Duration Refinement by Jointly Optimizing State and Longer Unit Likelihood

Boyang Gao1∗,2, Yao Qian1, Zhizheng Wu1∗,3, Frank K. Soong1

1Microsoft Research Asia, Beijing, P.R. China
2Department of Computer Science and Technology, Harbin Institute of Technology, P.R. China
3College of Software, Nankai University, P.R. China
{yaoqian,frankkps}@microsoft.com

Abstract

We refine the duration model in HMM-based TTS by extending the work of Wu [1]. The model is refined by jointly maximizing the duration likelihoods of state, phone and syllable units. Both Gaussian and gamma distributions are employed. In synthesis, the state durations are generated by the same joint optimization procedure. By considering the duration of state and longer units jointly, the accumulation of errors in estimated state durations is regulated in the optimization procedure. Experiments on Mandarin and English databases show that the refined model yields more accurate duration predictions, compared with the baseline state duration model. The improvement of phone RMSEs are 2.2ms and 1.1ms or 11% and 5.6%, in English and Mandarin synthesis, respectively. The perceptual test on synthesized English and Mandarin speech further confirms that the refined duration model outperforms the baseline system.

Index Terms: HMM-based TTS, duration model, gamma distribution

1. Introduction

In HMM-based speech synthesis, speech parameter modeling and generation is based upon the maximum likelihood (ML) criterion. Given the state sequence, a parameter trajectory is generated, under the dynamic feature constraints, to maximize its likelihood [2]. The state sequence is actually obtained from the sequence of state durations predicted by the estimated state duration model. The duration, which is pertinent to the perceived quality of synthesized speech, needs to be cogently predicted. However, the current HMMs cannot predict duration information very accurately and the resultant supra-segmental quality of synthesized speech suffers [3]. The state duration of a standard HMM is explicitly modeled with a single Gaussian distribution which is estimated by using state occupancy counts in the Baum-Welch re-estimation procedure [4].

There are many studies on how to improve duration models. In [3], a Hidden Semi-Markov model (HSMM), which models explicitly the duration distribution, was proposed. Different from the conventional HMM framework where the state duration is implicitly defined by the state transition probability as an exponential distribution, HSMM defines explicitly the state duration in a nonparametric form of histograms or a parametric form like Gaussian distribution. Gaussian distribution, however, with its infinite support extending from negative infinity to positive infinity, is inappropriate for modeling the positive semi-definite nature of duration information. Gamma distribution, which can model positive definite random variables, is more appropriate for modeling state duration and was proposed before [5, 6, 7]. In [1], phone duration is integrated with state duration in the generation procedure where duration likelihood of state and phone is jointly maximized.

In speech production, durations of a short unit like state is actually regulated by the durations of longer units, e.g., phone, syllable and word, etc. The duration assignment of different units is actually done in a highly regulated, hierarchical manner. In this paper, motivated by the work of Wu [1], we extend the joint optimization to include state, phone and syllable for HMM-based TTS. By analyzing the duration distribution of longer units, we found that gamma distribution fits the data better than Gaussian distribution. We then improve the conventional HMM duration model by a new jointly optimized duration model over all constituent units.

The rest of the paper is organized as follows. In Section 2 we compare Gaussian and gamma distributions in their model fitting to the duration distributions of longer units. In Section 3 we describe the method that combines phone and syllable duration with state’s meanwhile maximizing likelihood of duration sequence. Results of experiments and analysis are shown in Section 4. The conclusion and remarks are presented in the final section.

2. Duration Modeling for Long Term Units

To study the duration of longer units in Mandarin and English, we first analyze the distribution of phone and syllable durations. The data is obtained by forced aligning speech training data with acoustic HMMs trained for TTS. Syllables are clustered into groups according to their length, the number of phones. The distributions (histograms) of phone and syllable durations resemble more like a gamma than a Gaussian distribution as shown in

Figure 1: Histograms for duration of different units.

The figure shows typical examples of duration histograms
of English 3-phone syllables and diphthong "ao". The distribution is further tested by a Chi-Square statistics for measuring the goodness-of-fit [8]. In HMM-based speech synthesis, models of rich contexts are used to characterize the co-articulation effects. In practice, however, limited by insufficient training data, we tie models of rich and detailed contexts into generalized ones so as to predict unseen contexts robustly. State tying via a clustered decision tree is commonly used. We test the distribution of durations in each leaf node of decision tree by the Chi-Square test of goodness-of-fit. Table 1 shows the percentage of leaf nodes where gamma distributions fit better than Gaussians, i.e., with a smaller value of $\chi^2$ statistics. However, in English the difference between Gaussian and gamma is more distinctive than in Mandarin.

