Explicit Duration Modeling for Cantonese Connected-digit Recognition

Yu Zhu and Tan Lee
Department of Electronic Engineering
The Chinese University of Hong Kong
{yzhu,tanlee}@ee.cuhk.edu.hk

Abstract
This paper describes a study on using explicit duration models in hidden Markov model (HMM) based Cantonese connected-digit recognition. An HMM does not give explicit control to the temporal structure of speech. As a result, the recognition output may exhibit unreasonable duration pattern, which is often accompanied with the presence of recognition errors. We propose to use a duration model that models the relative duration of the tail part of a Cantonese digit, together with conventional word-level duration models. The duration models are integrated into the Viterbi search algorithm for speech recognition. Experimental results show that proposed method leads to substantial reduction of recognition errors, especially for slowly spoken utterances.

1. Introduction
HMM is the most predominant technique for automatic speech recognition (ASR). It models speech signals as two concurrent stochastic processes. While the spectral variation is captured by probability density functions at individual states, the duration of a speech segment is modeled implicitly by the state transition probabilities. An HMM does not give explicit control to the temporal structure of speech. As a result, the recognition output may exhibit unreasonable (in terms of probability) duration pattern, which is often accompanied with the presence of recognition errors. Short sounds with simple phonetic composition tend to be inserted or deleted. Substitution errors may occur if the discrimination between certain sounds relies heavily on their duration. To alleviate this problem, the use of explicit duration models has been proposed to constrain duration variation and preclude extreme cases. In [1][2], it was proposed to replace the self-transition probability of HMM with explicit state duration distribution. In [3-6], the domain of duration modeling was extended to supra-segmental unit, e.g. syllable, word. These methods showed significant improvement on speech recognition performance.

In this paper, we focus on explicit duration modeling for Cantonese connected-digit recognition. Connected-digit recognition has many practical applications that often require very high accuracy. Despite its limited vocabulary size, it is not straightforward to attain the desired accuracy mainly because the combination of digits is unrestricted.

Cantonese is a major Chinese dialect spoken in Southern China and Hong Kong. It is a monosyllabic language. As shown in Table 1, each Cantonese digit is pronounced as a monosyllable sound. The syllable compositions are generally very simple. In particular, digits “2” and “5” can be regarded as a single vowel segment and a single nasal segment respectively. In speech recognition, these mono-phone digits tend to be inserted between themselves, e.g. “2” may be recognized as “22” or vice versa.

Table 1: Phonetic transcriptions of the 10 Cantonese digits

<table>
<thead>
<tr>
<th>digit</th>
<th>IPA</th>
<th>LSHK</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ling4</td>
<td>lin⁴</td>
</tr>
<tr>
<td>1</td>
<td>jat1</td>
<td>ja¹</td>
</tr>
<tr>
<td>2</td>
<td>ji6</td>
<td>ji⁶</td>
</tr>
<tr>
<td>3</td>
<td>saam1</td>
<td>saam¹</td>
</tr>
<tr>
<td>4</td>
<td>sei3</td>
<td>se³</td>
</tr>
<tr>
<td>5</td>
<td>ng5</td>
<td>ng⁵</td>
</tr>
<tr>
<td>6</td>
<td>luk6</td>
<td>luk⁶</td>
</tr>
<tr>
<td>7</td>
<td>cat1</td>
<td>c²</td>
</tr>
<tr>
<td>8</td>
<td>baat3</td>
<td>baat³</td>
</tr>
<tr>
<td>9</td>
<td>gau2</td>
<td>ga²</td>
</tr>
</tbody>
</table>

On the other hand, the two mono-phone digits are phonetically very similar to the tail part of other digits, e.g. “0”, “3” and “4”. It is likely, for example, to recognize “4” as “42”, i.e. a “2” is inserted. Table 2 shows the baseline recognition performance of an HMM based Cantonese connected-digit recognition system. Among the recognition errors, insertion and deletion account for over 80%. About 70% of deletion errors and 70% of the insertion errors are caused by the two mono-phone digits.

Table 2: Baseline performance of Cantonese connected-digit recognition. The number in brackets refers to those caused by mono-phone digits

<table>
<thead>
<tr>
<th>Word acc</th>
<th>Sent acc</th>
<th>Deletion</th>
<th>Substitution</th>
<th>Insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.47</td>
<td>88.23</td>
<td>80(58)</td>
<td>57</td>
<td>230(158)</td>
</tr>
</tbody>
</table>

We propose to use absolute duration models and relative duration models for Cantonese connected-digit recognition. Specifically, the relative duration of the tail part of a Cantonese digit is found to be an important and useful feature. It is observed that many recognition errors have unreasonable tail part duration, and HMM give very loose constraint to it. Consequently, a specific relative duration model is established to model the variation of this feature.

