On Use of Duration Modeling for Continuous Digits Speech Recognition

Rong Dong, Jie Zhu

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
jiezhu@guomai.sh.cn

ABSTRACT

In this paper, we describe our duration model techniques in HMM based speech recognizer. With this approach, a large amount of deletion and insertion errors can be reduced in Mandarin continuous digits recognizer. We address a simple duration penalty function, which can be explicitly combined into Viterbi-Beam search with negligible incremental computation overload. Different parametric distributions are investigated to accurately approximate the syllable-level duration information. A relative Rate of Speech (ROS) based duration normalization scheme is proposed to eliminate variation caused by different speaking rate. In order to directly incorporate this normalization strategy, an online dynamic ROS estimation method is introduced into real-time recognition application. Experimental results demonstrated significant performance improvement has been achieved. The word error rate (WER) was reduced 52.1%, compared with our baseline recognition system.

1. INTRODUCTION

Continuous digits recognition plays a vital role in most speech recognition applications due to its low perplexity and implied application. In addition, this kind of task usually demands extremely high accuracy, reliability and focuses on the acoustic match rather than language level processing.

Examining our preliminary experimental results in a Mandarin continuous digits recognizer, we have found that the only three monophonemic digits (“1” [yi], “2” [er] and “5” [wu]) comprise most of the insertion and deletion errors. Totally, 53.56% errors stem from the following three reasons:

1) Fail to separate repeatedly pronounced “1”, “2” or “5”, e.g., “11” was recognized as “1”.
2) One digit is sometimes falsely recognized as two segments, e.g., “2” was recognized as “22”.
3) Digits “1” and “5” also tend to be obscured in context of other digits, such as in “65” [liu wu], “95” [jiu wu], “71” [qi yi], etc.

The similar phenomena are frequently observed when English digit “oh” or “eight” is uttered [1]. These problems also exist in Korean digits recognition system [2]. These errors cannot be corrected even by intensive discriminative training due to the inherent unrealistic duration assumption in conventional HMMs.

Instead, a good syllable-based duration model can effectively eliminate those excessively long or short syllables occur in the partial hypothesis and manipulate a high performance continuous digits recognizer.

So far, considerable related works have been done in English speech system. Bounded Uniform and a variety of exponential family probability distributions, such as Gaussian, Gamma and Poisson distributions [1][3][4][5], are suggested to model English duration information. Various mechanisms have been developed to incorporate duration modeling in an HMM-based recognizer. Duration information can be explicitly modeled [1][5] or employed in an indirect manner [6]. It was even used in post-processing stage of speech recognition. One such example is reported in [1].

Considering the potential implementation perplexity and the expected recognition accuracy, we have opted for an explicit framework to impose duration constraints directly into Viterbi-Beam search module.

For Chinese Mandarin, recent duration study is mainly made on Gaussian hypothesis. One objective of this work is to analyze three alternative probability density functions and find out an optimal distribution model to capture the variation in syllable duration of Mandarin digits.

Still, this kind of duration is not robust enough to handle with various speaking-rate utterances. Using the relative Rate of Speaking (ROS) to normalize the observed duration is a solution to overcome this problem.

The rest of this paper is organized as follows. Section 2 describes how to integrate duration model into Viterbi-Beam decoding algorithm. In section 3, we focus on three parametric distributions to accurately model duration of Mandarin digit. Duration normalization for different speech rate are discussed in Section 4. A set of experiments and evaluations are given in Section 5.

2. EXPLICIT DURATION MODELING

2.1. Duration Model and Its Explicit Implementation

One generic limitation associated with the standard HMMs is that they implicitly model the probability of state occupancy by Geometric density as shown in Eq. (1). The probability of \( \tau_{ij} \), consecutive observations in state \( i \) can be written as:

\[
\rho_1(\tau_{ij}) = a_{ij} = (1 - a_{ii})
\]

(1)

where \( a_{ii} \) is the self-loop transition probability.

This underlying assumption is usually inappropriate in most of the case. An explicit time duration distribution for each state or model is particularly motivated to improve this issue. In this work, we only discuss duration modeling for syllable models rather than internal HMM states.