<table>
<thead>
<tr>
<th>Gamma better (%)</th>
<th>phone</th>
<th>syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>62.9</td>
<td>61.9</td>
</tr>
<tr>
<td>Mandarin</td>
<td>38.6</td>
<td>35.5</td>
</tr>
</tbody>
</table>

In modeling duration of different units, full-context units are clustered via regression trees [7]. The question set used in clustering phone and syllable duration into decision tree includes tones and break for Mandarin; stress, accent and POS for English; quin-phone, the position of phone, syllable and word in phrase and sentence, and the length of word and phrase for both Mandarin and English. Duration of each leaf node is modeled either with a Gaussian or gamma distribution. Gamma distribution has the form of

$$p(x) = \frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x/b}$$

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x} dx$. The expectation and variance of random variable $x$ under gamma distribution are $E(x) = ab$ and $Var(x) = ab^2$, where $a$ and $b$, $a = \mu^2/\sigma^2$, $b = \sigma^2/\mu$, are functions of $\mu$ and $\sigma^2$, the mean and variance of duration for a leaf node.

3. Duration Estimation by Jointly Maximizing the Likelihoods of Different Units

State durations can be estimated more precisely if they can be regulated by the durations of longer and higher level units like phone and syllables. The approach was originally proposed by Wu [1] where the state duration is estimated jointly with the phone duration. In this study we extend it further to include syllable duration and adopt gamma pdf as the distribution function. The likelihood of state durations is jointly maximized in conjunction with the weighted likelihoods of phone and syllable durations. The log likelihood $L$ of duration is maximized as

$$\max L = \log P(d | q, \lambda)$$

$$= \sum_j \left[ \sum_n \left[ \sum_k \log p_{j,n,k}(d_{j,n,k}) + \alpha \log p_{j,n}(d_{j,n}) \right] + \beta \log p_j(d_j) \right] \tag{1}$$

subject to

$$\sum_k d_{j,n,k} = d_{j,n} \tag{2}$$

$$\sum_n d_{j,n} = d_j \tag{3}$$

where $d$ is the sequence of state duration; $q$ is the corresponding state sequence; $\lambda$ is the duration model; $d_{j,n,k}$ is the duration of state $k$, phone $n$, and syllable $j$. Correspondingly, $p_{j,n,k}(\cdot)$ is the probability density function of $d_{j,n,k}$; $p_{j,n}(\cdot)$ and $p_j(\cdot)$ are similarly defined. Two parameters, $\alpha$ and $\beta$, are to weight phone and syllable durations likelihood.

We use Gaussian and gamma distributions for modeling phone and syllable durations to refine Gaussian model of state duration. Gaussian is still used despite the fact that the duration histograms of phone and syllable look more gamma like than Gaussian for two reasons: Gaussian has been used for modeling duration information as a default or benchmark distribution; and Gaussian is used for its mathematical tractability with the first and second order moments. To maximize likelihood $L$ in Gaussian distribution we set

$$\frac{\partial L}{\partial d_{j,n,k}} = \frac{d_{j,n,k} - \mu_{j,n,k}}{\sigma_{j,n,k}^2} + \alpha \frac{d_{j,n} - \mu_{j,n}}{\sigma_{j,n}^2} + \beta \frac{d_j - \mu_j}{\sigma_j^2} = 0 \tag{4}$$

$$d_{j,n,k} = \mu_{j,n,k} + \left[ -\alpha \frac{d_{j,n} - \mu_{j,n}}{\sigma_{j,n}^2} - \beta \frac{d_j - \mu_j}{\sigma_j^2} \right] \sigma_{j,n,k}^2 \tag{5}$$

In Eq. (5) $d_{j,n}$ and $d_j$ are unknown variables and can be solved by applying the constraints in (2) and (3) to (5). Two resultant linear equations are solved to obtain

$$d_{j,n} = \frac{\sigma_{j,n}^2 \sum_k \sigma_{j,n,k}^2 M_{j,n,k} + \alpha \sum_n \mu_{j,n} V_{j,n}}{\sum_k \sigma_{j,n,k}^2 + \alpha \sum_n \mu_{j,n} V_{j,n}}$$

$$d_j = \frac{\sigma_j^2 \sum_n \sigma_{j,n}^2 M_{j,n} + \alpha \sum_k \mu_{j,n} V_{j,n} + \beta \sum_k \mu_{j,n,k} V_{j,n}}{\sigma_j^2 + \beta \sum_k \sigma_{j,n,k}^2 V_{j,n}} \tag{7}$$

where

$$M_{j,n} = \sum_k \mu_{j,n,k}$$

$$V_{j,n} = \sum_k \sigma_{j,n,k}^2$$

$$D_{j,n} = \sigma_j^2 + \alpha \sum_k \sigma_{j,n,k}^2$$

When gamma distribution is employed instead of Gaussian to model the durations of state, the formulas have too complicated form for practical implementation. In addition, five-state HMM phone model is used in our system and the state duration in terms of the number of frames per state ranges from 1 to 5 observed in 90% of states. It is difficult to tell the difference between Gaussian and gamma distributions on such a small scale.
Therefore, we only use gamma distribution to model phone and syllable durations and replace the corresponding Gaussian models with gamma models for maximizing the likelihoods in Eq. (1). Similar as Eq. (5), we have

\[ d_{j,n,k} = \mu_{j,n,k} + \rho_{j,n} \cdot \sigma_{j,n,k}^2 \]  

