Bayesian Duration Modeling and Learning for Speech Recognition

Jen-Tzung Chien and Chih-Hsien Huang

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan 70101, ROC
jtchien@mail.ncku.edu.tw & acheron@chien.csie.ncku.edu.tw

Abstract

We present the Bayesian duration modeling and learning for speech recognition under nonstationary speaking rates and noise conditions. In this study, the Gaussian, Poisson and gamma distributions are investigated to characterize duration models. The maximum *a posteriori* (MAP) estimate of gamma duration model is developed. To exploit the sequential learning, we adopt the Poisson duration model incorporated with gamma prior density, which belongs to the conjugate prior family. When the adaptation data are sequentially observed, the gamma posterior density is produced for twofold advantages. One is to determine the optimal quasi-Bayes (QB) duration parameter, which can be merged in HMM's for speech recognition. The other one is to build the updating mechanism of gamma prior statistics for sequential learning. EM algorithm is applied to fulfill parameter estimation. In the experiments, the proposed Bayesian approaches significantly improve the speech recognition performance of Mandarin broadcast news. The batch and sequential learning are investigated for MAP and QB duration models, respectively.

1. Introduction

There is no doubt that the performance of automatic speech recognition is significantly degraded in adverse conditions. Speaking rate is one of the mismatch sources between training and recognition data. One successful approach to recognize fast speech aims to estimate the adaptive state duration parameters in hidden Markov models (HMM's). In [8], the temporal constraints of state duration were imposed in Viterbi decoding. Considerable improvement was attained for noisy speech recognition. Except recognizing fast and noisy speech, the duration parameters are essential for HMM modeling of normal speech. In standard HMM, the state duration probability decreases exponentially with time. When the speech signal stays in state i with self-transition probability a_{ii} for τ frames, the implicit duration probability density is formed by a geometric distribution $d_i(\tau) = a_{ii}^{\tau-1}(1-a_{ii})$. However, this exponential state duration density is inappropriate for most signals. The pioneer work of Ferguson [4] discovered the explicit duration modeling for HMM's through calculating the nonparametric $d_i(\tau)$ for all states i and duration lengths τ. The physical speech duration was characterized without state self-transition. Because of inducing too many parameters, this method suffered from insufficient training data and increased computational load. Also, the alternative of using parametric density function was widely discussed. This approach had the significant advantage for reliable parameter estimation. In addition to Gaussian density, Russell and Moore [7] applied Poisson distribution while Levinson [6] applied gamma distribution for parametric duration modeling. In [1], the state as well as the word parametric duration models was merged in a modified Viterbi algorithm. In general, the parametric duration models using different distribution functions have been demonstrated effective for speech recognition. In this paper, we present the Bayesian duration learning framework for robust speech recognition. The MAP and QB estimates are developed and realized via the expectation-maximization (EM) algorithm.

2. Parametric Duration Modeling

In this study, we externally append the state duration densities $D = \{d_i(\cdot)\}$ to standard HMM parameters consisted of initial state probabilities $\pi = \{\pi_i\}$, transition probabilities $A = \{a_{ij}, i \neq j\}$ and observation densities $B = \{b_i(\cdot)\}$. Given a set of training data $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, we are estimating the joint HMM parameters $\lambda = (\pi, A, B, D)$ based on the maximum likelihood (ML) criterion. With the parametric models, we are able to perform Bayesian duration model learning.

2.1. ML Parameter Estimation

ML estimation aims to find the optimal parameters λ_{ML} via maximizing the observation likelihood function

$$\lambda_{ML} = \arg\max_{\lambda} p(X|\lambda). \tag{1}$$

Because HMM is inherent with incomplete data problem, it is necessary to apply EM algorithm [2] to realize ML parameter estimation. Accordingly, we introduce the HMM state sequence $\mathbf{q} = (q_1, q_2, \ldots, q_T)$, $1 \le q_t \le N$, to generate the complete data $(X, \mathbf{q})$. The E-step is to calculate the expectation of complete data given new estimate $\hat{\lambda}$ after having current estimate λ

$$\begin{aligned} Q(\hat{\lambda}|\lambda) = E[\log p(X, \mathbf{q}|\hat{\lambda})|X, \lambda] &= \sum_{i=1}^{N} P(q_1 = i|X, \lambda) \log \hat{\pi}_i \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} P(q_{t\in t_s} = i|X, \lambda) \log \hat{d}_i(\tau_t) \\ &+ \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\sum_{t=2}^{T} P(q_{t-1} = i, q_t = j|X, \lambda) \log \hat{a}_{ij} \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} P(q_t = i|X, \lambda) \log \hat{b}_{q_t}(\mathbf{x}_t). \end{aligned} \tag{2}$$

