{P(T_{curr} = t, T_{next} = t+1)}{P(T_{next} = t+1)} = 1 \quad (2.10)$$

$$P(T_{next} = t+1, T_{curr} = t) = P(T_{next} = t+1) = P_{T_{next}}(t+1) \quad (2.11)$$

□

Proposition 2.2. The Time-Distribution $P_T(t)$ is a decreasing function with respect to time, *i.e.* $P_T(t+1) \le P_T(t)$.

Proof. Firstly, we define a set of functions $f_k(t) = 1(t \le L_k)$. It is obvious that:

$$f_k(t+1) \le f_k(t) \quad \text{for all } t \quad (2.12)$$

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

$$\sum_{k=1}^{K} f_k(t+1) \le \sum_{k=1}^{K} f_k(t) \quad (2.13)$$

$$\left(\sum_{k=1}^{K} f_k(t+1) / \sum_{k=1}^{K} L_k\right) \le \left(\sum_{k=1}^{K} f_k(t) / \sum_{k=1}^{K} L_k\right) \quad (2.14)$$

$$\Rightarrow \quad P_T(t+1) \le P_T(t) \quad (2.15)$$

□

Corollary 2.1. $P_{T_{next}|T_{curr}}(t+1 \mid t) = \dfrac{P_T(t+1)}{P_T(t)}$.

Proof. Using proposition (2.1), we have:

$$P_{T_{next}|T_{curr}}(t+1 \mid t) = \frac{P_{T_{next},T_{curr}}(t+1,t)}{P_{T_{curr}}(t)} = \frac{P_{T_{next}}(t+1)}{P_{T_{curr}}(t)} = \frac{P_T(t+1)}{P_T(t)} \quad (2.16)$$

Two events $T = t$ and $T_{curr} = t$, and also events $T = t+1$ and $T_{next} = t+1$ are equivalent.

□

Corollary 2.2. Probability of generating a sequence of minimum length d is $P(D \ge d) = \dfrac{P_T(d)}{P_T(1)}$.

Proof. If D is a variable for the sequence length, then:

$$\begin{aligned} P(D \ge d) &= P(1 \mid 0).\prod_{t=2}^{d} P(t \mid t-1) \\ &= \frac{P_T(2)}{P_T(1)}.\frac{P_T(3)}{P_T(2)}.\cdots.\frac{P_T(d)}{P_T(d-1)} = \frac{P_T(d)}{P_T(1)} \end{aligned} \quad (2.17)$$

□

Corollary 2.3. Probability of generating a sequence of exact length d is $P_D(d) = \dfrac{P_T(d) - P_T(d+1)}{P_T(1)}$.

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

$$\begin{aligned} P_D(d) &= P(D = d) = P(1 \mid 0).\left\{\prod_{t=2}^{d} P(t \mid t-1)\right\}.(1 - P(d+1 \mid d)) \\ &= \frac{P_T(2)}{P_T(1)}.\frac{P_T(3)}{P_T(2)}.\cdots.\frac{P_T(d)}{P_T(d-1)}.\left(1 - \frac{P_T(d+1)}{P_T(d)}\right) = \frac{P_T(d) - P_T(d+1)}{P_T(1)} \end{aligned} \quad (2.18)$$

□

The Corollary 2.2 is a way for converting duration distribution function $P_D(.)$ to time distribution function $P_T(.)$:

$$P_T(d) = P_T(1).P(D \ge d) \quad (2.19)$$

$$P_T(1) = \frac{K}{\sum_{k=1}^{K} L_k} = \left(\frac{\sum_{k=1}^{K} L_k}{K}\right)^{-1} = \frac{1}{E\{D\}} \quad (2.20)$$

According to Corollary 2.1 and Eq. (2.19), we can derive survival probabilities using duration distribution function as follows:

$$P_{T_{next}|T_{curr}}(t+1 \mid t) = \frac{P_T(t+1)}{P_T(t)} = \frac{P(D \ge t+1)}{P(D \ge t)} \quad (2.21)$$

Equation (2.21) is compatible with some result achieved in [6, page 1115], and verifies the propositions and corollaries in another way.

3. SIMULATION OF TI-HBM

Simulating TI-HBM means that how an observation sequence $X = \{x_1, x_2, ..., x_t, ..., x_L\}$ is generated by TI-HBM. For this purpose, an algorithm in Fig. (1) is followed.

