Context-dependent Duration Modeling with Backoff Strategy and Look-up Tables for Pronunciation Assessment and Mispronunciation Detection

Hongyan Li, Shen Huang, Shijin Wang, Bo Xu

Digital Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{hyli, shenhuang, sjwang, xubo}@hitic.ia.ac.cn

Abstract

This paper makes an intensive study on the contextual modeling methods of duration information, for the purpose of improving the performance of pronunciation assessment and mispronunciation detection. The main ideas include: 1) we extend the relations among duration sequence with different level of contextual constraints, and bring them into a unified framework. 2) A backoff mechanism is introduced to resolve the problem of data sparseness and unbalanced distribution. 3) Rather than the traditional parametric functions, we use the discrete modeling for empirical duration distributions based on look-up tables, which can improve the model precision and accelerate the computation speed. The experimental results show the effectiveness of the above methods. The proposed word-dependent duration models can yield 0.0782 in absolute CC (correlation coefficient) improvement and 4.58% in absolute EER (equal error rate) reduction for the tasks of pronunciation assessment and mispronunciation detection respectively, both compared with the baseline method with conventional context-independent case.

Index Terms: duration model, backoff, look-up table, pronunciation assessment, mispronunciation detection

1. Introduction

Duration information has been successfully used in many applications, such as speech recognition [1, 2, 3], automatic segmentation [4] and utterance filtering [5]. Recently, there has been increased interest in computer-assisted language learning (CALL), and the duration is often regarded as one of the pronunciation quality features [6].

In practice, the acoustic confidences still play the most significant role for pronunciation assessment task, especially mispronunciation detection. As we know, the acoustic models focus on the probabilistic behaviors of the given observation sequences. However, most of the acoustic models tend to decompose the segment of the observation sequence into frame based units and then measure them in an independent way due to the limitation of model complexity and parameter estimation, for example, the GMM-UBM, GLDS-SVM [7] and TRAP-NN [8] models in our previous work are all frame based methods to obtain acoustic scores. Obviously, these assumptions lack the information about the whole segment more or less, and the frame based operations can be seen as a compromise between the good modeling ability and the low model complexity. In our opinions, a good way to solve the above shortcomings is to use the segmental information. The duration, which is highly correlative to the characteristics of the segment, is preferred as the useful complement for the traditional acoustic features. Another advantage of duration is that it has simple structure or representation (usually with one dimension of duration in frames or normalized duration), so we can adopt more subtle and complex modeling methods to describe its distribution without increasing too much system load.

Typically, duration is often modeled with a certain distribution assumption (e.g. exponential, Gaussian, Gamma, etc.), and then the likelihood of duration model is combined with the acoustic scores to improve the performance [3, 9]. D. Povey [2] pointed out that the durations in speech are difficult to model accurately with some probability distribution functions (pdf) because they have a very non-Gaussian shape, including a long tail and a sharp cutoff below the minimum duration limitation of the HMM framework, and this makes a discrete distribution attractive for modeling durations. In fact, directly modeling the duration with a pdf is a viable solution for small corpus tasks (due to the data sparseness), but for large corpus it is more appropriate to use discrete distribution for duration modeling. With a large corpus in our work, we propose a discrete duration modeling method based on look-up tables for practical applications.

Compared with the recognition task, it needs higher requirements on pronunciation models for the tasks of pronunciation assessment and mispronunciation detection. In particularly, mispronunciation detection is more sensitive to pronunciation models because we might make a hard decision for the correctness of the given phone or word. However, most of the duration features for pronunciation assessment task are based on simple context-independent models without the constraint limitation. Factually, the relation among durations in a word or sentence has an important impact on duration modeling. Consequently, in order to improve the precision of duration models, we attempt to model the phone durations with contextual constraints or within a word. Moreover, for some longer words (with three or more syllables), the context-dependent duration models are even more necessary. Since there obviously will not be enough training data for each context, an improved hierarchical backoff mechanism will be explored in our work.