By taking logarithms of the model parameters, the conventional Viterbi score \( \delta_1(j) \), when the observation \( O_t \) occurred in state \( j \) of model \( s \), can be defined recursively:

\[
\delta_1(j) = \max_{i \in N} \left[ \delta_{t-1}(i) + \log(a_{ij}) + b_{j}(o_t) \right]
\]

(2)
This frame-synchronous Viterbi algorithm can be easily modified to incorporate explicit duration model as an additional penalty for accumulated score of the path at the end of each model, when \( j \) is the exiting state of model \( s \). It can be rewritten as follows

\[
\bar{\delta}_s(j) = \text{Max}_{1 \leq c \leq N} \{ w \log P_s(\tau_{s,i}) + \bar{\delta}_s(i) + \log(a_{ij}) + b_j(o_j) \} \tag{3}
\]

where \( \tau_{s,i} \) is duration of model \( s \) at time \( t \). \( P_s(\tau_{s,i}) \) denotes the expected duration penalty. The weighting factor \( w \) is used to control the duration contribution to the total Viterbi score.

### 2.2. Penalty Function in Viterbi-Beam Search

Penalty function \( P_s(\tau_{s,i}) \) in Eq. (3) represents duration probability when exits from model \( s \) at time \( t \). Therefore, it can be directly defined as duration p.d.f. associated with model \( s \). In the Bounded Uniform case,

\[
P_s(\tau_{s,i}) = \begin{cases} D_{\text{min}} \leq \tau_{s,i} \leq D_{\text{max}} & 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}
\]

\( D_{\text{min}} \) and \( D_{\text{max}} \) are the minimum and maximum corresponding duration observed on the training data. \( P_s(\tau_{s,i}) = 0 \) results in an impossible hypothesis which will be removed from the remaining search space. Note that, this form of model only detects extreme unlikely duration. More accurate duration penalty can be introduced by sophisticated parametric distribution such as Gaussian, Gamma or Poisson density function which will be discussed in Section 3.

However, allowing for the beam pruning strategy during search, we find it is unreasonable to apply this duration score merely once at the end of model. As mentioned in Burstein’s implementation [4], it should be gradually added at each frame. Unfortunately, in his work, the duration penalty only takes effect after a maximal threshold. That is, only unrealistic long duration observed on the training data. \( P_s(\tau_{s,i}) \) is duration of model \( s \) at time \( t \).

In the very beginning, this penalty score is a positive value added in each frame, therefore the longer occupancy is favored in this stage. In contrast, upon the downward density curve, the longer occupancy, the more negative penalty is accumulated in the local score, it means the more likely to be pruned. At the end of model, \( P_s(0) \) is compensated to keep the overall duration penalty equals to \( P_s(\text{overall-duration}) \).

In practice, observation output probability is so dominant in comparison with the transition probability and duration penalty score. A thoroughly tuned weight \( w \) is thus rather important to make the duration modeling take effect in recognition. Recognition performance is sensitive to the choice of this weight.

### 3. Parametric Distribution

Since a variety of exponential form distributions have been proposed in the previous studies of duration modeling [1][3][4][5], we found it worthy to investigate which is the best approximation for Mandarin digit syllable duration. Here, three of the most popular parametric distributions, Gaussian, Gamma and Poisson are evaluated. Boundaries of digit words for each utterance in training set are automatically segmented by high reliable Viterbi force-alignment. The histogram of the duration is collected for each digit model. From the observed random sample \{x1, x2,..., xn\}, free parameters for each type of p.d.f. can be estimated with the standard maximum-likelihood (ML) criterion.

Figure 1 displays the empirical duration distribution for digit “5” ([wu]) along with its Gaussian, Gamma and Poisson fitting estimated under ML criterion.

From the visual presentation of curves in Figure 1, Gamma fit exhibits ability superior to Gaussian and Poisson in this case. The same conclusion is validated in other pronunciation of Mandarin digits.

Distortion quantity between the empirical distribution and the given p.d.f. are measured in terms of mean squared error (MSE). These values listed in Table 1 are averaged over all the syllable models in our recognition system.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Gaussian</th>
<th>Gamma</th>
<th>Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE (10^-5)</td>
<td>4.894</td>
<td>3.476</td>
<td>8.323</td>
</tr>
</tbody>
</table>

Table 1: Average fitting distortion

Moreover, it is supported by our experiments results which are shown in Section 5.

