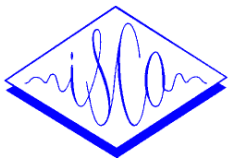

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

State Duration Constraint using Syllable Duration for Speech Recognition

Yumi Wakita* Eiichi Tsuboka

Central Research Laboratories, Matsushita Electric Industrial Co., Ltd
3-4 Hikaridai Seika-cho Soraku-gun Kyoto 619-02 JAPAN

Abstract

For speech recognition using HMMs, we propose an adaptive syllable duration constraint method. The method constrains syllable durations using a relation each syllable included in the same utterance [1]. The duration of t-th syllable $d(t)$ is predicted by using $d_o(1) \ldots d_o(t-1)$ the durations of syllables which have been recognized . After a syllable is recognized, if the durations of the t-th syllable is very different from the predicted value, the result is rejected. Advantages of this method are

- Its applicability is independent of the speed of speech.
- The durations of syllables within a sentence don't vary unnaturally.

These qualities are not found in non-adaptive duration constraint method. We confirmed that this method improves recognition rate [1]. If this syllable duration prediction (SDP) method can be used for constraining the duration of HMM states, the duration constraint can be integrated with matching and will bring SDP's improvements in recognition rate and computing time.

This paper proposes a new method of state duration constraint using SDP. At first the duration of s-th state of t-th syllable is predicted using the duration of t-th syllable which is predicted by SDP. Next the matching period of the state is constrained using the predicted state duration.

We evaluate this method using word and sentence recognition. For word recognition (100 words and 9 speakers, open test), the error reduction is 14% and the matching speed is 25% shorter. For sentence recognition (50 sentences and 6 speakers, open test), the error reduction is 46% and the matching speed is 50% shorter.

1 Introduction

In speech recognition using HMM, several methods have been proposed for incorporating state duration information and the effectiveness of these methods is shown in [2] [3] [4] [5]. However these methods model the duration of each state or each word only. The relation of each state which are included in the same utterance isn't considered. As the result of analyzing recognition errors, we find several sentences for which matching periods of syllables or words are unnatural. For decreasing these errors, the duration should be controlled by using the relation among the durations of separate syllables or durations of separate words within a sentence.

We have proposed a duration constraint method using syllable duration prediction. The syllable duration was predicted by linear combination of syllable matching periods already recognized within a sentence. After a syllable was recognized, if the t-th syllable matching period was very different from the predicted value, the result was rejected. This method could be applied only after matching, not simultaneously, becourse the unit used was a syllable. If it were possible to change the unit from syllable to HMM state, the duration control could be done simultaneously with matching. It would be efficient for recognition rate and matching time.

In general the syllable duration depends on the speed of speech regardness of syllable type. This is a basic assumption of the SDP method. But the relation between state duration and speed of speech depends on state type and position. In this paper, we propose a new method of the state duration control using state duration prediction. Assuming that the syllable duration depends on the speed of speech strongly, we predict the state duration by using the predicted syllable duration which is calculated by SDP and the relation between syllable duration and state duration.

We explain in section 2 the prediction method in consideration of the relation between state duration and speed of speech and in section 3 the performance of prediction. Section 4 describes the recognition method using state duration prediction and the experimental results.

2 Method of state duration prediction

2.1 Relation between state duration and syllable duration

In order to investigate whether the relation between state duration and speed of speech depends on state type and position, we check the relation between syllable duration and state duration assuming that a syllable duration depends on speed of speech strongly. We calculated the correlation coefficients between syllable matching periods and state matching periods. These matching periods are selected from only the sentences which were correctly recognized. We describe about the condtion of recognition in section 4.

For investigating variations in speaking style, we confirmed using both database of words and sentences separately. We use the database of 100 city names 6 males offered by JEIDA as word data and the database of the A-set 50 sentences 6 males in continuous speech

*Currently working at ATR Interpreting Telecommunications Research Labs.

corpus edited by the ASJ as sentence data.

The Table 1 shows the proportion of the data for which correlation coefficients are over 0.7 . The tendency of the results for word data is almost same as for sentences. We describe only the result for all speakers for word.

Table 1. The fraction of data of which correlation coefficient are over 0.7

speech style	speaker	state1	state2	state3	state4
sentence	ECL0001	0.44	0.20	0.31	0.74
	ECL0002	0.43	0.26	0.41	0.64
	NEC0001	0.45	0.29	0.43	0.68
	NEC0002	0.39	0.28	0.32	0.61
	TSU0001	0.51	0.21	0.36	0.63
	TSU0003	0.49	0.21	0.35	0.66
	all speaker	0.39	0.13	0.27	0.56
word	all speaker	0.22	0.21	0.37	0.53

The results are as follows:

1. The duration of 4-th state has the greatest effect on the syllable duration and the duration of 2nd state have little effect on the syllable duration. Even in the 4-th state, the correlation coefficients of the short vowels and the long vowels and the syllables containing unvoiced vowels are not so high.