Duration depends largely on the speaking rate. Duration models are established separately for three different speaking-rate categories, namely “fast”, “normal” and “slow”.

For speech recognition, HMM states form a search space from which the optimal path would be determined by the Viterbi search algorithm. Duration model is integrated by adding an additional probabilistic score to the conventional
path score. In this paper, it is integrated by applying duration penalties at each model end.

2. HMM-based Cantonese connected-digit recognition

An HMM models a speech utterance as two concurrent stochastic processes. As shown in Figure 1, for speech recognition, the HMM has a number of states that are arranged in a left-to-right topology. Temporally, the transition between states is governed by the transition probabilities $a_{ij}$. Spectrally, each HMM state is described by a probability density function $b_j(o_t)$, where $o_t$ is the input feature vector of a particular time frame.

![Figure 1: A hidden Markov model for speech recognition](image)

For Cantonese connected-digit recognition discussed in this paper, 10 HMMs are trained to represent the digits “0” – “9” respectively. Each HMM has 6 states and the probabilistic density function of each state is represented as a mixture of Gaussian distributions.

The HMM parameters are obtained by training with a large amount of speech data. For recognition, the HMM states form a search space that essentially covers all possible recognition outcomes. Given an input feature sequence $\{o_1, o_2, ..., o_T\}$, the optimal path in this search space is determined by the Viterbi decoding algorithm. Details of Viterbi decoder will be given in Section 4.

3. Design of duration models

3.1. Absolute duration model

In an HMM-based system, duration can be measured and modeled at state level or model level [1-6]. In an HMM, state transition probabilities implicitly model the state duration, and this leads to a Geometric distribution. Figure 2 shows the duration distribution of state 3 of the HMM for digit “3”. The empirical distribution is obtained by supervised segmentation of training data with the HMM, while the implicit distribution is computed from the transition matrix of the HMM. Obviously, the actual state duration derived from empirical data does not follow a Geometric distribution. In this work, we choose to use Gamma distribution to model absolute state duration.

On the other hand, in our application, model-level duration directly corresponds to the duration of a Cantonese digit. By applying model-level duration models, digit sequence with unreasonable digit duration would be penalized. In our design, bounded Gamma distribution is used for model-level duration modeling.

3.2. Relative duration model

Modeling of the absolute duration of a speech segment does not reflect possible internal adjustment between the sub-components of this segment. In other words, there are cases that the absolute model duration is reasonable but the duration occupied by its individual states are not in appropriate proportion. In [4], state duration normalized with respect to model duration was used to address this problem. It showed complementary effectiveness as compared with the use of absolute model-level duration.

We propose to use the tail part ratio for Cantonese connected-digit recognition, which measures the relative duration of the tail part of a digit. The tail part is defined to cover the last two states of the 6-state HMM. The tail part ratio can be considered as a variation of normalized state duration. The tail part defined above roughly corresponds to the last phonetic unit in a Cantonese digit. As mentioned earlier, the two mono-phone digits are very similar to the tail part of another three digits. If this tail part happens to be deleted or prolonged, the tail part ratio would be unreasonable. In this case, the tail part ratio model would be helpful.

![Figure 2: State duration distribution for state 3 of the HMM for Cantonese digit “3”](image)

![Figure 3: Tail part ratio distribution of the HMM for Cantonese digit “3”](image)
Figure 3 plots the empirical distribution of the tail part ratio for digit “3” and the distribution implicitly modeled in the respective HMM. It can be seen that the implicit duration distribution has a much larger variance than the empirical observation. This implies that unreasonable tail part ratios are usually allowed.

In summary, we establish the following duration models for Cantonese connected-digit recognition:
1) AM: absolute model-level duration model, using bounded Gamma distribution
2) AS: absolute state duration model, using Gamma distribution.
3) NS: normalized state duration model, using bounded Gamma distribution.
4) TP: tail part ratio model, using bounded gamma distribution model.

All the above duration models are digit-dependent, considering that different digits have different phonetic compositions.

3.3. Speaking-rate dependent model
Speech data are divided into three categories of different speaking rate: “fast”, “normal” and “slow”. Duration models are built separately for these categories. For speech recognition, the speaking rate category of a test utterance is first identified based on the number of digits per second, which can be calculated from the baseline recognition output. Subsequently, corresponding duration models are chosen to be used in the Viterbi decoding.