The duration density $d_i(\cdot)$ is incorporated instead of using self-transition probability a_{ii}. Transition probability parameters contain $A = \{a_{ij}, i \neq j\}$. In case of left-to-right HMM without state skipping, this duration penalty is activated when $\mathbf{x}_t$ stays at the starting frame t_s of a state. Let the state at moment t_s last for τ_{t_s} frames. Generally, the starting frame labels t_s of observations X can be found via Viterbi segmentation. Using the segmental ML method, we may determine the optimal state sequence with respect to the observation sequence X. The posterior probabilities are $P(q_1 = i|X, \lambda) = \delta(q_1 - i)$, $P(q_{t-1} = i, q_t = j|X, \lambda) = \delta(q_{t-1} - i)\delta(q_t - j)$ and

$P(q_t = i|X,\lambda) = \delta(q_t - i)$ where $\delta(\cdot)$ is Kronecker delta function. Specially, the posterior probability $P(q_{t\in t_s} = i|X,\lambda)$ of time t staying at frame t_s of state i is expressed by

$$P(q_{t\in t_s} = i|X,\lambda) = \delta(q_t - i)\delta(t - t_s) \equiv \xi_{t\in t_s}(i). \quad (3)$$

ML estimates of $\{\hat{\pi}_i\}$, $\{\hat{a}_{ij}\}$ and $\{\hat{b}_i(\cdot)\}$ have been well-known. This paper is focused on ML estimation of $\{\hat{d}_i(\cdot)\}$.

2.2. ML Estimation for Different Duration Parameters

1) Gaussian Duration Parameters D_α**.** Discrete duration length τ of state i can be simply represented by a continuous Gaussian density, $\tau \sim N(\mu_i, \sigma_i^2)$,

$$d_i(\tau|\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(\tau-\mu_i)^2}{2\sigma_i^2}\right], \quad (4)$$

where μ_i and σ_i^2 are mean and variance, respectively. Only positive value $\tau \geq 0$ is allowed in (4). In this case, duration parameters are formed by $D_\alpha = \{\mu_i, \sigma_i^2\}$. The expectation function for state i yields

$$Q_\alpha(\hat{\mu}_i,\hat{\sigma}_i^2|\mu_i,\sigma_i^2) \propto \sum_{t=1}^{T}\xi_{t\in t_s}(i)\cdot\left[\log\hat{\sigma}_i^2 + \frac{(\tau_t-\hat{\mu}_i)^2}{\hat{\sigma}_i^2}\right]. \quad (5)$$

When performing maximization step, the new ML estimates of duration mean and variance are respectively produced by

$$\hat{\mu}_i = \textstyle\sum_{t=1}^{T}\xi_{t\in t_s}(i)\cdot\tau_t \Big/ \sum_{t=1}^{T}\xi_{t\in t_s}(i) = \bar{\tau} \quad (6)$$

$$\hat{\sigma}_i^2 = \textstyle\sum_{t=1}^{T}\xi_{t\in t_s}(i)\cdot(\tau_t-\hat{\mu}_i)^2 \Big/ \sum_{t=1}^{T}\xi_{t\in t_s}(i). \quad (7)$$

These results are sample mean $\bar{\tau}$ and sample variance of duration lengths $\{\tau_{t_s}\}$ of a state.

2) Poisson Duration Parameters D_β**.** More precisely, the state duration variable τ can be represented using discrete Poisson distribution

$$d_i(\tau|\mu_i) = \frac{\mu_i^\tau}{\tau!}\exp[-\mu_i], \quad (8)$$

as suggested by Russell and Moore [7]. Only parameter μ_i exists in a Poisson distribution. Poisson duration parameters are constructed by $D_\beta = \{\mu_i\}$. The expectation function becomes

$$Q_\beta(\hat{\mu}_i|\mu_i) \propto \sum_{t=1}^{T}\xi_{t\in t_s}(i)\cdot[\tau_t\log\hat{\mu}_i - \hat{\mu}_i]. \quad (9)$$

Taking differentiation with respect to $\hat{\mu}_i$, we can derive the new parameter estimate, which has the same formula as (6).

