TRIGRAM DURATION MODELING IN SPEECH RECOGNITION*

Yun Tang, Wenju Liu and Bo Xu

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{tangyun, lwj}@nlpr.ia.ac.cn, xubo@hitic.ia.ac.cn

ABSTRACT

Rate of speech (ROS) is a very important factor in speech recognition. In this paper we present a new speech rate measure method which first normalizes the duration of different acoustic units to standard duration and then builds a trigram duration model to measure the speech rate of sentence. We propose two methods based on the standard duration to compensate the influence introduced by speech rate variation in data corpus and get 11% error rate reduction in mandarin digit string recognition.

1. INTRODUCTION

Acoustic model aims at finding the probabilistic behavior of the given observation sequence. However, most of acoustic models tend to decompose the segment of observation sequence into frame based unit and measure them in a dependent way due to the limitation of model complexity and parameter estimation, such as the state-based measure in HMM[2] or region-based measure in SSM[1]. These assumptions lose the information about the whole segment more or less. They are compromises between fine modeling ability and low complexity of model. A good way to compensate these weak points is to use the segmental level information, which can be defined as the effect of different combinations of energy, $F_0$, and duration produced on spoken language. These features highly relate to the whole segment and are useful complements for the traditional acoustic feature. Another advantage of these features is that they usually have simple structure or representation, so we can use more complex model to describe them without increasing too much load to the system. In this paper, we will focus on using duration information to detect the speech rate and improve the recognition results. The performance of a speech recognition system dramatically degrades when the speech rate is different from the average speech rate in training corpus. The speech rate is highly correlative to the duration sequence of acoustic unit (syllable, word etc.) in the sentence. However, it is not the case that duration in fast speech rate is simply shorter than that in slow speech rate even for the same acoustic unit [4]. Many speech phenomena, such as the acoustic unit identity of current segment and its surrounding segments, stress/unstressed property of segments, sentence/phrase prominence, within-phrase position, etc., had been proposed as factors that contribute to the variation of segmental duration [6]. Martinez F. et al [5] have presented a speech rate measure method based on duration of each phone. They first get the ratios of phone duration to the average duration and then measure speech rate by average performance of these ratios. N. Morgan et al [8] estimate ROS directly from the waveform or acoustic features of the input utterance. A new method is proposed to measure the speech rate in this paper, which builds context model on the duration sequence in order to reveal the variation of duration in the sequence. We first normalize the duration of different acoustic units to standard duration and then build a context duration model to measure the speech rate of sentence. Then we propose two methods based on the standard duration to compensate the influence introduced by speech rate variation in speech corpus. The organization of this paper is as follows. In the next section, we will introduce the experiments environment. Trigram model on duration will be presented in section 3. The speech rate classifier and compensation techniques will be proposed in next two sections. Finally, in section 6 we give our conclusions and future work in this area.

2. MANDARIN DIGIT STRING DATABASE

Our methods are verified on a Mandarin digit string recognition system. The Mandarin digit string database used for experiments includes 55 males and each one has 80 utterances. The length of each utterance varies from 1 to 7 digits and the average length is 4. The vocabulary is

*This work was supported in part by the China National Nature Science Foundation (No. 60172055) and the Beijing Nature Science Foundation (No. 4042025).
3. TRIGRAM DURATION MODEL

The N-gram models are successfully applied in language model, which can capture the constraints within the language by simply conditioning the probability of a word on a small fixed number of predecessors. We extend this property of Ngram to grasp the relation among durations in a sentence. Given three succeeded segments 123, where 1 is the first segment and 3 is the last segment, the correspondent durations are $l_1,l_2,l_3$ separately and their corresponding acoustic units are $a_1,a_2,a_3$; the duration probability of segment $(l_1,a_1)$ followed segment sequence $l_1,a_1,l_2,a_2$ is $P(l_1,a_1|l_1,a_1,l_2,a_2)$.

