INTRODUCE SEGMENTAL INNER TIME WARPING INTO PARAMETRIC TRAJECTORY SEGMENT MODEL FOR LVCSR

Jia Lei   Xu Bo

National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China 100080

ABSTRACT

In this paper, a parametric trajectory segment model (PTSM) with segmental inner time warping is proposed to improve the recognition accuracy of large vocabulary continuous speech recognition (LVCSR). The proposed PTSM utilizes the state boundary information provided by HMM system during decoding to do segmental inner time warping. Good alignment between different length realizations of a same phone unit can be obtained by this method. Based on the effective alignment, a new distance measure of measuring the average value of the norm of the residual error is used in k-means clustering to decide the parameters of the mixture density of PTSM. For two LVCSR tasks, the HMM system working with the proposed PTSM can give a consistent improvement over either the HMM system working with the traditional PTSM or the HMM system working alone.

1. INTRODUCTION

HMM has been successfully used in the current speech recognition system. One success of HMM is that it provides a dynamic time warping mechanism to warp the different length realizations of a phone unit to a suitable state sequence. Good time alignment between different length realizations can be achieved by this dynamic time warping. Although HMM has made great progress in speech recognition, it has an inherent assumption that successive observation within a state are independent and identically distribute (IID)[1]. Segmental models are proposed in order to overcome the inherent IDD shortcomings of HMM by modeling the full length of a phone in a whole unit, but it introduce the difficulty of aligning different length realizations of a same phone unit. How to align different length realizations of a same phone unit is crucial for segmental models.

In recent years, there has been an effort of adopting segmental models in LVCSR in order to improve accuracy of the recognition system[1,2,3]. In [1], the author introduces non-linear dynamic time warping into non-parametric segmental model to align different length realizations of a same phone unit. Improved performance can be obtained by this dynamic time warping. In [2,3], parametric trajectory segment model (PTSM) is applied in LVCSR tasks, which just normalize different length realizations by start point and end point alignment. All segment data of a same phone unit are normalized uniformly. Compared with the dynamic time warping mechanism in [1], the alignment is very ineffective.

In this paper, we try to normalize different length realizations of a same phone unit more effectively by introducing segmental inner time warping into the traditional PTSM. The segmental inner time warping method normalizes different length realizations of a phone unit by utilizing the state information provided by HMM system during decoding. Not only the start and the end point but also the segmental inner point can be aligned by this normalization method. Based on effective alignment of different segment data, a more reasonable distance measure is proposed, which measure the distance from a segment data to a segmental trajectory by the average value of the norm of the residual error. Better k-means clustering result can be gotten using the distance.

The paper is organized as follows: In sections 2, algorithm of the PTSM with inner time warping is carefully described. In section 3, a new distance measure is introduced in order to decide the parameters of the mixture density of the PTSM. In section 4, some experiments are carried out to prove that the effectiveness of the segmental inner time warping. Finally the conclusion is draw in section 5.

2. INTRODUCE INNER TIME WARPING INTO PARAMETRIC TRAJECTORY MODEL

2.1 Traditional PTSM

In [4], the author proposed a PTSM which models a speech segment by a polynomial function of $t$. If the speech segment at time $t$ is denoted as $C_t$, and the length of the segment is $N$, the segment can be modeled below:
\[ C_r = ZB_r + E(\Sigma) \]  

(1)

where \( Z \) is a \( N \times R \) design matrix which deals with normalizing different length of segments data uniformly between times 0 and 1, \( B_r \) is a \( R \times D \) trajectory coefficient matrix and \( E \) denotes residual error. \( R \) is the regression order of the PTSM. \( D \) is the dimension of feature vector.

In the above PTSM, different length realizations of a phone unit are uniformly normalized. The equal step normalization is not verified by the actual acoustic distribution of the segment data because the acoustic distribution of a segment data is not uniform, which is the reason that HMM use multiply state to model a phone unit. Compared with the dynamic time warping mechanism of HMM, it is very ineffective for the PTSM to normalize different segment data uniformly.

2.2 The PTSM with segmental inner time warping

In the framework proposed by [2], segmental model can utilize the start and end information provided by HMM system to do normalization in LVCSR task, but the state sequence information provided by HMM during decoding are ignored. Here, we proposed a segmental inner time warping method to normalize the different length of segment data by using the state information provided by HMM system during search.

