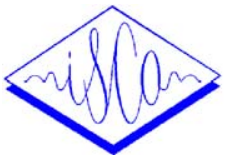

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A STUDY OF DURATION IN CONTINUOUS SPEECH RECOGNITION BASED ON DDBHMM

Zhao Qingwei[1], *Wang Zuoying*, *Lu Dajin*
Department of Electronic Engineering, Tsinghua University, Beijing, 100084

ABSTRACT DDBHMM solved the defects of traditional HMM. Based on DDBHMM, the problem of how to effectively utilize the duration information is studied in detail. The approach on estimating the duration distribution is introduced firstly, then the data file is classified according to the speak rate. The recognition experiment shows that, the duration information behaves best on the data of low speak rate, behaves normal on the data of medium speak rate and has little effect on the data of fast speak rate. Therefore, the most importance of duration is that by it the more accurate state segmentation point could be obtained and then the recognition rate can be improved. At the same time, the robustness of the system to speaking rate is improved with the employment of the duration information. Furthermore, the method of classified duration and normalized duration is also put forward and studied in detail, it shows that both of the two method can improve the effect. In order to study the dependency between the duration, the method of using the Bigram of the duration is proposed and analyzed. At last, the approach of post processing duration is studied, it shows that only based on DDBHMM, and utilizing the duration information synchronously in the recognition process, then the performance can be improved greatly.

1. INTRODUCTION

Traditional HMM is homogeneous Markov process, its state self transition probability a_{ij} is a constant independent of time. Then duration τ of state i satisfy exponential distribution:

$$P_i(\tau) = a_{ii}^{\tau}(1-a_{ii}) \qquad (1)$$

This is inappropriate according to the nature of speech. Since this defect of tradition HMM is cognized by people, the research on duration is the focus in the field of speech recognition[2][3][4][5]:

Though all the usual methods obtained some improvements, but there still exists contradictory phenomena in theory on their models. In fact, the state transition probability a_{ij} and the state duration probability is dependent on each other. Since the state duration has stable distribution[1], the more rational method to determine HMM parameter is that, first estimate the state duration distribution function $\{P_i(\tau)\}$, then deduce the state transition matrix $\{a_{ij}\}$. This leads to a duration distribution based hidden Markov model(DDBHMM)[6]. Because this model correctly processes the dependence of $a_{ij}(k)$ and $P_i(\tau)$, so it describes the physical nature of speech signal much better.

The utilization of duration information based on DDBHMM in the case of isolate syllable is studied in [7]. Since the continuous speech is not replete, the duration of continuous speech is different from isolate syllable speech. Therefore, in this paper, the effective utilization of duration information in continuous speech recognition based on DDBHMM is first studied.

In this paper, the estimate method of duration distribution is introduced, the research method of sorting the speech data according to different speech speed is put forward. Then the method of classified duration and normalized duration is put forward and studied. In order to study the correlation between the duration, the approach of Bigram of duration is also proposed. In addition, the method of post processing is studied. The recognition effect of all of the above method is examined and analyzed through experiments, and the conclusion about the utilization of the duration is obtained.

2. ESTIMATION OF DURATION DISTRIBUTION AND INTRODUCTION OF RECOGNITION ALGORITHM

2.1 Estimation of the duration distribution

The duration distribution is obtained during the training process. The training algorithm based on DDBHMM includes two procedure, on the first step, the training data is segmented and the optimal segmentation point can be obtained, on the second step, the model parameter can be derived by the principle of maximum likelihood estimation. Therefore, the segmented training data is employed to get the histogram of the duration distribution, and then the various statistical parameter can be computed, such as mean and variance, and so on.

The duration distribution described by parameter is often used in the recognition process. Because if the histogram distribution is needed to estimate, then the large amounts of parameter will require quantities of training data which could not be satisfied by the real database. However, if the duration distribution is described by parameter, then only small amounts of

[1] Now work at Intel China Research Center, Beijing. *Email: qingwei.zhao@intel.com*

parameter is demanded to estimate, so the training data will be relatively enough and the description of the duration distribution will be robust.