2. Context-dependent duration modeling

Tang and Liu [12] have introduced a concept of “tri-gram duration model” to measure the rate of speech. In our work, with reference to N-gram language models in speech recognition, we will extend the property of N-gram to depict the relation among durations with different level of contextual constraints, and then bring them into a unified framework.

Given the phone sequence of a word by \( a_1, a_2, \ldots, a_n \), and the corresponding durations are \( l_1, l_2, \ldots, l_n \), the duration probability of phone \( a_i \) is calculated as follow:

\[
P(l_i | a_i | C) = \frac{P(l_i, a_i, C)}{P(l_i, a_i)}
\]  

(1)

where \( C \) stands for the contextual constraints. Here, we can define \( C_{tr} \), \( C_{cn} \) and \( C_{wd} \) as context-independent, context-
dependent (tri-gram) and word-dependent duration models respectively, i.e.,

\[ C_{CD} = \{ \text{none} \} \]  \hspace{1cm} (2)

\[ C_{WD} = \{ l_i, a_i, l_{i+1}, a_{i+1} \} \]  \hspace{1cm} (3)

\[ C_{WD} = \{ l_1, a_1, \ldots, l_N, a_N \} \]  \hspace{1cm} (4)

Then, the corresponding duration probabilities upon different duration models can also be obtained.

In term of discrete distribution, we classify the duration of each given segment sets evenly into \( T \) discrete duration levels by their real length (e.g. normalized by rate of speech), and the discrete duration probability can be represented as follow,

\[ P(l, a | C) = \{ p', l', \ldots, p', l'_T \} \]  \hspace{1cm} (5)

where \( p'_i \in [0,1] \) is the probability of the \( i \) level duration.

Similarly, formula (1) can be rewritten in discrete form as follow,

\[ P(l', a | C) = \frac{P(l', a, C)}{\sum_{i=1}^T P(l'_i, a_i, C)} \]  \hspace{1cm} (6)

### 2.1. Improved backoff mechanism for context-dependent duration models

Obviously, the main difficulty of the proposed context-dependent duration models, as with other probability estimation problems (e.g. language models), is the sparseness of the data. The solutions to this kind of problems for other tasks such as language modeling [13, 14] may also have relevance for the duration modeling. Therefore, we introduce a hierarchical backoff mechanism to solve the problem of unseen symbols, which might occur in the context of \( C_{CD} \) or \( C_{WD} \) duration models. That is, for unseen contexts for a given phone duration, or when the number of samples for a give contextual case is below a predefined threshold, the higher level contextual models will back off to lower level contextual models, e.g., \( C_{WD} \) backed off to \( C_{CD} \) or even \( C_{CD} \).

Towards the tasks of pronunciation assessment and mispronunciation detection, there exist 44 phonemes and a silence (the symbol “sil” is fixed as context cues for word boundaries) in the phone set. So, if we need to measure the duration distribution on phone level, we should compute \( 44 \times T \) and \( 45 \times (44 \times T) \) probabilities for \( C_{CD} \) and \( C_{WD} \) (tri-gram with silence), respectively. As for the case of \( C_{WD} \), suppose there are \( M \) words in the vocabulary, and the word \( M \), has \( S_j \) syllables, the number of the calculated probabilities will increase to \( \sum_{i=1}^M \sum_{j=1}^{S_j} \times T \). For example, if we normalize the duration length to 30 levels and have 1000 words with 3 syllables, that is \( T = 30, M = 1000, S_j = 3 \), the number of probabilities for \( C_{CD} \), \( C_{CD} \) and \( C_{WD} \) are 1320, 2,673,000 and 90,000, respectively. Additionally, taking into account the phonological positions of vowels and consonants, it will reduce by half the number of calculations for \( C_{WD} \) approximately.