4. Integrating duration models into Viterbi decoder
For speech recognition, HMMs that represent different speech units are connected according to grammatical rules. The HMM states then form a search space from which the optimal path would be determined by the Viterbi search algorithm. In the search process, the probabilistic score of a path is contributed by $b_j(o_j)$ and $a_{v(i)}$. An additional score generated by duration models can be added to the path score. If the recognized speech units in this path exhibit unreasonable duration patterns, the duration score may be used to penalize the path. Penalties can be applied to the N-best recognition output at post-processing stage [7]. That is, duration score is used at the end of the utterance. Alternatively it can also be applied to word end [4] or even at state transitions [3] [6]. In this work, we choose to apply duration score at the model end during path extension in the forward Viterbi search.

Let $(t,v,j)$ denote a node in the search space, where $t$ is the frame index, $v$ indexes the HMM, $j$ indexes the state in this HMM. Each node is described by the accumulated path score $S$, the identity of the previous node and the relevant duration information. At a model-end node, the absolute model duration, the absolute and relative duration of each state in the current model, the relative duration of tail part are obtained.

If duration score is not incorporated, the intra-model and inter-model path extensions are governed respectively by the following recursive equations,

$$S(t,v,j) = \max_{1 \leq i \leq N_t} \{S(t-1,v,i) \times a_{v(i)} \times b_j(o_j)\} \quad (2-1)$$

$$S(t,v,1) = \max_{1 \leq i \leq N_t} \{S(t-1,u,M_i) \times a_{v(i)} \times b_j(o_j)\} \quad (2-2)$$

where $u$ is the index of the previous model, $M_i$ is the number of states in model $u$. The introduction of duration score is applied to the model end, therefore, the inter-model recursion equation is modified to,

$$S(t,v,1) = \max_{1 \leq i \leq N_t} \{S(t-1,u,M_i) \times a_{v(i)} \times b_j(o_j) \times (D_u)\} \quad (2-3)$$

$D_u$ is the duration score contributed by one or more of the suggested duration models. For example, if we want to incorporate tail part ratio model, $D_u$ is equal to $P_d(r)$, where $P_d(r)$ is the probability that the tail part ratio of model $u$ has a value of $r$. Similarly $D_u = \prod P_d(d_i)$ if the state duration models are used, where $d_i$ is the state duration value, $P_d(d_i)$ is the probability that duration of state $s$ in model $u$ has a value of $d_i$. Weight factor $w$ is used to control the contribution of the duration score. Experimental observations reveal that spectral match score always has a much larger dynamic range than that of the duration score. Consequently, the duration modeling effect tends to be overshadowed by that of spectral modeling. $w$ can be used to emphasize the duration score and make it influential.

5. Experimental results and analysis
In this paper, all experiments on Cantonese connected-digit recognition are based on CUDIGIT, which was developed at the Digital Signal Processing Laboratory, the Chinese University of Hong Kong [8]. The speech data were sampled at 16 KHz with 16 bit resolution. Only male speakers’ utterances are used. These include 11,387 training utterances from 20 speakers and 2,847 test utterances from another 5 speakers. A total of 11 HMMs are trained for the 10 Cantonese digits and silence. The acoustic feature vector has 39 components, which include 12 Mel-Frequency Cepstral Coefficient (MFCC), log-energy, and their first and second order derivatives. The experiments are categorized into three groups as described below.

Group 1: Use of absolute duration models

Table 3 gives the recognition results on the use of absolute duration models, i.e. AM and AS as named in Section 3. Many trials of experiments are carried out with different weights on duration scores. The best weights for different speech-rate categories may be different. The results in Table 3 (and Table 4 and 5 as well) are the overall performance with the best weights used for each category.

It can be seen that insertion errors are reduced substantially from 230 (the baseline) to 58 when both model-level and state-level duration information are used. However, at the same time, more deletion errors are introduced (from 80 to 125). We examine a few cases of errors and found that the deleted digits have short duration, which tends to be penalized by the duration models. In particular, the number of deletions for the slow category is doubled because the duration modeling favors long-duration digits. The number of substitution errors decreases slightly. On the whole, the digit accuracy is improved from 96.47% to 97.77%, and the utterance accuracy is improved from 88.23% to 92.34%.
Deletion errors increase more or less at the same time. This implies that they are to certain extent complementary to each other. The performance of using the tail part duration is better than that of using normalized state duration, when they are combined with the absolute model duration.