In most cases, both GAMMA and GAUSS function can approximate the raw duration distribution. In a relative sense, the approximation effects of GAMMA function is better, that is, it is nearer to the raw distribution. This is consistent to the conclusion of current references [11]. Therefore, in this paper the GAMMA function is employed.

The statistics of the duration parameter indicates that, there is strong law of the duration distribution, for example, the duration value of unvoiced stops (b,d,g) is small, and the duration value of fricatives (s, sh, x, f) is larger, this is consistent to the results from [1].

2.2 recognition algorithm

Suppose the T frames of input observation feature vector of continuous speech signal is $X=\{x(t)\}_{t=1}^{T}$, and the sentence A is consisted of L syllables, $W_1 W_2 W_3 \cdots\cdots W_L$ (L is unknown), then based on DDBHMM, according to the principle of maximum a posterior, among all of the candidate sentences, the most probable sentence which corresponding to the speech signal should be:

$$\hat{A}=\arg\max_{A} P(X/A)$$
$$=\arg\left\{\max_{0<L<\infty}\max_{(w_1,w_2,\ldots,w_L)}\max_{(S_1,S_2,\cdots,S_J)}\left[\prod_{k=1}^{L}\prod_{j=1}^{J}P_{k_j}(\tau)\prod_{t=S_{k_j}+1}^{S_{k_j+1}}b_{k_j}(x_t)\right]\right\} \quad (5)$$

where $P(X/A)$ is the probability of observation vector X on condition of sentence A, L is the sentence length (unknown), $P_{k_j}(\tau)$ is the duration of state j of syllable k. $b_{k_j}(x_t)$ is the probability density of observation vector under the case of syllable k and state i, S_{k_j} $(j=1,2,\cdots,J+1)$ is the state segmentation point of syllable k.

If the simplified DDBHMM is adopted, that is, the duration distribution is supposed to be uniform, then formula (5) can be changed as:

$$\hat{A}=\arg\left\{\max_{0<L<\infty}\max_{(w_1,w_2,\ldots,w_L)}\max_{(S_1,S_2,\cdots,S_J)}\left[\prod_{k=1}^{L}\prod_{j=1}^{J}\prod_{t=S_{k_j}+1}^{S_{k_j+1}}b_{k_j}(x_t)\right]\right\} \quad (6)$$

The detailed recognition algorithm is proposed based on the frame synchronous quick pruning algorithm[7][8]. It will not be introduced here.

3. CLASSIFYING ACCORDING TO THE SPEAKING SPEED AND THE RECOGNITION RESULTS

3.1 Classifying according to the speaking speed

In the case of continuous speech, the speaking speed may influence the recognition performance. In order to study the effect the duration information under different speaking speed, the speech data is classified. This work can proceed automatically, first the speech data is recognized with the simplified DDBHMM, then the state duration will be obtained by the state segmentation point, finally the speaking speed of one speaker can be computed as follows:

$$avtime=\frac{\sum_{i=1}^{I}\sum_{k=1}^{K_i}d_{i,k}}{\sum_{i=1}^{I}\sum_{k=1}^{K_i}m_{i,k}} \quad (7)$$

where $d_{i,k}$ is the duration of state k of sentence i, $m_{i,k}$ is the mean of the duration of the state k class of sentence i, K_i is the state number of sentence i. I is the total sentence number。

The larger $avtime$ is, the slower the speaker speed is.

The speaking speed of all of the speakers in the database is tested (every person corresponding to one data file), and the speakers are classified into three parts (A, B, C, From slower to faster):

A Class: $avtime<0.95$

B Class: $0.95\le avtime\le 1.05$

C Class: $avtime>1.05$

3.2 Test Experiments

The speech database is provided by the office of 863 intelligent computer topic, the speech material is selected from *people' daily*. The pronunciation of the speakers is mandarin and in declamatory form, some of the speakers have dialect background. The speech signal is transferred to SoundBlaster16 by CD1-41 type microphone. And the speech signal is converted into wave file by sampling in 16Khz and quantizing in 16bit. The recording environment is quiet office, the amplitude of the noise is small than 50dB.