1. $t = 1$, $P_{T_{next}|T_{curr}}(1 \mid 0) = 1$.
2. The Bernoulli process continues with survival probability $P_{T_{next}|T_{curr}}(t \mid t-1)$ (otherwise, it is terminated with probability $1 - P_{T_{next}|T_{curr}}(t \mid t-1)$).
3. At time t, state q_t is selected with probability $P(q_t \mid t)$.
4. In state q_t, a vector x_t is generated using a Gaussian mixture probability density function $p(x_t \mid q_t, t)$ which is usually assumed to be time-independent, *i.e.* $p(x_t \mid q_t, t) \simeq p(x_t \mid q_t)$.
5. $t = t+1$.
6. Go to step (2).

Figure 1. Algorithm for simulating TI-HBM

If $\mathcal{T} = \{1,2,...,t,...,L\}$ 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:

$$\begin{aligned} P(\mathcal{T}, X, Q) &= (1 - P(L+1 \mid L)).\prod_{t=1}^{L} P(t \mid t-1).P(q_t \mid t).p(x_t \mid q_t, t) \\ &\simeq (1 - P(L+1 \mid L)).\prod_{t=1}^{L} P(t \mid t-1).P(q_t \mid t).p(x_t \mid q_t) \\ &= \left(\frac{P_T(L) - P_T(L+1)}{P_T(1)}\right).\prod_{t=1}^{L} P(q_t \mid t).p(x_t \mid q_t) \\ &= P_D(L).\prod_{t=1}^{L} P(q_t \mid t).p(x_t \mid q_t) \end{aligned} \quad (3.1)$$

The above equation can be written in another form:

$$P(\mathcal{T}, X, Q) = P(\mathcal{T}).P(Q \mid \mathcal{T}).P(X \mid Q, \mathcal{T}) \simeq P(\mathcal{T}).P(Q \mid \mathcal{T}).P(X \mid Q) \quad (3.2)$$

$$P(\mathcal{T}) = P_D(L) = \frac{P_T(L) - P_T(L+1)}{P_T(1)} \quad (3.3)$$

$$P(Q \mid \mathcal{T}) = \prod_{t=1}^{L} P(q_t \mid t) \quad (3.4)$$

$$P(X \mid Q, \mathcal{T}) \simeq P(X \mid Q) = \prod_{t=1}^{L} p(x_t \mid q_t) \quad (3.5)$$

where $P(\mathcal{T})$ is probability of generating a sequence with exact length L. It can be seen that the $P(\mathcal{T})$ is a function of $P_T(t)$ parameters only. On the other hands, parameters $P_T(t)$ are optimally and globally determined by Eq. (2.4) and are fixed (constant values). Therefore, $P(\mathcal{T})$ will be treated as constant value in log-likelihood function of TI-HBM.

4. THREE ESSENTIAL PROBLEMS IN TI-HBM

For employing TI-HBM in real-world applications, three essential problems (similar to those in the HMM case) must be solved: Efficient Probability Evaluation, Optimal State Sequence Estimation (Decoding), and Parameter Estimation (Training).

4.1. Efficient Evaluation of Probability $P(\mathcal{T},X)$

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

$$\begin{aligned}P(\mathcal{T},X)&=(1-P(L+1\mid L)).\prod_{t=1}^{L}P(t\mid t-1).p(x_t\mid t)\\&=\left(\frac{P_T(L)-P_T(L+1)}{P_T(1)}\right).\prod_{t=1}^{L}p(x_t\mid t)\\&=\left(\frac{P_T(L)-P_T(L+1)}{P_T(1)}\right).\prod_{t=1}^{L}\sum_{i=1}^{N}p(x_t,i\mid t)\\&=\left(\frac{P_T(L)-P_T(L+1)}{P_T(1)}\right).\prod_{t=1}^{L}\sum_{i=1}^{N}P(i\mid t).p(x_t\mid i,t)\\&\simeq\left(\frac{P_T(L)-P_T(L+1)}{P_T(1)}\right).\prod_{t=1}^{L}\sum_{i=1}^{N}P(i\mid t).p(x_t\mid i)\\&=P_D(L).\prod_{t=1}^{L}\sum_{i=1}^{N}P(i\mid t).p(x_t\mid i)\end{aligned}\quad(4.1)$$

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 $\mathcal{O}(N^2L)$ [1], while in TI-HBM, the order for evaluating $P(\mathcal{T},X)$ is $\mathcal{O}(NL)$. Since the state transition process in TI-HBM is not Markov-dependent, therefore the dynamic programming-type search is not needed for computing $P(\mathcal{T},X)$.