During acoustic model training process, force alignment helps to locate the start point, end point and state sequence of a triphone unit. We compute the average state occupying percentage of a triphone in the whole triphone duration. The average state occupying percentage information is used to do inner time warping in the training of PTSM of the triphone. The normalization is performed in a piecewise manner. Each segment data is divided into the four inner segment which correspond to the four state of the HMM models of the segment data. Whole segment data of a triphone is normalized between 0 and 1, but each four inner segment data is normalized according to its average state occupying percentage of the triphone. The acoustic training process utilized the new normalizing methods to devise the design matrix \( Z \), which is used in deciding the other model parameters. To reflect the temporal variability between the four inner segment data, four different variance matrix corresponding to the four inner segment data is used in order to obtain a more powerful modeling capability. Because of the effective normalization, realization variability of a phone unit is much reduced and the robustness and accuracy of the parametric segmental trajectory models is much improved. The new proposed PTSM with segmental inner time warping is carefully described below:

1) Calculate the average state occupying percentage Assume that HMM model \( m \) have \( K \) occurrences of the segment data in training data set. For \( k \in [1, K] \),

\[ Y_i = \{y_i, y_{i+1}, \ldots, y_i^{(m,k)}\} = \{Y_i(1), Y_i(2), \ldots, Y_i(S)\} = \{y_1^{(i,1)}, \ldots, y_1^{(i,k)}; y_2^{(i,1)}, \ldots, y_2^{(i,k)}; \ldots; y_S^{(i,1)}, \ldots, y_S^{(i,k)}\} \]

is the \( k \) th segment data of the model \( m \), where \( y_{i,t} (1 \leq t \leq L(k)) \) is the \( t \) th frame feature raw vector of the \( k \) th segment data \( Y_i \) and \( L(k) \) is the length of the \( k \) th segment data \( Y_i \). For the given segment \( Y_i \), the inner \( i \) th segment which corresponds to the \( i \) th state of the HMM model is \( Y_i(i) = \{y_{i,1}, y_{i,2}, \ldots, y_{i,1}\} (1 \leq i \leq S) \), where \( S \) is the max state index of HMM model \( m \) and \( l_i(i) \) is the length of the inner \( i \) th segment.

The occupancy percentage of the \( i \) th inner segment in the segment data \( Y_i \) is defined as:

\[ \{P_i(k) = l_i(i)/L(k), 1 \leq i \leq S\} \]

(2)

The average state occupying percentage of the \( i \) th state in the whole duration of the model \( m \) is defined as:

\[ \{\overline{P}_i = \sum_{k=1}^{K} P_i(k)/K \| i \leq S\} \]

(3)

2) Calculate the design matrix \( Z \).

For the segment data \( Y_i \), the design matrix \( Z_i \) is defined in a piecewise manner. Each piece corresponds to one state of the HMM model. For the \( i \) th inner segment of \( Y_i \):

If \( i = 1 \)

\[ Z_i(1) = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & (1 \times \overline{P}_i/l_i(1)) & \ldots & (1 \times \overline{P}_i/l_i(1))^8 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (l_i(1) \times \overline{P}_i/l_i(1)) & \ldots & (l_i(1) \times \overline{P}_i/l_i(1))^8 \end{bmatrix} \]

else

\[ Z_i(i) = \begin{bmatrix} z_{i,1}(i,1) & \ldots & z_{i,1}(i,R) \\ 1 & \ldots & z_{i,1}(i,1) \\ \vdots & \ddots & \vdots \\ 1 & \ldots & z_{i,1}(i,R) \end{bmatrix} \]

(4)

where \( z_{i,t}(r) = (\sum_{n=1}^{K} \overline{P}_i + t \times \overline{P}_i/l_i(i))^{r} \).

The design matrix is defined below:

\[ Z_i = [Z_i(1)^T, Z_i(2)^T, \ldots, Z_i(S)^T]^T \]

(5)

3) Calculate the model parameters.

Replicate the first frame of every segment data and add the replica of the first frame to the start point of each segment data. The \( k \) th augmented segment data is denoted as:

\[ \tilde{Y}_i = \{y_i, y_{i+1}, \ldots, y_i^{(m,k)}\} = \{\tilde{y}_i, \tilde{y}_{i+1}, \ldots, \tilde{y}_i^{(m,k)}\} \]

\[ \tilde{Y}_i(1), \tilde{Y}_i(2), \ldots, \tilde{Y}_i(S) = \{\tilde{y}_1^{(i,1)}, \ldots, \tilde{y}_1^{(i,k)}; \ldots; \tilde{y}_S^{(i,1)}, \ldots, \tilde{y}_S^{(i,k)}\} \]