The female speech data is employed in the experiments of this paper. There are 83 persons, each person corresponding to one data file. Every file includes 520 sentences, where 65 files is used to train, and 18 files is used to recognition (each of the A,B,C class includes 6 files).

In the front end, 45 dimensions are adopted, including 14 dimensions MFCC features, normalized energy and their first and second order differences. The semi-syllable acoustic units are exploited in the system, including 100 initial units and 164 final units. The last recognition result is in single syllable form without tone.

The baseline system is based on the simplified DDBHMM represented by formula (6). The other cases are based on the DDBHMM represented by formula (5). Therefore, the experiments in the paper are to compare the various approaches with duration information to the baseline system (uniform duration distribution).

In table 1, 2 and 3, the recognition effects in different cases with duration information are listed. From the table, it can be seen that for the slow speed speaker, the recognition with duration information got excellent results, the insert error rate is reduced from 6.77% to 2.72%, the relative reduction is 59.8%. The substitute error rate is reduced from 23.14 to 21.80, the relative reduction is 5.5%. Considering the low delete error rate, the change of the delete error rate will not be discussed. For the normal speed speaker, the relative reduction of the insert error is 40.7%, and that of the substitute error is 2.7%, this is also a good result. For the fast speed speaker (A), the

improvement is low, the insert error is reduced from 1.78 to 1.26, the relative reduction is 29.2%, and the substitute error is not reduced.

From the recognition result, it can be seen that, with the duration information, for the slow speed speaker and the normal speed speaker, the insert error of the recognition result is greatly reduced. This shows that the duration information can be used to accurately determine the syllable segment point, and then improve the recognition rate.

For the fast speed speaker, the duration information do not obtain better result, it can be interpreted as follows. In the cases of fast speaking, the absolute value of the insert error is low (about 1 percent), which indicates that the syllable segment point is basically correct. Hence the system can only depends on the discriminative ability of the duration information. However, for the fast speaking speech, the state duration distribution of many different semi-syllables is similar, thus the discriminative ability of the duration is not very strong, that is, the semi-syllable can not be distinguished by comparing different type of duration distribution.

In conclusion, the key importance of the duration information is that more accurate segmentation point can be gained by it, and therefore the recognition result can be improved.

Table1 Part C recognition performance with duration information

	Corr	Ins err		Del err	Sub err	
		rate	red		rate	red
base	76.50	6.77		0.36	23.14	
dur	77.78	2.72	59.8%	0.42	21.80	5.5%

Table 2 PartB recognition performance with duration information

	Corr	Ins err		Del err	Sub err	
		rate	red		rate	red
base	74.37	3.39		0.48	25.16	
dur	75.06	2.01	40.7%	0.50	24.44	2.7%

Table3 PartA recognition performance with duration information

	Corr	Ins err		Del err	Sub err	
		rate	red		rate	red
base	75.76	1.78		0.29	23.96	
dur	75.76	1.26	29.2%	0.28	23.96	0

4. CLASSIFIED DURATION AND NORMALIZED DURATION

In the above section, we employed uniform duration distribution for different class data. However, if different type of duration distribution is adopted according to different type of speech data, better results will be obtained. In detail, the train data is classified into three classes, and then estimate the A, B, C three type duration distribution, During recognition, the special type of duration will be made use of according to the speaking speed of the speaker. In table 4,5, the recognition result is listed in terms of this approach. (Because the data of class A is less influenced by the duration, it will not be discussed later). It can be seen that, the correct rate is improved further, for class C, the correct rate is improved from 77.78% to 78.02, for class B, the correct rate is improved from 75.06% to 75.29. At the same time, the insert error and substitute error of the system is reduced further.