$$P(l_1,a_1|l_1,a_1,l_2,a_2) = \sum_{a_3} P(l_2,a_2|l_1,a_1,a_3,l_3)$$

(1)

where $\Omega$ is acoustic unit set, $L$ is the maximum duration. This idea is clear, however, we need to deal with a main problem: the sparseness of the data. In our digital recognition system, there are 11 digit models and a silence model. It needs to count every possible trigram duration probability appeared in utterance, so we should get $(12 \times L)^3$ duration probabilities for a trigram model. There are a lot of smoothing methods in language model, such as linear interpolation smoothing or back-off n-gram [3]. Comparing with the trigram in language model, there is an extra $L$ in this model, which is always beyond 30 for mandarin digit words. So these smoothing methods are still not enough for our case. A possible way is to normalize the duration of different digits to standard duration and eliminate the deviation introduced by different digits. We classify the duration of each digit segment sets evenly into $K$ standard duration levels by their real length. The standard duration transition probability can be computed as follow:

Transform the segment duration into standard duration:

$$T(l_1,a_1) \rightarrow \{R_{i1},...,R_{ik}\}$$

$R_i \in [0,1]$ is the ratio that duration $l$ mapping to standard duration $i$.

$$N(l_1,a_1) = \sum_{R_i} C(R_i,a_1)$$

(2)

where $C(R_i,a_1)$ is a function that counts the occurrences of $a_1$ with duration $l_1$ in training corpus and $N(l_1,a_1)$ is the occurrences of $a_1$ with duration no more than $l$.

$$L(l_1,a_1) = K \cdot N(l_1,a_1) / N(\infty,a_1)$$

(3)

$$R_i(l_1,a_1) = \begin{cases} 0, i > L(l_1,a_1) or i < L(l_1,a_1) or i = 0 \\ 1, i \in \{L(l_1,a_1),L(l_1,a_1)\} \\ 1 - R_i(l_1,a_1), i = L(l_1,a_1) \\ L(l_1,a_1) - L(l_1,a_1), i = L(l_1,a_1) \end{cases}$$

(4)

where $\lfloor x \rfloor$ is the largest integer $\leq x$.

$$R_i(l_1,a_1) = \sum_{R_j} R_j(l_1,a_1)$$

(5)

Compute the transform probability:

$$P(l_1,a_1|l_2,a_2,a_3) = \sum_{R_1} \sum_{R_2} \sum_{R_3} R_1(l_1,a_1) R_2(l_2,a_2) R_3(l_3,a_3) P(t \mid i, j)$$

where $P(t \mid i, j)$ is the probability of duration $t$ preceded by standard duration $f$.

$$P(t \mid i, j) = \sum_{l_1,a_1} \sum_{l_2,a_2} \sum_{l_3,a_3} R_1(l_1,a_1) \cdot R_2(l_2,a_2) \cdot R_3(l_3,a_3)$$

(7)

where $A$ is valid duration trigram set in training corpus and $B$ is valid duration bigram set in training corpus. The constraints to the segment durations are only decided by the standard duration they transfer to and the specific digit model of the durations can be ignored in $(7)$. By this way, the number of probability needed can be reduced to $K^3$.

The position of the segment in the utterance is a main factor that influence the length of duration. The segment in the Intonational phrase (IP) has the strongest lengthening effects and the duration in prosodic phrase (PP) and utterance boundaries are lengthened too, but not as obvious as in IP [6]. We use the silence segment to identify these boundaries. If a segment is preceded (followed) by a silence segment, then this segment is a boundary. It is different from other segment durations that there are only two standard duration levels for the duration of silence. If the silence duration is up a given threshold, then we consider the silence segment is a pause between IPs or utterances and represent this silence by high-level silence model. If the silence duration low than or equal to the given threshold, we consider the silence segment is a pause between PPs and label this silence low-level silence notation. In experiments, we set the threshold to 10 frames.

4. SPEECH RATE CLASSIFIER

The speech rates of digit strings are sorted by utterance duration to 3 levels: fast (0), average (1) and slow (2). Then we build 3 duration trigram models for these 3 levels separately. Figure 1 is the duration trajectories of “9509038” in different representation. One is the original
duration; the other is the standard duration. The number of standard duration level for this experiment is 40. The trajectories below are similar for the same digit string and the trend is more obvious in the standard duration trajectory than in original duration trajectory.