We can see that, \( C_{CD} \) has much larger computational burden than \( C_{WD} \), and the burden of \( C_{WD} \) will be significantly increased with the capacity of vocabulary. When there are 100,000 words in vocabulary, the computational burden of \( C_{WD} \) will achieve to the same order of magnitude with that of \( C_{CD} \). That means, in term of computational burden, \( C_{WD} \) with 100,000 words is equivalent to \( C_{CD} \). On the other hand, considering the precision of duration models, it has no doubt that \( C_{WD} \) is the best, the next is \( C_{CD} \), and the least is \( C_{CD} \).

In the training phase of duration models, as shown in Figure 1, according to different situations (mainly the size of corpus), different contextual levels will be adopted automatically by using the hierarchical backoff mechanism. When there are enough samples for a certain word, \( C_{WD} \) will be preferred for this word immediately. And when the samples for each word are decreasing, and the phonemes are with a relatively balanced phonetic distribution, \( C_{WD} \) will back off to \( C_{CD} \). As for encountering a small size of corpus with more infrequent or unseen events of phone contexts, there is no alternative but to choose \( C_{CD} \), that is, \( C_{CD} \) will further back off to \( C_{CD} \). In order to reduce the complexity of the backoff operation, we use the forced backoff skills as described in [14]. The counts below which the duration models are backed off to lower contextual levels are referred to as our backoff thresholds.

In order to deal with the data sparsity problem, decision tree based clustering was used decades ago for acoustic modeling, which is not adopted here. The main reason is to simplify the problem, because the duration is 1-dimensional feature. In addition, we have accumulatd lots of speech data for many key words, it seems very natural to use the word-dependent strategy.

![Figure 1: Backoff mechanism for duration models.](image)

### 2.2. Parametric function vs. discrete distribution for duration modeling

As indicated in the introduction section, parametric modeling of durations with a certain pdf is favor for small corpus. The reason is that the parametric functions only have a few parameters to be tuned, which could reduce the amount of training data and improve the robustness of estimation [10]. The empirical distributions of phone durations in our training corpus are shown in Figure 2, which are obtained after performing a forced alignment of the speech signals and the orthographic transcripts via Viterbi decoding. Here /ah/ and /p/ are regarded as the representatives of vowels and consonants, respectively.

It can be seen that the distributions are mostly unimodal. It is also apparent that the phone durations could be helpful to differentiate the phones, which is the foundation for mispronunciation detection task. However, the distribution shapes of vowels and consonants have obvious difference for each other, that is, the shape of the majority of vowels is similar to Gaussian, while the shape of some consonants (especially for plosives) is more like exponential distribution which is a special case of Gamma distribution.

Though it is more difficult to accurately model the duration with a pdf, several parametric distributions have been proposed in the past for modeling durations, and the main purpose of them is to investigate which is the most appropriate. In [5, 10],...
Gamma is believed to produce a high quality fit to the empirical distributions. And in [11], the author concluded that the Gaussian mixture pdf (GMM) is a better approximation. In fact, the fitting quality is truly not stable upon different distribution functions, which is affected by the essential characteristics of the phone durations and the different corpus or environment. That is, the unstable fitting quality is one of the main shortcomings for parametric duration modeling in practical applications.

Figure 2: Duration distributions for /ah/ and /p/.

Consequently, the phone durations are preferred to be modeled in discrete form when enough data is available to properly count the probability of each duration level. However, since the largest corpora available will have the bias in duration distributions, it is necessary to smooth the data in order to provide better estimates of the more infrequent events. These infrequent events are always reflected as some distinct ripples in the distribution curve as shown in Figure 2. In most cases, the probability estimate is biased high for frequent events and biased low for infrequent ones. So, to correct this bias, we redistribute some probability mass from the frequent events to infrequent ones, using discounting operation by a discount coefficient $d$. Here, we adopt the linear discounting form [13] in our work, and the modified duration probability is gained as follow,

$$ p^* = p \times d $$

After discounting scheme, the empirical distribution would possess a relatively smooth contour (see Figure 2), which is favorable for further discrete modeling.