Table 4 PARTC recognition performance of the classified duration and normalized duration

	Corr	Ins err		Del err	Sub err	
		rate	red		rate	red
base	76.50	6.77		0.36	23.14	
unified	77.78	2.72	59.8%	0.42	21.80	5.5%
classfied	78.02	2.30	66.0%	0.25	21.73	6.6%
normalized	78.41	2.10	69.0%	0.30	21.28	8.3%

Table 5 PARTB recognition performance of the classified duration and normalized duration

	Corr	Ins err		Del err	Sub err	
		rate	red		rate	red
base	74.37	3.39		0.48	25.16	
unified	75.06	2.01	40.7%	0.50	24.44	2.7%
classfied	75.29	2.04	39.8%	0.51	24.21	3.8%
normalized	75.40	1.26	62.8%	0.60	23.99	4.7%

At the same time, another problem is also considered, that is, if the classified duration is adapted, then because the training data of each class is reduced, the duration of some states will not be trained sufficiently. To solve this problem, all of the training data is recommended to use (that is, the training data will not be classified). But the normalization will be added in the training process, then the duration distribution estimated will be processed again according to special method, to make it useable to different type of duration distribution.

The detail procedure is as follows: firstly the training data is segmented according to DDBHMM, then the segment point of every state is obtained, subsequently the duration of every state is normalized, at last, the normalized duration distribution will be estimated.

The normalized duration is defined as the ratio of the state duration to the syllable duration which the state belongs to.

$$\bar{\tau}_{i,j} = \frac{\tau_{i,j}}{d_i} \qquad (8)$$

By employing the normalized duration, the dependence of the duration parameter estimated on the speaking rate will be reduced.

When the normalized duration is exploited to the recognition, the duration distribution should be adjusted in term of the speaking speed of the speaker, that is, the new duration must be derived, which is equal to the normalized duration multiplied by the normalization value syl_len:

$$\tau_{i,j} = \bar{\tau}_{i,j} * syl\_len \qquad (9)$$

where the normalization value is determined in the light of the average syllable speed of the training data.

From table 4, it can be seen that the recognition performance of the new duration for Class C is improved distinctly. The correct rate is improved by 1.91 percent, from 76.50 to 78.41. The insert error is reduced by 69.0% relatively, the substitute error is

reduced by 8.3% relatively.

The recognition performance of the new duration for Class B is also improved. The correct rate is improved by 1.03 percent, from 74.37 to 75.40. The insert error is reduced by 62.8% relatively, the substitute error is reduced by 4.7% relatively.

On the other hand, it can be seen that the recognition performance of the new duration is improved further in contrast to the classified duration, which showed the validity of our approach.

At the same time, if the duration is not exploited, the system will exhibit a large insert error for the slow speaker and small insert error for the fast speaker. By effectively employing the duration information, the insert error of the system to different speaking rate is reduced to a consistent level (1%~2%), which in fact improved the robustness of the system to the speaking speed.

5. BIGRAM OF THE DURATION

In formula (5), it is supposed impliedly that the duration between different states is independent of each other. If the correlation between the duration is considered, then formula (5) can be extended as follows:

$$\hat{A} = \arg\left\{ \max_{0<L<\infty} \max_{(w_1,w_2,\cdots,w_L)} \max_{(S_1,S_2,\cdots,S_J)} \left[\prod_{k=1}^{L} P_k(\tau_1,\tau_2,\ldots,\tau_J) \prod_{j=1}^{J} \prod_{t=S_{k_j}+1}^{S_{k_j+1}} b_{k_j}(x_t) \right] \right\} \quad (10)$$

