label sequence $\boldsymbol{b} = (b_1, b_2, \ldots, b_N)$ of N phonemes. T is the length of the sentence. T denotes the transpose. Explicit duration modeling is used in hidden semi-Markov model (HSMM) for HTS proposed by Yoshimura et al. [8]. The likelihood is decomposed into two parts

$$\begin{aligned} \hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} \sum_{all \boldsymbol{q}} p\left(\boldsymbol{o} \mid \lambda, \boldsymbol{q}\right) p\left(\boldsymbol{q} \mid \lambda, \boldsymbol{b}\right) \\ &\approx \arg\max_{\boldsymbol{o}} p\left(\boldsymbol{o} \mid \lambda, \hat{\boldsymbol{q}}\right) p\left(\hat{\boldsymbol{q}} \mid \lambda, \boldsymbol{b}\right) \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{q}}$ is the optimal sequence of Gaussian distributions predicted by the duration model independent of $\boldsymbol{o}$ [3]. The search for all possible $\boldsymbol{q}$ is intractable. Therefore, (1) is decomposed