3) Gamma Duration Parameters D_γ**.** Furthermore, it is popular to apply gamma distribution

$$d_i(\tau|\eta_i,\nu_i) = \frac{\eta_i^{\nu_i}}{\Gamma(\nu_i)}\tau^{\nu_i-1}\exp[-\eta_i\tau], \quad (10)$$

to fit the statistics of state duration. In (10), $\Gamma(\cdot)$ is gamma function, η_i and ν_i are parameters and $\eta_i > 0$, $\nu_i > 0$. Gamma duration parameters contain $D_\gamma = \{\eta_i, \nu_i\}$. We have

$$Q_\gamma(\hat{\eta}_i,\hat{\nu}_i|\eta_i,\nu_i) = \sum_{t=1}^{T}\xi_{t\in t_s}(i)\cdot\left[\begin{array}{l}-\log\Gamma(\hat{\nu}_i) + \hat{\nu}_i\log\hat{\eta}_i + \\ (\hat{\nu}_i - 1)\log\tau_t - \hat{\eta}_i\tau_t\end{array}\right]. \quad (11)$$

In [1][8], ML estimates of gamma parameters $(\hat{\eta}_i, \hat{\nu}_i)$ were calculated using the empirical method. Because gamma distribution $d_i(\tau|\eta_i,\nu_i)$ has mean ν_i/η_i and variance ν_i/η_i^2, we may empirically compute the sample mean $\hat{\mu}_i$ of (6) and sample variance $\hat{\sigma}_i^2$ of (7) using training samples $\{\tau_{t_s}\}$ of state i. The new gamma parameter estimates can be determined by $\hat{\eta}_i = \hat{\mu}_i/\hat{\sigma}_i^2$ and $\hat{\nu}_i = \hat{\mu}_i^2/\hat{\sigma}_i^2$.

3. Bayesian Learning of Duration Models

In real world, speech durations are varied for different speakers, speaking rates, environments, etc. A practical way is to adapt the state duration models to test conditions. Bayesian learning provides the fundamental theory. The maximum *a posteriori* (MAP) and quasi-Bayes (QB) estimates are respectively derived for batch and sequential learning of state duration parameters D.

3.1. MAP and QB Parameter Estimates

1) MAP Batch Learning. Namely, the model learning is performed when all adaptation data X are presented. MAP duration parameters D_{MAP} are obtained by maximizing the posterior density $p(D|X)$, or equivalently, the product of likelihood $p(X|D)$ and prior density $g(D)$,

$$D_{MAP} = \arg\max_D p(D|X) = \arg\max_D p(X|D)g(D). \quad (12)$$

When applying EM algorithm, we should determine the expectation function

$$R(\hat{D}|D) = E[\log p(\hat{D},\mathbf{q}|X)|X,D] = Q(\hat{D}|D) + \log g(\hat{D}). \quad (13)$$

MAP estimates are obtained by maximizing $R(\hat{D}|D)$. This paper presents the MAP estimate for gamma duration parameters.

2) QB Sequential Learning. However, it is difficult to trace the newest model statistics from batch learning data under the changing environments. It should be preferable to conduct sequential learning using online collected data. QB sequential learning is developed. Consider a sequence of adaptation data, $\chi^{(n)} = (X_1,\cdots,X_{n-1},X_n) = (\chi^{(n-1)}, X_n)$. Using QB theory [5], the posterior density of $\chi^{(n)}$ is approximated by the product of current data likelihood $p(X_n|D)$ and the prior density with hyperparameters $\varphi^{(n-1)}$ estimated from historical data $\chi^{(n-1)}$, i.e.

$$\begin{aligned} D_{QB} &= \arg\max_D p(D|\chi^{(n)}) = \arg\max_D p(X_n|D)g(D|\chi^{(n-1)}) \\ &\cong \arg\max_D p(X_n|D)g(D|\varphi^{(n-1)}). \end{aligned} \quad (14)$$

The resulting expectation function is given by

$$R^{(n)}(\hat{D}|D) = Q^{(n)}(\hat{D}|D) + \log g(\hat{D}\,|\,\varphi^{(n-1)}). \quad (15)$$

Herein, only Gaussian and Poisson duration models are considered for QB estimation because they are associated with conjugate priors $g(D|\varphi^{(n-1)})$. Sequential updating of hyperparameters from $\varphi^{(n-1)}$ to $\varphi^{(n)}$ is achievable [5].