If the state duration is supposed to be one-order Markov source, that is, only the bigram between the duration is considered, then above formula can be rewritten as:

$$\hat{A} = \arg\left\{ \max_{0<L<\infty} \max_{(w_1,w_2,\cdots,w_L)} \max_{(S_1,S_2,\cdots,S_J)} \left[\prod_{k=1}^{L} \prod_{j=1}^{J} P_k(\tau_j/\tau_{j-1}) \prod_{t=S_{k_j}+1}^{S_{k_j+1}} b_{k_j}(x_t) \right] \right\} \quad (11)$$

where $P(\tau_1/\tau_0) = P(\tau_1)$

In this case, the conditional distribution of the duration should be estimated. From the principles of the probability, the conditional probability of the variable of two dimension Gauss distribution can be easily computed, thus in this paper the Gauss distribution is employed.

In table 6, the recognition result of the system with the Bigram of the duration is showed. Comparing with table 1 and 2, it can be seen that the performance of the Bigram of duration is nearly the same as the Unigram of the duration. It can be interpreted as follows, the dependency between the duration is not very strong, in addition, the training data is not very sufficient, so it is difficult to estimate very accurate Bigram parameter, which also influenced the effect the Bigram. Therefore, the feature of using the Bigram of duration, must be further studied.

Table 6 Recognition performance of the Bigram of duration

	correct	Ins err	Del err	Sub err
PARTC	77.68	2.35	0.26	22.06
PARTB	75.10	1.49	0.26	24.64

6.METHOD OF DURATION POST PROCESSING

In this section, the post processing method is tested and compared with previous method.

The multi candidate results can be obtained with current program. That is, for every adjacent syllable segment point in the optimal path, N syllable candidates can be derived, at the same time, the match distance and the segment points of the N candidates with the acoustic feature vectors can also be reserved. When the post processing method is used to duration information, the match distance of every candidates is added by its duration likelihood, which is then regarded as updated match distance, thereafter the multi candidates should be resorted with regard to the updated match distance.

In table 7, the recognition performance of the method of duration post processing is showed. It can be seen that the recognition performance of this method descends in contrast to the baseline system, because the method of post processing could not change the syllable segment point, and the discriminative ability of different state duration is not very strong, this is consistent with previous analysis.

Table 7 Recognition performance of the method of duration post processing

	correct	Ins err	Del err	Sub err
PART C	75.70	7.08	0.38	23.92
PART B	73.50	3.56	0.48	26.02

7. CONCLUSION

Based on DDBHMM, the effective application of the duration in continuous speech recognition is studied. The recognition experiments shows that, the use of duration information behaves best to slow speakers, better to normal speakers, not manifest to fast speakers. The importance of the duration lies in the fact that with the help of the duration, the more accurate syllable and state segmentation points can be obtained, and then the recognition performance can be improved. At the same time, the robustness of the system to speaking rate is improved with the employment of the duration information.

REFERENCES

1. Qi Shi-qian, Zhang Jia-lu, "A study of duration of Chinese consonants," in ACTA ACUSTICA, vol.7, No.1, pp.8-13, Jan., 1982 (in Chinese).
2. Ferguson J D. Variable duration models for speech. Proc Symp Applic Hidden Markov Models Text Speech, 1980: 143,179,
3 Levinson S E. Continuously variable duration hidden markov models for automatic speech recognition. Computer Speech and Language, 1986, 1(1): 29,45
4 Russell M J, R.K.Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. Proc.ICASSP, 1985: 5,8
5. D.Burshtein, "Robust parametric modeling of durations in hidden markov models", IEEE Trans. On SAP, Vol.4, No.3, 1996, pp240-242
6. Wang Zuoying, "Improved hidden Markov Model in Speech Recognition", 863 Smart Computer System Conference, China, Dec, 1988
7. Wang Zuoying, Gao Hongge. An inhomogeneous speech recognition algorithm. First International Conference on Multimodal Interface(ICMI' 96), 1996, October: 15,17
8. Zhao Qingwei, Wang Zuoying, Lu Dajin, "Improved algorithm with duration information for continuous speech recognition", Journal of Tsinghua University,1997,12,pp87-90