Another problem is that the discrete duration modeling has a large number of parameters than that of parametric functions. With the aim to reduce the model size and raise the computation speed, we try to construct a hash table for saving all the discrete probability values, and then get the corresponding probability of given phone by directly looking up the table in the testing phase. Take $C_{\text{WO}}$ for example, the address of the probability value can be obtained by a predefined hash function as follow,

$$ \text{addr} = \text{hash}(w, p, t) = \text{addr}0 + \sum_{i=1}^{\infty} T \times p \times T^i $$

where $\text{addr}0$ stands for the start address of the hash table, $w, p, t = 0, 1, 2, \ldots$, $w$ is the index of the word, $p$ is the index of the phone, $t$ is the duration level. With the address, the probability value is also found correspondingly. In the following section of experiments, the effectiveness of the proposed discrete duration modeling based on discounting and look-up tables will be illustrated in detail.

3. Experiments

3.1. Experimental setup

In our work, a high-volume database of non-native English speech recordings is constructed from lots of middle schools in the city of Shanghai and Jiangsu province of China under the project of Spoken English Testing for Chinese Students (SETCS). Here, we separate the database into four parts, as shown in Table 1. For TRAIN-AM set, the contents of recordings are composed of words, phrases, sentences and paragraphs, which are all carefully designed by considering the phonetic balance. For TRAIN-DM, TEST-PA and TEST-MD, the students are required to read a number of prompting paragraphs which are easy to understand for middle school students, and each reading paragraph has about 100 words. Besides, due to the model training phase, the samples in TRAIN-AM and TRAIN-DM sets are almost with the good pronunciation quality by using a front-end filtering mechanism.

The TEST-PA and TEST-MD sets are separately annotated by five experts. For TEST-PA, each recording sample is manually scored on discrete ratings as the overall assessment of the pronunciation quality, including four evaluation levels of good, medium, bad and flunk. For TEST-MD, the words of each recording sample’s manuscripts are annotated with tags for mispronunciation tokens, but the concrete erroneous forms are not further provided. When a word token is extremely ambiguous to judge its pronunciation correctness, a majority choice will be performed.

In the experiments, the acoustic HMM models used in forced alignment are trained with TRAIN-AM set. The duration models proposed in this work are all trained with TRAIN-DM set. The performances of different duration models for pronunciation assessment and mispronunciation detection are evaluated separately, using the correlation coefficient (CC) measure for the former, and the equal error rate (EER) measure for the latter.

Table 1. The construction of the used database.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAIN-AM</td>
<td>120 hours’ non-native speech, used for AM training and adaptation.</td>
</tr>
<tr>
<td>TRAIN-DM</td>
<td>16463 better samples selected from a high-volume database of 800,000 recordings.</td>
</tr>
<tr>
<td>TEST-PA</td>
<td>1314 samples with four discrete ratings of pronunciation quality by five annotators.</td>
</tr>
<tr>
<td>TEST-MD</td>
<td>595 samples with mispronunciation tags on word level by five annotators.</td>
</tr>
</tbody>
</table>

3.2. Experimental results

In our work, some acoustic features are used, such as LLR, GOP [6], GMM, SVM and NN scores [7, 8]. As for fluency features, ROS and ART [6] are also used. In Table 2, for the pronunciation assessment of oral readings, we give the correlation coefficients of various pronunciation features with the manual ratings. It can be see that the contextual strategy of duration is obvious in performance improvement, and the CC of DUR-WD can achieve about 0.7034. The DUR-CD and DUR-WD can also make a contribution to the fusion scheme. Of course, the acoustic features are still the dominating elements for pronunciation quality.