3.2. MAP Estimation for Gamma Duration Parameters

To derive MAP estimates of gamma parameters $D_\gamma = \{\eta_i, \nu_i\}$, we assume that parameters η_i and ν_i are independent with

Gaussian priors, $\eta_i \sim N(\eta_i | \mu_{\eta_i}, \sigma_{\eta_i}^2)$ and $\nu_i \sim N(\nu_i | \mu_{\nu_i}, \sigma_{\nu_i}^2)$. The expectation function $R_\gamma(\hat{\eta}_i, \hat{\nu}_i | \eta_i, \nu_i)$ is accordingly defined. When performing M-step, the new estimate $\hat{\eta}_i$ is obtained via

$$\frac{\partial R_\gamma(\hat{\eta}_i, \hat{\nu}_i | \eta_i, \nu_i)}{\partial \hat{\eta}_i} = \sum_{t=1}^{T} \xi_{t \in t_s}(i) \left[\frac{\hat{\nu}_i}{\hat{\eta}_i} - \tau_t \right] - \frac{\hat{\eta}_i - \mu_{\eta_i}}{\sigma_{\eta_i}^2} = 0 \,. \quad (16)$$

However, we could not derive the closed form solution of $\hat{\nu}_i$. The Newton's algorithm is applied to iteratively reach the optimal solution

$$\hat{\nu}_i^{(k+1)} = \hat{\nu}_i^{(k)} - \frac{f(\hat{\nu}_i^{(k)})}{f'(\hat{\nu}_i^{(k)})}, \quad (17)$$

where

$$\frac{\partial R_\gamma(\hat{\eta}_i, \hat{\nu}_i | \eta_i, \nu_i)}{\partial \hat{\nu}_i} \propto \sum_{t=1}^{T} \xi_{t \in t_s}(i) \left[-\frac{\Gamma'(\hat{\nu}_i)}{\Gamma(\hat{\nu}_i)} + \log \hat{\eta}_i + \log \tau_t \right] - \frac{(\hat{\nu}_i - \mu_{\nu_i})}{\sigma_{\nu_i}^2} = f(\hat{\nu}_i). \quad (18)$$

3.3. QB Estimation for Gaussian Duration Parameters

When using Gaussian duration model in (4) with parameters $D_\alpha^{(n)} = \{\mu_i^{(n)}, \sigma_i^{(n)^2}\}$, we assume that the parameter $\mu_i^{(n)}$ is random with Gaussian prior $\mu_i^{(n)} \sim N(m_\mu^{(n-1)}, \rho_\mu^{(n-1)^2})$ and $\sigma_i^{(n)^2}$ is fixed but unknown. After EM implementation for $\chi^{(n)}$, the new QB estimate $\hat{\mu}_i^{(n)}$ is formulated by

$$\hat{\mu}_i^{(n)} = \frac{\rho_\mu^{(n-1)^2} \sum_{t=1}^{T} \xi_{t \in t_s}(i) \tau_t^{(n)} + \sigma_i^{(n)^2} m_\mu^{(n-1)}}{\rho_\mu^{(n-1)^2} + \sigma_i^{(n)^2}} \,. \quad (19)$$

3.4. QB Estimation for Poisson Duration Parameters

In [7], the Poisson distribution was studied for duration model estimation. The QB sequential learning mechanism was exploited. Here, the prior density of the parameter $D_\beta^{(n)} = \{\upsilon_i^{(n)}\}$ of Poisson distribution is assumed to be gamma distribution

$$g(D_\beta^{(n)} \mid \varphi_\beta^{(n-1)}) \propto \upsilon_i^{(n)^{\nu_\upsilon^{(n-1)}-1}} \exp[-\eta_\upsilon^{(n-1)} \upsilon_i^{(n)}], \quad (20)$$

where hyperparameters are $\varphi_\beta^{(n-1)} = (\eta_\upsilon^{(n-1)}, \nu_\upsilon^{(n-1)})$. Notably, the prior density using gamma distribution belongs to conjugate prior family. This attractive property is useful to derive sequential learning mechanism of hyperparameters.