4.2. Optimal State Sequence Estimation

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

$$Q^*=\arg\max_{Q}P(\mathcal{T},X,Q)=\arg\max_{Q}P(X,Q\mid\mathcal{T})\quad(4.2)$$

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

$$\begin{aligned}P(X,Q^*\mid\mathcal{T})&=\max_{Q}P(X,Q\mid\mathcal{T})=\max_{q_1,q_2,\ldots,q_L}\left\{\prod_{t=1}^{L}P(q_t\mid t).p(x_t\mid q_t)\right\}\\&=\prod_{t=1}^{L}\max_{q_t}P(q_t\mid t).p(x_t\mid q_t)=\prod_{t=1}^{L}P(q_t^*\mid t).p(x_t\mid q_t^*)\end{aligned}\quad(4.3)$$

$$q_t^*=\arg\max_{q_t}\{P(q_t\mid t).p(x_t\mid q_t)\}\quad(4.4)$$

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

4.3 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 observation sequence of length L_k, then:

$$X^{(k)}=\left(x_1^{(k)},x_2^{(k)},\ldots,x_t^{(k)},\ldots,x_{L_k}^{(k)}\right)\quad(4.5)$$

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. (2.4)). 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 $P_{S|T}(i\mid t)$ parameters.

In practice, the estimator in Eq. (2.4) must be smoothed. One way is to parameterize $P_D(.)$ with a suitable distribution (*e.g.* Gamma distribution), and convert $P_D(.)$ to $P_T(.)$ by Eq. (2.19)-(2.20).

4.3.2. Training TI-HBM by EM Algorithm

We have used EM algorithm [7] for training TI-HBM parameters. The details of mathematical manipulations can be found in [8]:

$$\hat{P}(i\mid t)=\frac{\sum_{k=1}^{K}P\left(i\mid t,x_t^{(k)};\theta^{(n-1)}\right).1(t\le L_k)}{\sum_{k=1}^{K}1(t\le L_k)}\quad(4.6)$$

$$\hat{w}_{im}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(m,i\mid t,x_t^{(k)};\theta^{(n-1)}\right)}{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(i\mid t,x_t^{(k)};\theta^{(n-1)}\right)}\quad(4.7)$$

$$\hat{\mu}_{im}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(m,i\mid t,x_t^{(k)};\theta^{(n-1)}\right).x_t^{(k)}}{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(m,i\mid t,x_t^{(k)};\theta^{(n-1)}\right)}\quad(4.8)$$

$$\hat{C}_{im}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(m,i\mid t,x_t^{(k)};\theta^{(n-1)}\right).\left(x_t^{(k)}-\mu_{im}\right)\left(x_t^{(k)}-\mu_{im}\right)^T}{\sum_{k=1}^{K}\sum_{t=1}^{L_k}P\left(m,i\mid t,x_t^{(k)};\theta^{(n-1)}\right)}\quad(4.9)$$

$$\begin{aligned}P\left(i\mid t,x_t^{(k)};\theta^{(n-1)}\right)&=\frac{P\left(i,t,x_t^{(k)}\right)}{P\left(t,x_t^{(k)}\right)}=\frac{P\left(i,t,x_t^{(k)}\right)}{\sum_{j=1}^{N}P\left(j,t,x_t^{(k)}\right)}\\&=\frac{P(i\mid t).p(x_t^{(k)}\mid i)}{\sum_{j=1}^{N}P(j\mid t).p(x_t^{(k)}\mid j)}\end{aligned}\quad(4.10)$$

$$p(x_t^{(k)}\mid i)=\sum_{m=1}^{M}w_{im}\mathcal{N}(x_t^{(k)},\mu_{im},C_{im})\quad(4.11)$$

$$P(m,i\mid t,x_t^{(k)};\theta^{(n-1)})=P(i\mid t,x_t^{(k)};\theta^{(n-1)}).P(m\mid i,t,x_t^{(k)};\theta^{(n-1)})\quad(4.12)$$

$$P(m\mid i,t,x_t^{(k)};\theta^{(n-1)})=\frac{w_{im}\mathcal{N}(x_t^{(k)},\mu_{im},C_{im})}{\sum_{m'=1}^{M}w_{im'}\mathcal{N}(x_t^{(k)},\mu_{im'},C_{im'})}\quad(4.13)$$