Figure 3 illustrates the DET curves of different duration models for mispronunciation detection task, and the detail results of EER are listed in Table 3. We can see that the DUR-CD and DUR-WD can both decrease the EER value significantly compared to DUR-CL. However, in term of combination with the acoustic scores, the absolute EER reduction of the best fusion case is only about one percent compared to the acoustic scores. This means, even though there
is a certain complementary between the acoustic and duration features, that the effect of duration scores is limited in some sense, especially for the direct decision of pronunciation correctness of a single phone or word.

Table 2. The CC performance in TEST-PA.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic</td>
<td>LLR+GOP+GMM+ SVM+NN</td>
<td>0.7807</td>
</tr>
<tr>
<td>Flueney</td>
<td>ROS+ART</td>
<td>0.5375</td>
</tr>
<tr>
<td>Duration</td>
<td>DUR-CI</td>
<td>0.6252</td>
</tr>
<tr>
<td></td>
<td>DUR-CD</td>
<td>0.6453</td>
</tr>
<tr>
<td></td>
<td>DUR-WD</td>
<td>0.7034</td>
</tr>
<tr>
<td>Acoustic + Flueney + DUR-CI</td>
<td>0.8122</td>
<td></td>
</tr>
<tr>
<td>Acoustic + Flueney + DUR-CD</td>
<td>0.8243</td>
<td></td>
</tr>
<tr>
<td>Acoustic + Flueney + DUR-WD</td>
<td>0.8389</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. The EER performance in TEST-MD.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic</td>
<td>LLR+GOP+GMM+ SVM+NN</td>
<td>28.96</td>
</tr>
<tr>
<td>Duration</td>
<td>DUR-CI</td>
<td>38.13</td>
</tr>
<tr>
<td></td>
<td>DUR-CD</td>
<td>35.47</td>
</tr>
<tr>
<td></td>
<td>DUR-WD</td>
<td>33.55</td>
</tr>
<tr>
<td>Acoustic + DUR-CI</td>
<td>28.54</td>
<td></td>
</tr>
<tr>
<td>Acoustic + DUR-CD</td>
<td>28.23</td>
<td></td>
</tr>
<tr>
<td>Acoustic + DUR-WD</td>
<td>28.01</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: DET plot for different duration models in TEST-MD.

Table 4. The performance of different fitting methods for word-dependent duration empirical distributions (CC in TEST-PA, EER in TEST-MD, X and S are real-time ratio and model storage respectively for computation efficiency).

<table>
<thead>
<tr>
<th>Method</th>
<th>Performance</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>EER (%)</td>
</tr>
<tr>
<td>Exponential fitting</td>
<td>0.6923</td>
<td>34.67</td>
</tr>
<tr>
<td>Gaussian fitting with two mixtures</td>
<td>0.7001</td>
<td>33.98</td>
</tr>
<tr>
<td>Discrete modeling with look-up tables</td>
<td>0.7034</td>
<td>33.55</td>
</tr>
</tbody>
</table>

In order to explain the necessity of the discrete modeling for duration empirical distributions and the corresponding optimization strategy based on the look-up tables, Table 4 shows the performance comparison from various angles. In

term of the performance, the discrete modeling is a little superior to the exponential and Gaussian fitting. And, in term of computation efficiency, the discrete duration modeling is much faster than others by using the look-up tables. But the discrete duration modeling also has a relatively large model space due to the non-parametric operation. However, at present, the size of only 0.720 MB will have no influence on our practical applications.

4. Conclusions

We can draw some conclusions. Firstly, the backoff mechanism can effectively resolve the problem of data sparseness or unbalance in the training phase of duration models. Secondly, the discrete modeling of empirical distributions has improved the precision of duration models, and the strategy of look-up tables has accelerated the computation speed. Thirdly, the word-dependent duration scores play an important role for pronunciation assessment and mispronunciation detection tasks, against the background of oral reading testing for Chinese students.

5. Acknowledgements

This work is partially supported by the National Natural Science Foundation of P. R. China under Grant 90820303.

6. References