### 4. Duration Normalization

#### 4.1. Speaking Rate and Normalized Duration

There is strong evidence that speaking rate is a source of variation which can lead to significant increase in error rate in speaker independent speech recognition [8]. The variance of the distribution we measure directly by the empirical histogram as

![Figure 1: Modeled distribution contours for duration of digit “5” ([wu]).](image)
mentioned in Section 3 is too large to precisely characterize the duration of each syllable. Therefore we introduce the idea of duration normalization to enhance the robustness of speech recognition against the relative Rate of Speaking (ROS) in each utterance.

In our digits recognizer application, it is generally reasonable to assume that the speaking-rate is almost stable throughout a sentence (a string of digits). In light of this, an utterance-wise duration normalization strategy is adopted here. A simple “syllables-per-second” definition of ROS measurement is, for sentence $k$,

$$\text{ros}(k) = \frac{\text{Number Syllable}}{\text{Duration} - \text{Silence}}$$

Figure 2 is a plot of sentence level ROS in training set. The mean value is 3.7739 and the standard deviation is 0.7153.

$$\text{Figure 2: Histogram for sentence level ROS in training set.}$$

An average ROS is measured over all of the utterances in the training set and is denoted as $\mu_{\text{ros}}$, consequently the relative ROS denoted as $\hat{\text{ros}}(k)$ can be calculated

$$\hat{\text{ros}}(k) = \frac{\text{ros}(k)}{\mu_{\text{ros}}}$$

Fast speech tends to shorter syllable length, therefore expected duration should be inverse with the ROS. In order to compensate for the effect of ROS, the real duration $\tau_{i,j}^k$ of the $i$-th syllable in the $k$-th utterance should be scaled to $\hat{\tau}_{i,j}^k$, which is defined as follows

$$\hat{\tau}_{i,j}^k = \tau_{i,j}^k \times \hat{\text{ros}}(k)$$

The remaining process is the same as outlined above except that all the duration $\tau_{i,j}^k$ in training and recognition is substituted by the normalized duration $\hat{\tau}_{i,j}^k$.

With normalization process, the standard deviation of digit “5” duration is reduced from $8.676 \times 10^{-2}$s to $7.495 \times 10^{-2}$s. Relatively 13.61% reduction is achieved.

A new set of MLE for normalized duration distribution is given in Figure 3. Again, the results of MSE test are listed in Table 2.

$$\text{Figure 3: Modeled distribution contours for normalized duration of digit “5” ([wu]).}$$

$$\begin{array}{|c|c|c|c|}
\hline
\text{Model Type} & \text{Gaussian} & \text{Gamma} & \text{Poisson} \\
\hline
\text{MSE (10^{-5})} & 5.330 & 2.140 & 3.753 \\
\hline
\end{array}$$

Table 2: Average fitting distortion

It should be emphasized that the normalized duration distribution is approximated quite well with the Gamma p.d.f except the tail region.

4.2. Online Speech Rate Estimator

In real-time processing, the identity and segmentation of each model is actually unknown in advance, therefore the exact ROS value for the current test utterance is not available before the end of recognition. However it can be estimated during Viterbi search in a way called partial trace back. Using the available partial ancestor shared by all the active hypothesis, ROS can be updated with new arrival recognition results.

5. EXPERIMENTAL EVALUATION AND RESULTS

Recognition experiments are performed in continuous Mandarin digits tasks. The speech materials used in these works are collected from 63 male speakers in a quiet room. Each speaker pronounced 100 digit strings with length varying from 1 to 16.

Data from 42 speakers are used for training and the other 20 speakers as test set. The feature vector for this 8kHz task consists of 12 MFCCs with energy and corresponding delta and acceleration values.

The recognition vocabulary includes “0” to “9”, “A” [yao] (an alternative Mandarin pronunciation of “1”) and silence. We select a simple model proto consisting of 6 states, left-to-right continuous HMM for each digit. An unconstrained grammar allows any digit sequence with optional silence between digits. Expressed in HTK format, the grammar is “sil <$\text{digit}$ | sil”.