Table 1 is the distribution of speech rate difference labeled by original utterance duration and trigram context duration model for all the strings in data corpus. L is the speech rate labeled by original utterance duration and L' is the results of trigram context duration model in different number of standard duration level. More than 95% prediction results are same or difference limited to 1 level. As the number of standard duration level increasing, the accuracy of classifying the speech rate is increased too, which proves that the duration smoothing method works.

5. SPEECH RATE COMPENSATION TECHNIQUES

Segment model (SM) is a family of methods using segmental rather than frame-based features (e.g. in HMM) to represent the underlying characteristics of observation sequence. In SM, a general model for a random length observation sequence $s = \{ x_1, x_2, ..., x_t \}$ generated by reference $\alpha$ according to the density is

$$ p(s | \alpha) = p(s | l, \alpha) p(l | \alpha) \tag{8} $$

where $p(s | l, \alpha)$ is the output densities describing observation sequences in acoustic model and $p(l | \alpha)$ is the duration distribution.

Usually, the decoding process of SM is solved by a dynamic programming. The score of best segment sequence $J'$ ending at time point $m$ can be expressed as

$$ J'_m = \max_{t,n} \left( J'_n + \ln[p(x^n | \alpha)(m-t) + \ln[P(\alpha)] + C \right) \tag{9} $$

where $C$ is the segment penalty cost.

Figure 1 Duration sequence trajectories of 9509038

At the end of the observation sequence, we obtain the optimal $J'$ and trace back to get the best segment sequence.

The speech rate measure method we proposed above need to pre-know the segment of the acoustic unit. It is a problem for HMM based system before getting the recognition results; however, SM can produce the duration information while decoding the observation sequence. So it is better to realize the compensation methods based on a SM system. The baseline we used in this work is a stochastic segment model based system (SSM). Details of this model please refer to [1]. There are 40 regions for each reference and 10 mixture components for each region. The number of standard duration level is 60.

The penalty cost $C$ in (9) highly relates to the speech rate in SM. See Table 2. Large penalty cost can reduce the insertion error rate and in the mean time increase the deletion error rate, and vice versa. In conventional SM system, the penalty cost is a constant value and can’t adjust by the variation of speech rate. Two methods based on dynamic adjusting the penalty cost to compensate the variation of speech rate have been examined.

Table 1 The distribution of speech rate difference labeled by original utterance duration and trigram context duration model in different number of standard duration

<table>
<thead>
<tr>
<th>L-L'</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard duration level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.0067</td>
<td>0.1858</td>
<td>0.6458</td>
<td>0.1767</td>
<td>0.0017</td>
</tr>
<tr>
<td>40</td>
<td>0.0083</td>
<td>0.1858</td>
<td>0.6467</td>
<td>0.1575</td>
<td>0.0017</td>
</tr>
<tr>
<td>60</td>
<td>0.0150</td>
<td>0.1875</td>
<td>0.6491</td>
<td>0.1475</td>
<td>0.0008</td>
</tr>
<tr>
<td>80</td>
<td>0.0177</td>
<td>0.1841</td>
<td>0.6650</td>
<td>0.1483</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

5.1. Soft Integrating

The first is a two-pass method. The speech rate is predicted by the segment sequence obtained in the first pass and the penalty cost $C$ is adjusted to the corresponding level to get the final results by second pass. The penalty cost for the first pass sets to slow speech rate: -1000 and the penalty costs for 3 speech rate (fast, average and slow) are -1000, -2000 and -3000 separately. An adjusting parameter $\lambda$ is used to balance the duration errors introduced by penalty cost of slow speech rate. The function (6) is modified as:

$$ P^* (l_1, \alpha_1 | l_1, \alpha_1, l_2, \alpha_2, r) = $$

$$ P^* (l_1, \alpha_1 | l_1, \alpha_1, l_2, \alpha_2, 0) + (2-r) \cdot \lambda $$

(10)

where $r$ is the speech rate level index

The speech rate level is the level with maximum trigram duration probabilities.
\[
\sum_{i=1}^{n} p^*(l_{i-1}, \alpha_{i-1} | l_{i}, \alpha_{i}, r) = r^* = \max \sum_{i=1}^{n} p^*(l_{i-1}, \alpha_{i-1} | l_{i}, \alpha_{i}, r)
\]  
(11)

Results can be seen on Table 3. The best result using the Soft Integrating method is only a little better than the best result in Table 2. However, comparing with the results that the segment penalty cost \( C \) set to -1000 in Table 2, we can see this compensation technique can achieve a 12.4% error reduction in recognition.