(8)

where

\[ \rho_{j,n} = -\alpha \left( \frac{1}{b_{j,n}} + \frac{1 - a_{j,n}}{d_{j,n}} \right) - \beta \left( \frac{1}{b_j} + \frac{1 - a_j}{d_j} \right) \]

and \(a_{j,n}, b_{j,n}, a_j, b_j\) denote the parameters of the gamma distribution associated with \(d_{j,n}\) and \(d_j\). Applying constraints in (2) and (3) to (8). The optimization ends up with solving two quadratic equations:

\[
d_j^2 - \sum_n \left( (P_{j,n}^2 - 4C_{j,n})d_j^2 + 2N_{j,n}P_{j,n}d_j + N_{j,n}^2 \right) + d_j \sum_n P_{j,n} + \sum_n N_{j,n} = 0 \]  

(9)

\[
d_j^2 - \sum_n \left( (P_{j,n}^2 - 4C_{j,n})d_j^2 + 2N_{j,n}P_{j,n}d_j + N_{j,n}^2 \right) + d_j \sum_n P_{j,n} + \sum_n N_{j,n} = 0 \]  

(10)

where

\[ C_{j,n} = a(1 - a_{j,n})V_{j,n} \]

\[ P_{j,n} = -M_{j,n} + \frac{a_j}{b_j}V_{j,n} + \frac{\beta}{b_j}V_{j,n} \]

\[ N_{j,n} = \beta(1 - a_j)V_{j,n} \]

and \(M_{j,n}, V_{j,n}, D_{j,n}\) are defined the same as in Gaussian scenario. Note that Eq. (10) has no closed form solution, however, \(\sum_n \left( (P_{j,n}^2 - 4C_{j,n})d_j^2 + 2N_{j,n}P_{j,n}d_j + N_{j,n}^2 \right) \) can be well approximated by a linear function of \(d_j\) derived from Taylor series expanding, which leads to the solvable quadratic equation:

\[
d_j^2 + \sum_n P_{j,n} + \sum_n N_{j,n} = 0 \]  

(11)

Where \(f_{j,n}(x) = \sqrt{(P_{j,n}^2 - 4C_{j,n})x^2 + 2N_{j,n}P_{j,n}x + N_{j,n}^2} \) and \(d_j\) is the point at which Taylor series expanding is performed and in practice we use \(\mu_j\). Solving Eq. (9) and (11), we can obtain \(d_{j,n,k}\). Limited by space, we refer detail derivations of the solutions to [9] for both cases of Gaussian and gamma distributions.

4. Experiments and Results

4.1. Experiments setups

Two phonetically and prosodically rich, broadcast news style corpora in American English and Mandarin are used in our experiments. The two corpora are divided into three sets: training, developing and testing parts with corresponding sizes in number of sentences given in Table 2. The training set is used for training the refined duration model. Developing set is used to determine the appropriate weights \((\alpha, \beta)\) in the refined model. The testing set is used to measure the performance of the refined duration model.

The performance of the proposed approach is measured objectively and subjectively. The objective measure of the root mean squared errors (RMSE) of phone durations between natural speech of the original speaker and generated speech is calculated to measure duration distortions. The underlying phone and syllable durations of the natural speech are obtained by forced alignment with the acoustic models trained in our baseline HMM-based speech synthesis system. Subjectively, a preference test is conducted to compare speech sentence pairs synthesized by the refined model and the baseline. The duration in the baseline system is from a state modeled with Gaussian distribution. The phone and syllable durations are obtained by accumulating the corresponding state durations. We tested the proposed method for both Gaussian and gamma distributions.