5.1. Baseline Vs. HTK reference
Our baseline system is compared with the corresponding HTK reference recognizer under the same data sets and the same configuration. It appears that our baseline recognizer moderately outperforms HTK counterpart.

<table>
<thead>
<tr>
<th>System</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no duration modeling)</td>
<td>94.78%</td>
</tr>
<tr>
<td>HTK reference</td>
<td>94.10%</td>
</tr>
</tbody>
</table>

Table 3: Recognition performance in terms of word accuracy on baseline system and HTK counterpart

5.2. Bounded, Gaussian, Gamma and Poisson Form Duration Modeling

In this experiment, we summarize the improvement under different duration parametric model. The four representative distribution forms are examined in our recognizer. For this step, we use absolute syllable duration both in training and testing procedure.

<table>
<thead>
<tr>
<th>Model</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no duration modeling)</td>
<td>94.78%</td>
</tr>
<tr>
<td>+ Bounded Duration Model</td>
<td>94.98%</td>
</tr>
<tr>
<td>+ Gaussian Duration Model</td>
<td>96.63%</td>
</tr>
<tr>
<td>+ Gamma Duration Model</td>
<td>97.23%</td>
</tr>
<tr>
<td>+ Poisson Duration Model</td>
<td>96.02%</td>
</tr>
</tbody>
</table>

Table 4: Best recognition performance in terms of word accuracy with different duration parametric model

Table 4 demonstrates that Bounded Duration Model does not hurt nor bring much contribution to the overall recognition performance. This model assumption is admittedly too loose to detect those unreasonable long or short syllable occupancy during recognition. As indicated in Section 3, Gamma fit is much closer to the real distribution and accordingly leads to more desirable results in our application.

5.3. Normalized Duration

Table 5 illustrates the further improvement in word accuracy due to utterance-wise ROS normalization. Gamma method offers even higher recognition rate. Gaussian and Poisson methods share similar result in this experiment.

<table>
<thead>
<tr>
<th>Model</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no duration modeling)</td>
<td>94.78%</td>
</tr>
<tr>
<td>+ Bounded Duration Model (normalized w/ exact ROS)</td>
<td>94.98%</td>
</tr>
<tr>
<td>+ Gaussian Duration Model (normalized w/ exact ROS)</td>
<td>96.93%</td>
</tr>
<tr>
<td>+ Gamma Duration Model (normalized w/ exact ROS)</td>
<td>97.69%</td>
</tr>
<tr>
<td>+ Poisson Duration Model (normalized w/ exact ROS)</td>
<td>96.98%</td>
</tr>
</tbody>
</table>

Table 5: Recognition performance in terms of word accuracy for explicit normalized duration modeling with exact ROS

Using online ROS estimation strategy, recognition performance is somewhat degraded.

<table>
<thead>
<tr>
<th>Model</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Bounded Duration Model (normalized w/ estimated ROS)</td>
<td>94.98%</td>
</tr>
<tr>
<td>+ Gaussian Duration Model (normalized w/ estimated ROS)</td>
<td>96.77%</td>
</tr>
<tr>
<td>+ Gamma Duration Model (normalized w/ estimated ROS)</td>
<td>97.50%</td>
</tr>
<tr>
<td>+ Poisson Duration Model (normalized w/ estimated ROS)</td>
<td>96.62%</td>
</tr>
</tbody>
</table>

Table 6: Recognition performance in terms of word accuracy for explicit normalized duration modeling with estimated ROS

6. CONCLUSIONS

In conclusion, this work yields a practical syllable-level duration modeling strategy for Mandarin continuous digits recognizer. A feasible penalty function is addressed which can be explicitly and efficiently added in Viterbi score. Four duration models are investigated. Among all of these models, Gamma distribution is verified to optimally describe duration information for Mandarin digits rather than the commonly used Bounded Uniform, Gaussian and Poisson assumption. It is also proved that accurate modeling guarantees much gain in recognition performance. ROS normalization also generates further improvement in our recognizer.

7. REFERENCES