Given Poisson duration models and the gamma prior pdf, the expectation function becomes

$$\begin{aligned} R_\beta(\hat{\upsilon}_i^{(n)} | \upsilon_i^{(n)}) &\propto \sum_{t=1}^{T} \xi_{t \in t_s}(i) \left[\tau_t^{(n)} \log \hat{\upsilon}_i^{(n)} - \hat{\upsilon}_i^{(n)} \right] + (\nu_\upsilon^{(n-1)} - 1) \log \hat{\upsilon}_i^{(n)} - \eta_\upsilon^{(n-1)} \hat{\upsilon}_i^{(n)} \\ &= \left[\left(\sum_{t=1}^{T} \xi_{t \in t_s}(i) \tau_t^{(n)} + \nu_\upsilon^{(n-1)} \right) - 1 \right] \log \hat{\upsilon}_i^{(n)} - \left[\sum_{t=1}^{T} \xi_{t \in t_s}(i) + \eta_\upsilon^{(n-1)} \right] \hat{\upsilon}_i^{(n)} \\ &= (\nu_\upsilon^{(n)} - 1) \log \hat{\upsilon}_i^{(n)} - \eta_\upsilon^{(n)} \hat{\upsilon}_i^{(n)} \end{aligned} \quad (21)$$

The posterior and the prior densities of Poisson distribution are belonging to gamma distribution [2], then the updated hyperparameters $\varphi_\beta^{(n)} = (\eta_\upsilon^{(n)}, \nu_\upsilon^{(n)})$ are represented

$$\begin{aligned} \eta_\upsilon^{(n)} &= \sum_{t=1}^{T} \xi_{t \in t_s}(i) + \eta_\upsilon^{(n-1)}, \\ \nu_\upsilon^{(n)} &= \sum_{t=1}^{T} \xi_{t \in t_s}(i) \tau_t^{(n)} + \nu_\upsilon^{(n-1)}. \end{aligned} \quad (22)$$

The updated $\hat{\upsilon}_i^{(n)}$ is calculated by

$$\hat{\upsilon}_i^{(n)} = \frac{\sum_{t=1}^{T} \xi_{t \in t_s}(i) \tau_t^{(n)} + \nu_\upsilon^{(n-1)} - 1}{\sum_{t=1}^{T} \xi_{t \in t_s}(i) + \eta_\upsilon^{(n-1)}} \,. \quad (23)$$

4. Experiments

4.1. Speech Databases and Experimental Setup

The speaker adaptation task is conducted to examine the performance of the proposed sequential learning methods of duration model. It is performed on a continuous Mandarin speech recognition task. We adopted the microphone-based TCC300 speech database to estimate the SI HMM's and the associating duration parameters of different parametric distributions and to obtain the baseline recognition rate. There were totally 15851 utterances (about 16 hours) covering short and long sentences. We randomly selected 1500 utterances for recognition and the other 14266 for training. In order to reveal the improvement of the recognition performance using MAP and QB algorithms under different speaking rates and noise condition, there was another testing database prepared. It is the broadcast news database, provided by Academia Sinica, Taiwan. It was recorded from radio stations and its speaking rate was different from that of TCC300. Three speakers were selected for the evaluation of the MAP adaptation and QB sequential learning mechanism. For each speaker, there were totally 30 adaptation utterances. In the QB sequential learning, we used five utterances in each epoch and there were totally 6 epoches during the sequential learning process. The syllable error rate (SER) was averaged over these speakers. Those adaptation utterances per speaker were used for HMM's and duration parameter sequential learning. We performed supervised learning.

All utterances were sampled at 16 kHz with 16-bit resolution. Each frame was characterized by twelve Mel-frequency cepstral coefficients (MFCCs), one log energy and their derivatives. The cepstral mean subtraction (CMS) was applied for each utterance. In the following experiments, we carry out the HMM-based speech recognition with/without duration models and the MAP adaptation and QB sequential learning of different parametric duration distributions.

4.2. Comparison of Speaking Rates and Baseline Recognition Performances

First, the average speaking rates of two databases are compared in Table 1. Obviously, the speaking rate of broadcast news is much higher than those of TCC300. The performances of baseline speech recognition system with/without duration models are compared in Table 2. In this case, only ML duration estimation is performed. Clearly, the recognition performance using gamma duration model is superior to those using Poisson and Gaussian duration models.

Database		TCC300	Broadcast news
Speaking rate (syl/sec)	Male	3.55	5.50
	Female	4.86	5.47

Table 1. Comparison of speaking rates of two databases

	Without duration	With durations		
		Gaussian	Poisson	gamma
SER	38.2	36.9	36.4	35.6

Table 2. Syllable error rates (%) of the recognition system with/without duration models

4.3. Evaluation of MAP Adaptation and QB Sequential Learning

In Figure 1, we plot the recognition results of MAP adaptation. In this experiment, we conduct the adaptation for each testing speaker. Totally, thirty adaptation utterances are used to adapt the HMM's and the associated duration parameters. In this figure, the performance improvement order from the best to the worst is the result using gamma duration with Gaussian prior, the one using Poisson duration with gamma prior and the one using Gaussian duration with Gaussian prior of batch adaptation. Although gamma duration achieves better recognition result, our other experiments show that the processing time using gamma duration is more expensive.