4.2. Experiment results and analysis

To evaluate the performance of the refined model on longer units, we use parametric duration models in predicting the corresponding phone and syllable durations. For a specific unit, the maximum likelihood estimate (prediction) is the mode of its model. For Gaussian distribution, the mode is the mean and for gamma distribution, the mode is \(\mu = \frac{\alpha}{\beta} \). The RMSEs of phone and syllable durations predicted by the baseline, Gaussian and gamma distributions for developing sets in both English and Mandarin corpora are shown in Table 3. It shows that the refined duration model of longer units (phone and syllable) outperforms the baseline model for both English and Mandarin corpus. The Gaussian distribution outperforms gamma distribution in Mandarin corpus. By further checking the goodness of fit shown in Table 1, although more leaf nodes fit gamma better than Gaussian, some bad cases which are much far from gamma than Gaussian are observed.

<table>
<thead>
<tr>
<th># sentences</th>
<th>training</th>
<th>developing</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>2,586</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Mandarin</td>
<td>5,100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2: Training, developing and testing sets (# sentences) of the two speech corpora.

<table>
<thead>
<tr>
<th>RMSE (ms)</th>
<th>English</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>19.0</td>
<td>18.1</td>
</tr>
<tr>
<td>Syllable</td>
<td>34.5</td>
<td>26.7</td>
</tr>
<tr>
<td>Phone</td>
<td>18.8</td>
<td>19.1</td>
</tr>
<tr>
<td>Syllable</td>
<td>33.0</td>
<td>27.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE (ms)</th>
<th>English</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>20.3</td>
<td>19.1</td>
</tr>
<tr>
<td>Syllable</td>
<td>33.3</td>
<td>30.3</td>
</tr>
</tbody>
</table>

Table 3: The RMSEs of phone and syllable durations of the developing set.

To find the optimal \((\alpha, \beta)\) values for maximizing the joint likelihoods of state, phone and syllable duration, we use the development set via a grid search in the two dimensional space of \((\alpha, \beta)\). The grid search for finding the best \((\alpha, \beta)\) value in gamma distribution is in Figure 2. The optimal \((\alpha, \beta)\) locates at \((1.4, 0.6)\), which reduces RMSE of 20.3ms of the baseline to 18.15ms. The optimized weights are used in generating ML predicted durations for synthesis according to Eq. (1). The RMSE results of predicted phone duration are shown in Table
Figure 2: Full grid search of optimal (α, β) values which minimize RMSE for the gamma distribution.

4. The proposed refined duration model performed better in the English corpus than in Mandarin one with respective relative improvements of 11.0% and 5.6%.

Table 4: RMSE and relative improvement (%) for baseline and the refined duration model

<table>
<thead>
<tr>
<th></th>
<th>RMSE(ms) &amp; RI(%)</th>
<th>English</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td></td>
<td>ph</td>
<td>ph &amp; syl</td>
</tr>
<tr>
<td>Gamma</td>
<td></td>
<td>17.8(10.1)</td>
<td>17.6(11.0)</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>19.8</td>
<td></td>
</tr>
</tbody>
</table>

The breakdown of RMSE improvement for the test sentences shows that about 77% of sentences predicted by the refined duration model have lower RMSE than those of the baseline. Figure 3 shows the scattered diagram of RMSE pairs of phone durations predicted by the baseline and the refined model, plotted for each individual phone in English. Note that with only two exceptions of two phones “v” and “g”, predicted phone duration is universally improved by the refined model over that of the baseline. Similar but less improvements are observed in Mandarin phones.

Figure 3: Scatter diagram of phone RMSE pair (predicted by baseline and refined model) of all English phones

The refined duration model is further evaluated by a perceptual test. 50 Mandarin and 50 English sentences, which are selected from the testing set and synthesized by the baseline and the new refined duration model, are evaluated in an AB preference test. In the 100 sentences, the RMSE of phone duration between the baseline and the refined duration model is larger than 3ms. 10 subjects composed of 3 English, 3 Mandarin language experts (LE) and 4 bilingual graduate students (STU) participate in the preference test. There are three preference choices: 1) the former is better; 2) the latter is better; 3) no preference (The difference between the paired sentences cannot be perceived or the difference can be perceived but it is difficult to choose which one is better). The preference scores between the baseline and the refined duration model are shown in Table 5. It shows that the speech synthesized by the refined duration model outperforms the baseline system perceptually.

Table 5: The preference score of the baseline and the refined duration model

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>LE</th>
<th>STU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>Refined duration</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>No preference</td>
<td>32</td>
<td>60</td>
</tr>
</tbody>
</table>

5. Conclusion

We propose a refined duration model which jointly optimizes the likelihoods of state, phone and syllable durations. The joint optimization procedure is similarly applied to duration generations in synthesis. Both Gaussian and gamma distributions are investigated as model parameters. The refined model improves duration prediction: the RMSEs of phone durations are reduced by 2.2 and 1.1 ms, or 11% and 5.6% relative reductions, in English and Mandarin synthesis. The synthesized speech by our refined duration model also receives a higher preference score in a perceptional test, compared with that of the baseline.

6. References