5.1. Hard Integrating

The second method integrates the trigram value into segment score directly so the decoding procedure of recognition can get help from the trigram duration model. However, it can greatly increase the complexity of the algorithm, since the optimal result at current time point is relating to the recognition results of the preceding two segments. We approximately take the trigram duration probability at current point which is only influence to the current segment and integrate this value just as the segmental penalty cost does.

\[
C^* = C + p_{tr} \cdot \gamma
\]  
(12)

where \( p_{tr} \) is the score for current duration trigram and \( \gamma \) is a control parameter, \( C \) sets to the average speech rate: -2000.

We also can take the N-best list to approximately solve this problem by resoring the N-best candidate segment sequence and choosing the best result. In experiment \( \gamma \) is 5000.

HDI method achieves best performance in test, which reduces 11% error comparing with the baseline system.

6. CONCLUSIONS

In this paper we give a new method to predict the rate of speech by considering the relations between durations in a sentence and develop two approaches for speech recognition to compensate the influence introduced by variation of speech rate. These compensation methods accommodate the speech rate of utterance by adjusting the segment cost factor and achieve 11.0% error reduction in recognition.

Our future work will concentrate on the improvement of both the speech rate classifier based on duration’s relation and the corresponding compensation techniques. Such as we can build trigram duration model for the acoustic units with similarity in duration connection relation separately and not build a trigram duration model the whole acoustic unit set. The criteria to class the acoustic unit set is still unknown and need further study. With respect to compensation techniques, using the segment penalty cost to adjust the speech rate is only a primary method. More

Table 2  Segment penalty factor as the function of digit string recognition

<table>
<thead>
<tr>
<th>C</th>
<th>String Corr</th>
<th>In err</th>
<th>Del err</th>
<th>Sub err</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1000</td>
<td>94.58%</td>
<td>0.45%</td>
<td>0.22%</td>
<td>1.03%</td>
</tr>
<tr>
<td>-2000</td>
<td>95.00%</td>
<td>0.33%</td>
<td>0.31%</td>
<td>1.04%</td>
</tr>
<tr>
<td>-3000</td>
<td>94.42%</td>
<td>0.23%</td>
<td>0.46%</td>
<td>0.99%</td>
</tr>
</tbody>
</table>

Table 3 Adjusting Parameter \( \lambda \) as function of digit string recognition

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>String Corr</th>
<th>In err</th>
<th>Del err</th>
<th>Sub err</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>95.00%</td>
<td>0.35%</td>
<td>0.28%</td>
<td>1.04%</td>
</tr>
<tr>
<td>0.10</td>
<td>95.25%</td>
<td>0.25%</td>
<td>0.28%</td>
<td>1.03%</td>
</tr>
<tr>
<td>0.20</td>
<td>94.92%</td>
<td>0.28%</td>
<td>0.36%</td>
<td>1.04%</td>
</tr>
</tbody>
</table>

Table 4 Hard Directly Integrating (HDI) methods vs. N-best method

<table>
<thead>
<tr>
<th></th>
<th>String Corr</th>
<th>In err</th>
<th>Del err</th>
<th>Sub err</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI</td>
<td>95.55%</td>
<td>0.12%</td>
<td>0.24%</td>
<td>1.03%</td>
</tr>
<tr>
<td>5-best</td>
<td>95.00%</td>
<td>0.33%</td>
<td>0.31%</td>
<td>1.04%</td>
</tr>
<tr>
<td>20-best</td>
<td>95.25%</td>
<td>0.23%</td>
<td>0.30%</td>
<td>1.04%</td>
</tr>
</tbody>
</table>

parameters and models considered when pre-knowing the speech rate may achieve better results.

7. REFERENCES