Then the segmental parameter can be estimated by the following equation(6) and (7):

\[ \hat{B}_m = [\sum_{i=1}^{K} Z_i(T\tilde{Y}_i)]^{-1}[\sum_{i=1}^{K} Z_i^T\tilde{Y}_i] \]

(6)
\( \Sigma_n(i) = \sum_{k=1}^{K}(\hat{Y}_k(i) - Z_i(i) \hat{B}_n)\times(\hat{Y}_k(i) - Z_i(i) \hat{B}_n)^T / \sum_{k=1}^{K}L_k(i) \)

\[ 1 \leq i \leq S \] (7)

4) Calculate the likelihood of a segment data \( Y_i \).

\[
P(Y_i | \mathbf{B}_n, \Sigma_n) = \prod_{l=1}^{s} \prod_{i=1}^{L(l)} \exp(-0.5 \times (\hat{Y}_l(i) - Z_i^{l-1}(i) \hat{B}_n) \times \hat{\Sigma}_n(i)^{-1} \times (\hat{Y}_l(i) - Z_i^{l-1}(i) \hat{B}_n)^T / ((2\pi)^{0.5} | \hat{\Sigma}_n(i)^{0.5} | )
\]

where \( Z_i^{l-1}(i) \) is the \( l-1 \) th row vector of \( Z_i(i) \).

3. DISTANCE MEASURE OF MEASURING THE AVERAGE VALUE OF THE NORM OF RESIDUAL ERROR

Parametric multiply mixture trajectory segment models have been used for LVCSR in [2,3]. Several distance measures are used in the K-means clustering process to get the parameters of the mixture density. The Multivariate Gaussian distance is used successfully in [5]. The paper [3] have proposed a normalized method of linear interpolating the different segment data to the same length and a weighted Euclidean distance measure can be used after the normalization. Compared with the Likelihood distance [3], improved performance can be obtained by the normalization method. In section 2, we have proposed another normalization method, which divided the whole segment data into four inner segments and normalized the four segments respectively. Based on the new normalization method, a new distance measure is proposed. The distance is defined as the average value of the norm of the residual error \( E \). The K-means clustering based on the new distance measure is described below:

**Initialization:** The partitioning of the segment data is initialized on average and at random.

**Estimate a centroid:** The centroids are estimated using equation (6),(7).

**Re-partitioning of the data:** Repartition of the data is done based on the following distance measure:

Assume the parameters of the polynomial segmental model \( m \) is \( m(\mathbf{B}_n, \Sigma_n) \), the distance from a segment data \( Y_i \) to the model \( m \) is

\[
\text{Dis}(m | Y_i) = \sum_{l=1}^{s} \sum_{i=1}^{L(l)} \left( \hat{Y}_l(i) - Z_i^{l-1}(i) \hat{B}_n \right) \times \left( \hat{Y}_l(i) - Z_i^{l-1}(i) \hat{B}_n \right)^T / L(k)
\]

(9)

where \( Z_i^{l-1}(i) \) is the \( l-1 \) th row vector of \( Z_i(i) \).

4. EXPERIMENTS AND RESULTS

The main features of our HMM LVCSR is listed below: 12 dimension MFCC, 1 dimension normalized log energy plus 1 dimension pitch, and their 1 and 2 order derivative. Decision Tree based gender dependent class-triphone models are trained from the training Database DB863.

Each triphone is modeled by a 4-state left to right HMM. Open LM is trained from the corpus with 387 million words. Our system has a vocabulary of 40,000 words.

There are 7300 triphones appearing in the training data speech corpus. State clustering[6] is used to tie the rarely seen triphones to its “nearest” neighbors. After tying, 3478 triphones are left. Multiply Mixture trajectory segment model is built for each triphone. The acoustic feature of the segmental models is the power, delta power, 12 MFCC and 12 delta MFCC[2]. During acoustic model training process, Force alignment helps to locate the starting point, end point and state boundary information of a triphone. All the trajectory models are the 2 order \((R=2)\) polynomial trajectory model. The last two experiments are using the clustering method described in section 3 to get the parameters of the multiply mixture trajectory segment models. In LVCSR task, weighted acoustic scores of the segmental models are added to the end frame of each triphone to get a more efficient pruning.

4.1 Distance measure comparison

This experiment is carried out to compare the performance of the distance measures of Multivariate Gaussian distance [5], the weighted Euclidean distance between segment data after linear interpolation[1], and the distance in equation (9) proposed in section 3. We denote the three distance measures as Dis1, Dis2, Dis3.

The 16 triphones of 16 different finals with the similar left and right context are used to test the three distance measures. The proposed multiply mixture PTSM is used for the recognition of the 16 different triphones. The result of each item in table 1 is the best result which can be obtained by the proposed multiply mixture PTSM using k-means clustering based on the specified distance.