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

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* [8] was used (available via ELDA web site). The *FarsDat* contains utterances of 304 speakers from 10 dialect regions inside Iran. Each speaker has uttered 20 sentences. 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). 32 phone models were trained. Feature vectors are 13 cepstral coefficients (c_0 – c_{12}) derived from Perceptual Linear Prediction analysis, plus 1st, 2nd, and 3rd-order derivatives. 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 $L_{\max}$ was $2L_{\max}^{train}$. The $P_D(.)$ was parameterized (smoothed) with a Gamma distribution and truncated outside the interval $L_{\min}\le t\le L_{\max}$ ($L_{\min}=3$), and then converted to $P_T(.)$ using Eq. (2.19)-(2.20) in interval $1\le t\le L_{\max}+1$. The survival probabilities were then computed using smoothed $P_T(.)$. Both HMM and TI-HBM models were trained by EM algorithm. The HMM parameters were initialized with θ_0^{HMM}. By using θ_0^{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 $P(i\mid t)$ 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 $P(t \mid t-1)$ are used instead of $P_D(d)$, because phone durations (d's) are not known before the end of the search. In practice, $[P(t \mid t-1)]^{DSF}$ and $[1-P(t \mid t-1)]^{DSF}$ was used. This is equivalent to putting a weight on duration-distribution function, *i.e.* using $[P_D(L)]^{DSF}$ instead of $P_D(L)$. The DSF parameter was optimally set to 3. After Estimating $P(i \mid t)$ by EM algorithm, these parameters were extended to interval $L_{\max}^{train} < t \le L_{\max}$ like as follows:

$$P(i \mid t) = P(i \mid L_{\max}^{train}) \quad \text{for all } i \text{ and } L_{\max}^{train} < t \le L_{\max} \qquad (5.1)$$

In recognition phase, we have used an array $t_0(ph)$ for keeping the entrance time into phone ph along the partial best path. Therefore, relative time-index t' instead of t was used in parameters $P(t)$ and $P(i \mid t)$:

$$\begin{aligned} P_T^{(ph)}(t') &= P_T^{(ph)}(t - t_0(ph) + 1) \\ P_{S|T}^{(ph)}(i \mid t') &= P_{S|T}^{(ph)}(i \mid t - t_0(ph) + 1) \end{aligned} \qquad (5.2)$$

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.

Table 1. Phone recognition accuracy (%) for the test set

No. of Gaussians per state	HMM	TI-HBM
2	68.31	68.38
4	71.63	71.74
8	74.17	74.19
16	75.68	75.94
24	76.21	76.70
32	76.84	77.22

Table 2. Elapsed time (*sec*) for decoding 200 seconds of speech

No. of Gaussians per state	HMM	TI-HBM	Speed-up
2	6.37	3.53	80.45%
4	8.61	5.81	48.19%
8	12.57	10.61	18.47%
16	20.34	20.25	0.44%

In another experiment, we compared recognition time for both HMM and TI-HBM models. Table (2) shows the elapsed time for *decoding* 200 seconds 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 probability (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).

6. CONCLUSION

In this paper, a new acoustic model named Time-Inhomogeneous Hidden Bernoulli Model was introduced as an alternative to Hidden Markov Model 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 speech recognition, by defining a set of new parameters called Joint State-Time Distribution.
2. Dynamic Programming search is eliminated at 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 (*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.

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 Iran Telecommunication Research Center (ITRC) under contract T-500-9269.

REFERENCES

[1] L.R. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[2] G. Linares, B. Lecouteux, D. Matrouf, P. Nocera, "Phone Duration Models for Fast Broadcast News Transcription," *Proc. IEEE ICASSP*, PA, USA, 2005.

[3] D. Povey, "Phone Duration Modeling for LVCSR," *Proc. IEEE ICASSP*, Montreal, Canada, 2004.

[4] J. Pylkkönen, M. Kurimo, "Using Phone Durations in Finnish Large Vocabulary Continuous Speech Recognition," *Proc. NORSIG*, Espoo, Finland, June 2004.

[5] K.S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, 2nd Edition, John Wiley & Sons, 2001.

[6] P.M. Djurić, J.-H. Chun, "An MCMC Sampling Approach to Estimation of Nonstationary Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 50, no. 5, pp. 1113-1123, 2002.

[7] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, B*, vol. 39, no. 1, pp. 1-38, 1977.

[8] J. Kabudian, M.M. Homayounpour, S.M. Ahadi, "Time-Inhomogeneous Hidden Bernoulli Model," *Computer Speech and Language*, Under Review.

[9] M. Bijankhan, J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, M. Tebyani, "FarsDat – The Speech Database of Farsi Spoken Language," *Proc. 5th Australian Int. Conf. Speech Science and Technology (SST)*, pp. 826-831, Perth, Australia, 1994.