2. The tendency of the result 1 is found in all results, independently of the speakers and the speaking style.

Figure 1 shows the distribution of the correlation coefficients of the 4-th state which shows the highest correlations and the 2nd state which shows lowest correlations. This figure shows clearly that the relation between state duration and syllable duration depends on state type and position.

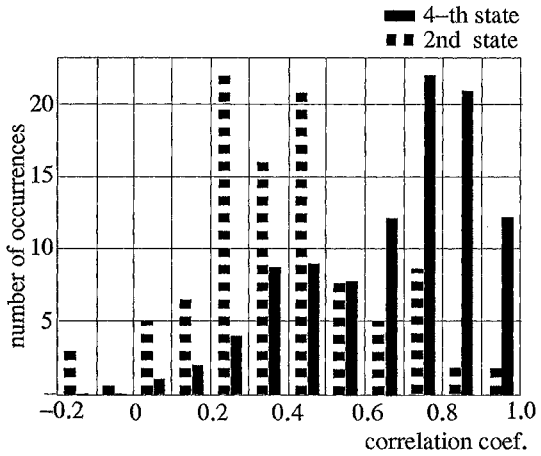

Fig. 1. The distribution of the correlation coef. of the 4–th state and 2nd state. (50 sentences , 6 males)

2.2 Method of Prediction

The estimation method of phoneme duration for investigating the relation between phoneme duration and speed of speech has been proposed [6] [7]. In general the relation between phoneme duration and speed of speech depends on the phoneme type. Based on this assumtion, the phoneme duration is estimated by a regression model using the average of vowel durations. In this paper we use a regression model using the state duration instead of the phoneme duration and we use the predicted syllable duration instead of the average of vowel durations. The formula for state duration prediction is as follows:

$$\overline{d_{state}}(t,s) = \overline{d_{sy}}(t)a(t,s) + b(t,s) \quad (1)$$

$\overline{d_{state}}(t,s)$: the predicted state duration of t-th syllable of s-th state

$\overline{d_{sy}}(t)$: the predicted t-th syllable duration

$a(t,s), b(t,s)$: weighting value of s-th state of t-th syllable

$a(t,s), b(t,s)$ are calculated by the method of least squares using training data before prediction. In the case that the sentence or word period can be known before prediction, formula 2 is used for the syllable duration prediction. If these periods are unmeasureable, formula 3 is used.

$$\overline{d_{sy}}(t) = L\frac{\sum_j (\alpha_j f_j(t))}{\sum_{i=1}^{N}\left(\sum_j \alpha_j f_j(i)\right)} \quad (2)$$

$f_j(i)$: the average duration of all the same syllables as t-th syllable for j-factor

α_j : weighting value of factor j

L : sentence or word period

$$\overline{d_{sy}}(t) = \frac{1}{m}\sum_{i=t-m}^{t-1}\left(d_o(i)\sum_j \alpha_j \frac{f_j(t)}{f_j(i)}\right) \quad (3)$$

$d_o(i)$: the matching period of i-th syllable.

3 Evaluation of Prediction Accuracy

For evaluating the effectiveness of formula 1, we calculate the difference between predicted value and true value. We compare our proposal method with non-prediction method which calculates the difference between the average value of each state and the true value. The true value is the matching period of states selected from words or sentences which were correctly recognized. The weighting value *a(t,s)* , *b(t,s)* in formula 1 are calculated using the true values of the state matching periods.

For training the HMM and estimating the weighting value by word data, the 100 cities name 48 males (TW set) offered by JEIDA are used. For prediction, the database of 100 words 9 males (EW set) are used. The words and speakers in EW set are different from TW set. For sentence data, the A-set 50 sentences 6 males (TS set) which are the same data as section 2 are used for training. And H-set 50 sentences 6 males (ES set) whose corpus is the same as A-set are used for prediction.

Table 2 and Table 3 show the standard devition (SD) and the worst-case error of the distribution of prediction errors for the state duration.

Table 2. The standard deviation (SD) and the worst-case error (word)

dataset	method	SD	worst-case error
TW set (closed)	average	33.6 msec	340 msec
	proposal	27.1 msec	220 msec
EW set (open)	average	35.1 msec	310 msec
	proposal	30.9 msec	240 msec

Table 3. The standard deviation (SD) and the worst-case error (sentence)

dataset	method	SD	worst-case error
TS set (closed)	average	21.8 msec	220 msec
	proposal	16.6 msec	150 msec
ES set (open)	average	27.4 msec	310 msec
	proposal	20.3 msec	240 msec

4 Recognition using state duration prediction

4.1 Our standard recognizer

The words and sentences consist of connected syllable FVQ-HMM [8]. A word or a sentence is recognized by viterbi algorithm using this dictionary. The duration isn't controlled. Table 4 describes the feature of the recognition system.