<table>
<thead>
<tr>
<th>Mixture num</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dis1</td>
<td>13</td>
</tr>
<tr>
<td>Dis2</td>
<td>20</td>
</tr>
<tr>
<td>Dis3</td>
<td>22</td>
</tr>
</tbody>
</table>

In table 1, the last two distance measures with segment data normalization seems to perform better than the distance measure in [5]. And the proposed distance measure with the segmental inner time warping gives a 12% error reduction over the weighted Euclidean distance measure with the normalization method of linear interpolation.

4.2 Triphone recognition

This experiment is carried out in order to test the relative discriminative power between the traditional PTSM and the new PTSM with segmental inner time warping. Denote POLY(old) and PLOY(new) as the two kind of parametric
trajectory segment model respectively. In this experiment, 100 triphones of the most confusing finals and 100 triphones of the most confusing initials are chosen to compare the performance of the two PTSM. The chosen triphones of the final have the similar left and right context, which is the same case as the triphones of the initial. Test1 is to recognize the 100 triphones of final and Test2 is to recognize the 100 triphones of initial.

<table>
<thead>
<tr>
<th></th>
<th>Error Rate (Test1)</th>
<th>Error Rate (Test2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLY(old)</td>
<td>0.37</td>
<td>0.58</td>
</tr>
<tr>
<td>POLY(new)</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>HMM</td>
<td>0.39</td>
<td>0.51</td>
</tr>
</tbody>
</table>

From error reduction result of Test1 in Table 2, we can see that the performance of the proposed POLY(new) is better than HMM and the traditional POLY(old). From the error reduction result of Test2 in Table 2, we can see the recognition accuracy of HMM is better than the two kind of segmental models. This may be due to the trajectory assumption in the initial of Mandarin is not supported well as that in the finals. Nevertheless, POLY(new) gives a consistent better performance than that of the POLY(old) in the two test.

4.3 LVCSR test

Two test set is used to test the performance of the POLY(new) of LVCSR. Test set1 have 240 sentences spoken by 6 people with standard accent, the recording condition of which is well prepared. Test set2 have 600 sentences spoken by 10 persons with a slight accent, the recording condition of which is the ordinary lab condition.

Table 3 Recognition of LVCSR

<table>
<thead>
<tr>
<th></th>
<th>Test set1</th>
<th>Test set2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>87.05%</td>
<td>78.45%</td>
</tr>
<tr>
<td>HMM+ POLY(old)</td>
<td>87.52%</td>
<td>79.36%</td>
</tr>
<tr>
<td>HMM+POLY(new)</td>
<td>88.10%</td>
<td>80.00%</td>
</tr>
<tr>
<td>HMM+POLY(new) OF Final</td>
<td>88.04%</td>
<td>79.97%</td>
</tr>
</tbody>
</table>

The last Item of the Table3 is that only the segmental models of finals are applied in test. From the Table3, we can witness the HMM system working with the POLY(new) can give nearly 10% error reduction for the test set1 and 7% error reduction for the test set2 over the HMM recognition system, while the HMM system working with POLY(old) can only give an 5% and 3% error reduction over HMM recognition system. This supports our viewpoint well that the segmental model with inner time warping can achieve good time alignment between the various segment data realization and good performance can be obtained in this way.

From the last two items in Table3, we can see that the improvement of the PTSM with inner time warping combined with HMM is largely contributed by the PTSM of the finals for the Mandarin LVCSR task. We can make the conclusion that only the final have a trajectory which can be modeled with the parametric polynomial segment model and the trajectory of the initial is more likely to be random or not in a polynomial form.

5. CONCLUSION

In this paper, we have proposed a parametric trajectory segment model with inner time warping. The segmental inner time warping can achieve good time alignment between the various length realizations of a same phone unit. A more reasonable distance measure is proposed and is proven to be useful in improving the performance of the multiply mixture PTSM. Two triphone recognition task show that the new proposed parametric trajectory model have more powerful modeling capability than the traditional parametric trajectory model. The final LVCSR experiments show that the HMM system combined with the proposed segmental trajectory model give a 10% and 7% error reduction over the HMM system working alone, while the HMM system combined with the traditional segmental trajectory model only give a 5% and 3% improvement.

6. ACKNOWLEDGEMENT

The research work described in this paper was supported by the National Key Fundamental Research Program(the 973 Program) of China under the grant G19980300504 and the National Natural Science Foundation of China under the grant 69835003.

7. REFERENCES