### Table 3: Recognition results for Group 1

<table>
<thead>
<tr>
<th></th>
<th>Digit acc %</th>
<th>Utterance acc %</th>
<th>Del</th>
<th>Sub</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>96.47</td>
<td>88.23</td>
<td>80</td>
<td>57</td>
<td>230</td>
</tr>
<tr>
<td>AM</td>
<td>97.72</td>
<td>92.31</td>
<td>116</td>
<td>55</td>
<td>66</td>
</tr>
<tr>
<td>AS</td>
<td>97.47</td>
<td>91.46</td>
<td>100</td>
<td>51</td>
<td>112</td>
</tr>
<tr>
<td>AM+AS</td>
<td>97.77</td>
<td>92.34</td>
<td>125</td>
<td>49</td>
<td>58</td>
</tr>
</tbody>
</table>

Group 2: Use of relative duration models

The recognition results on the use of relative duration models i.e. NS and TP, are given in Table 4. It seems that the relative duration models are equally effective as the absolute duration models in terms of overall accuracy. However, the distributions of different types of errors are quite different between absolute and relative models. The relative duration models achieve a better balance between insertion and deletion than the absolute ones. This may be due to that relative duration is less likely to take extreme values. There is no noticeable difference between NS and TP.

### Table 4: Recognition results for Group 2

<table>
<thead>
<tr>
<th></th>
<th>Digit acc %</th>
<th>Utterance acc %</th>
<th>Del</th>
<th>Sub</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>96.47</td>
<td>88.23</td>
<td>80</td>
<td>57</td>
<td>230</td>
</tr>
<tr>
<td>NS</td>
<td>97.40</td>
<td>91.18</td>
<td>90</td>
<td>59</td>
<td>121</td>
</tr>
<tr>
<td>TP</td>
<td>97.34</td>
<td>91.22</td>
<td>89</td>
<td>58</td>
<td>129</td>
</tr>
</tbody>
</table>

Group 3: Use of both absolute and relative duration models

As shown in Table 5, when the absolute duration model and the relative duration model are used together, further improvement can be achieved as compared with the case of using either one of them. This implies that they are to certain extent complementary to each other. The performance of using the tail part duration is better than that of using normalized state duration, when they are combined with the absolute model duration.

### Table 5: Recognition results for Group 3

<table>
<thead>
<tr>
<th></th>
<th>Digit acc %</th>
<th>Utterance acc %</th>
<th>Del</th>
<th>Sub</th>
<th>ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>96.47</td>
<td>88.23</td>
<td>80</td>
<td>57</td>
<td>230</td>
</tr>
<tr>
<td>AM+NS</td>
<td>97.90</td>
<td>92.80</td>
<td>111</td>
<td>56</td>
<td>51</td>
</tr>
<tr>
<td>AM+TP</td>
<td>97.93</td>
<td>93.01</td>
<td>110</td>
<td>55</td>
<td>50</td>
</tr>
</tbody>
</table>

In all of the experiments, performance improvement is mainly contributed from the reduction of insertion errors. The deletion errors increase more or less at the same time.

The number of substitution errors doesn’t change too much. This is as expected because there are no digit pairs in Cantonese that heavily rely on duration features to distinguish each other.

The effectiveness of duration models varies for different speaking-rate categories. In particular, for the fast category, the best weight for duration score is found to be 0. In other words, the use of duration models tends to degrade the recognition performance for the fast category. On the other hand, as shown in Table 6, duration modeling gives a significant performance improvement for the slow category. The utterance accuracy is improved from 85.96% to 95.78%. The performance improvement for normal category is not as noticeable as that of the slow category.

### Table 6: Recognition results for the slow and normal category

<table>
<thead>
<tr>
<th></th>
<th>Digit acc %</th>
<th>Utterance acc %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>94.68</td>
<td>85.96</td>
</tr>
<tr>
<td>AM+TP</td>
<td>98.45</td>
<td>95.78</td>
</tr>
</tbody>
</table>

### 6. Conclusions

This work investigates the use of duration information in Cantonese connected-digit recognition. Explicit duration models have been built and integrated into the speech recognition process. Both absolute duration models and relative duration models are found to be effective in improving the recognition performance. The improvement is mainly due to the reduction of insertion errors. The combined use of absolute model-level duration and the tail part ratio attains the best recognition accuracy. Performance improvement is particularly significant for the slow speech.

### 7. References