4.2 Proposed recognition system using state duration prediction

We combined the state prediction method described in section 2 with our standard recognizer. After training the syllable HMM, the recognition is done using training data for measuring the true duration of each state. Next *a(t,s)*, *b(t,s)*, the weighting values of each state and *e*, the prediction error of each data are calculated by the method of least squares using the formula 1. Next $e_{thold}(t,s)$, the error threshold of each state are decided using the prediction errors. The condition for e_{thold} is that 90% of the prediction errors of each state are less than the e_{thold} of each state. In recognition, at first the t-th syllable duration is predicted by using SDP and next the state duration is predicted by using formula 1. The prediction is done frame by frame simultaneously with matching. Next the limits of the state matching period is defined using the e_{thold} by the condition 4 as follows:

$$\overline{d_{state}}(t,s) - e_{thold}(t,s) < d_o(t,s) < \overline{d_{state}}(t,s) + e_{thold}(t,s) \quad (4)$$

Table 4. The feature of recognizer

Analysis	Sampling : 12kHz Window length : 21.3 msec Shift length : 10.0 msec Dimension : 10 order mel-cepstrum + 10 order delta-mel-cepstrum + delta power
Syllable model	Number of models : 132 syllables Model : FVQ-HMM [8] Codebook size : mel-cepstrum 256, delta-mel-cepstrum 256, delta power 64 States : 4 with loop and 1 for termination
Word	Training : 100 city names 48 males (TW set) Recognition : The same as training (TW set) and 100 words 9 males (EW set)
Sentence	Training : A-set in the corpus by ASJ 50 sentences 6 males (TS set) Recognition : H-set in the same corpus as TS set 50 sentences 6 males (ES set)

4.3 Evaluation of recognition result

We evaluated the proposed recognition method using state duration prediction by comparing it with our standard method and other state duration constraint methods. The same data are used as Table 4. For ES-set (H-set of continuous speech corpus by ASJ), we made a sentence dictionary which can accept the sentences which aren't spoken to the last word. Table 5 shows the example of sentences which are acceptable using this dictionary. If a sentence which isn't spoken complately is counted as a sentence, the 50 sentences in ESset becomes 354 sentences. In this evaluation, the 50 sentences are test using the 354 sentences dictionary.

Table 5. The example of sentences which are acceptable

h49-1	roojinwa
h49-2	roojinwa gyofutosite
h49-3	roojinwa gyofutosite subarasii
h49-4	roojinwa gyofutosite subarasii hitodato
h49	roojinwa gyofutosite subarasii hitodato omou

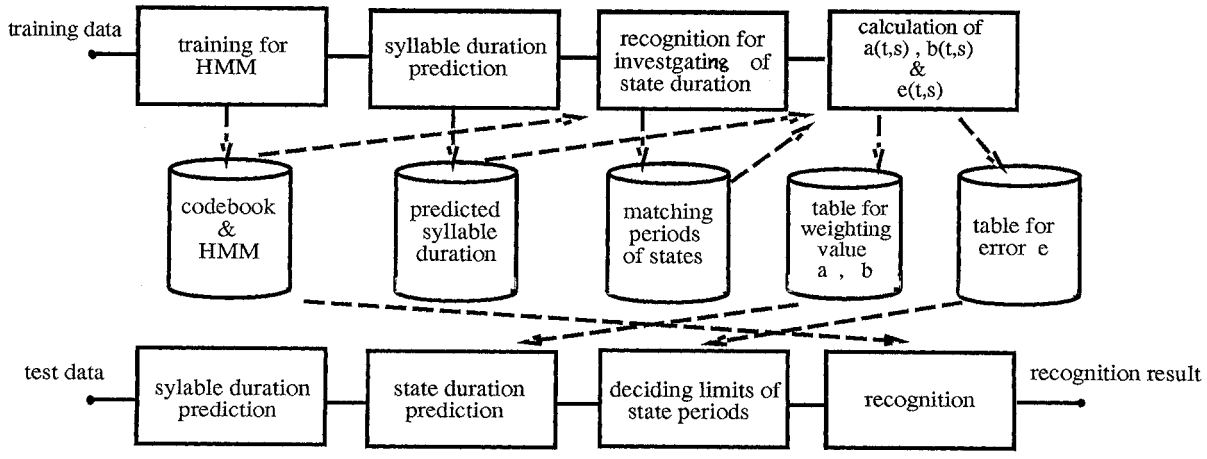

Fig. 2 The structure of the recognizer using state duration prediction

The condition for duration constraints we compared are as follows:

method1 standard method (non-duration control)

method2 only one limit is decided for all states

method3 limits of state periods are decided for each state

method4 limits of state periods are decided by proposed method

The conditions of the limits of state periods in method2 and method3 are that 90% of the matching periods of all states (method2) or each state (method3) in training data are included within the limits. Table 6 shows the recognition results.