Figure 2 shows the recognition results using QB sequential learning approach. In this part, we conduct the QB sequential adaptation for each testing speaker. Five utterances are used in each adaptation epoch for each speaker. The hyperparameters of HMM's and duration parameters are also estimated in each epoch. From this figure, the performance improvement of incremental adaptation can be observed. The performance of sequential learning using Poisson duration with gamma prior is better than the one using Gaussian duration with Gaussian prior. It should be noted that the adaptation performance using Poisson duration with gamma prior is always superior to the one using Gaussian duration with Gaussian prior either in MAP batch adaptation or in QB sequential learning if we exclude the gamma duration.

5. Conclusion

We proposed the joint Bayesian learning framework of HMM's as well as duration parameters. Three cases of Gaussian, Poisson and gamma densities for duration modeling were evaluated. ML estimation of duration parameters was described. In order to adapt duration models to a new speaker, we performed Bayesian learning and compensated the effect of speaking rate. MAP estimation of duration parameters was derived for speaker adaptation from clean TCC300 database to broadcast news speech database. The adaptation of HMM mean vectors and duration parameters did enhance the recognition performance. Furthermore, QB estimates for Gaussian and Poisson duration models were formulated because they were associated with conjugate priors. The reproducible prior/posterior property was applied to establish updating mechanism for prior statistics. Sequential learning of duration models was achievable using sequentially collected adaptation data in each epoch. The Poisson-based duration model is the best choice because its prior and posterior distributions are belonging to gamma distribution. The best performance was achieved when HMM mean vectors and duration parameters were simultaneously adapted. In the future, we will further investigate duration modeling using alternative distributions, e.g. alpha-stable distributions.

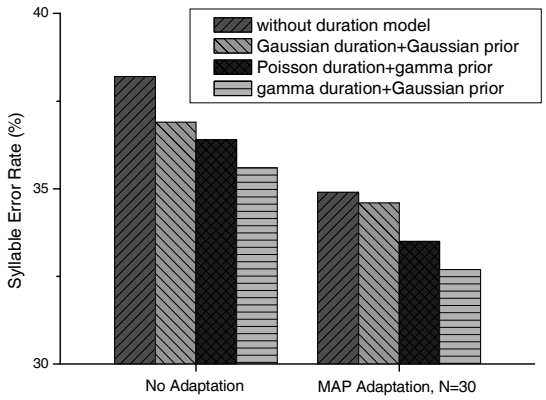

Figure 1. Performances of MAP adaptation using different duration distributions and priors

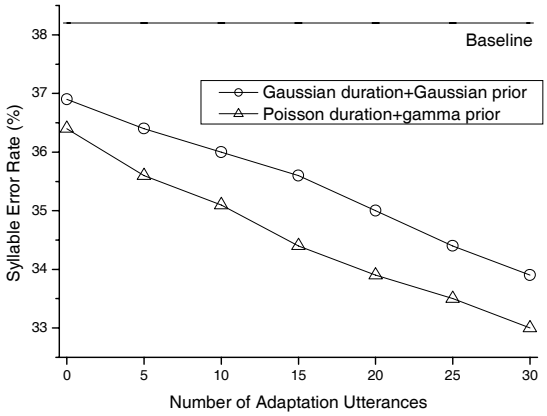

Figure 2. Performances of QB sequential learning under different duration models

6. References

[1] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 240-242, May 1996.

[2] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

[3] P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society (B)*, vol. 39, pp. 1-38, 1977.

[4] J. D. Ferguson, "Variable duration models for speech", *Proceedings of Symposium on the Application of Hidden Markov Models to Text and Speech*, pp. 143-179, 1980.

[5] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 2, pp. 161-172, March 1997.

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

[7] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition", *ICASSP*, pp. 5-8, 1985.

[8] N. B. Yoma, F. R. McInnes, M. A. Jack, S. D. Stump and L. L. Ling, "On including temporal constraints in Viterbi alignment for speech recognition in noise", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 179-182, February 2001.