Table 6. The recognition rate

	method1	method2	method3	method4
TWset(word)	97.9 %	98.1 %	98.0 %	98.4 %
EWset(word)	90.9 %	91.8 %	91.8 %	92.2 %
ESset(sentence)	95.0 %	97.0 %	95.3 %	97.3 %

The results are

1. By using the proposed method, the recognition rates improve for both words and sentences and for both the open test and the closed.

2. In the case of using method3, some errors vanish, but some new errors also occur, and the overall correct recognition rates don't improve compared with the method2. The reason is that the limits of the matching periods are too narrow and some states which are included in the correct sentences are rejected.

3. In the case of using the proposed method (method4), some confusable words or sentences which are difficult to be recognized using the standard system are improved. The example of confusable pairs are (/kiryuu/ , /chiryuu/) (/gobou/ , /mukou/) etc. in word data, and for sentences, (h49-4, h49) etc. in Table 5 .

4. For sentence data, we don't set limits for the state of pause models. Even if using duration control, misrecognized sentences in which the matching of pause models are wrong aren't improved. It's necessary to consider a duration control method for pause parts.

5. The limits of matching periods can be decided shorter by using the proposed method. For comparing the limit of each method, we calculated the range r defined as follows. The range for method2 is the difference between the maximum value and minimum value of the mathicng periods for all states. This range is constant for all states. The range for method3 is defined by formula 5. For our proposed method (method4), the range value is defined using formula 6.

$$r(t,s) = max\{d_o(t,s)\} - min\{d_o(t,s)\} \quad (5)$$

$$r(t,s) = 2 \times e_{thold}(t,s) \quad (6)$$

Table 7 shows the range r for method2 and the average value of the ranges $r(t,s)$ for method3 and method4. The range is 25% shorter than without state duration prediction (method3) for words and 50% shorter for sentences.

Table 7 The average of the ranges

	method1	method2	method3	method4
EWset(word)		230	80	60
ESset(sentence)		160	60	30

(msec)

5 Discussion

1. The relation between state duration and speed of spreech depends on syllable type and state position in syllable.

2. The proposed state duration prediction method using both the syllable duration and the relation between the state duration and syllable duration is effective in decreasing the prediction error.

3. The proposed recognition method which constrains the matching periods of state simultaneously with matching by using the prediction method is effective in improving the recognition rate. It also can decrease the calculation time.

6 Conclusion

We propose a method of state duration prediction and a recognition algorithm using the prediction method. And we confirm its effectiveness for improving the recognition rates for Japanese words and sentences.

Future work will be :

1. Improving the accuracy of duration prediction using other factors which affect the state duration strongly.

2. Applying the prediction method to other recognition algorithm.

References

[1] Takizawa(Wakita) Y., Tsuboka E. : Syllable duration prediction for speech recognition, ICSLP92 (Oct. 1992)

[2] Ferguson J.D. : Variable Duration Models for Speech, Proc. Symposium on the Application on Hidden Markov Models to Text and Speech pp. 143-179 (Oct. 1980)

[3] Levinson S.E. : Continuously Variable Duration Hidden Markov Models for automatic speech recognition Comput. Speech and Lang., 1, 1, pp. 29-35 (March 1986)

[4] Russel M.J. and Moore R.K. : Explicit modering of state occupancy in Hidden Markov Models for automatic speech recognition Proc. ICASSP85, pp.5-8 (1985)

[5] Lee K.F. : Automatic Speech Recognition, Kluwer Academic Publishers.

[6] Oosaka Y., Makino S., Sone T. : Spoken word recognition using information of duration by taking account of speaking rate and context of phoneme, Paper of Autumn Meeting of J.Acoust.Soc.Japan, (Oct. 1993)

[7] Matsuo H., Makino S., Kido K. : A Top-down Word Recognition System Using Phoneme Duration Model, IECIE Technical Report SP-89-90, pp.33-40 (1989)

[8] Tsuboka E., Nakahashi J. : On the Fuzzy Vector Quanrization based Hidden Markov Model, Proc. ICCASP94, 1, pp.637-640 (1994)