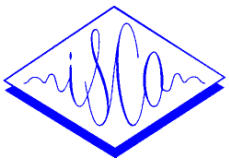

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Syllable Duration Prediction for Speech Recognition

Yumi Takizawa *Eiichi Tsuboka*

Central Research Laboratories,
Matsushita Electric Industrial Co., Ltd
3-1-1, Yagumo-Nakamachi,Moriguchi, Osaka 570, JAPAN

Abstract

In speech recognition using HMM, several methods have been proposed for controlling the state or word duration and their effectiveness is well known. However these methods model the duration of each state or word only, and don't consider the relation among durations of separate words within a sentence or separate states within a word.

This paper proposes a new method of syllable duration control for continuous Japanese speech recognition. It constrains the syllable duration using the relation of among each of syllables , and this method is effective even if the speed of speech changes . At first the syllable duration is predicted by using the matching periods which have already been spotted and using speaker independent factors which affect syllable duration. Next, the matching period of the predicted syllable is constrained using predicted duration. Using 50 sentences and 10 speakers, we evaluate the performance of prediction and recognition. As a result, this method improves sentence recognition rate by 5.2%.

10.21437/ICSLP.1992-367

1. Introduction

In speech recognition it's effective to use duration information for improvement of performance and calculation speed. In standard HMM, the probability of staying in a state decreases with time. Several duration models which are non-parametric model [1),2)] or parametric distributions with Gaussian, Poisson, Gamma or Log Gaussian [3)–6)] have been proposed. Also several methods which constrain word durations whose units are longer than a model have been proposed[7),8)]. However these methods can control the duration of each state or each word only.

As a result of analyzing spotting errors, we find several sentences for which matching periods of syllables or words are unnatural. And the distribution of syllable and word matching periods are too wide even though they are in the same sentence. For decreasing these errors, the duration should be controlled by using the relation among the durations of separate syllables or the durations of separate words within a sentence.

In this paper, we propose the method of duration control using syllable duration prediction. The syllable duration is predicted by a linear combination of syllable matching periods which have been recognized within a sentence. Weighting values of the linear combination are calculated using the factors which affect the syllable duration and are independent of the speaker. Using this prediction value the matching period is constrained.

We explain in section 2 the prediction method and in section 3 the performance of prediction in detail. Section 4 describes the recognition method using syllable duration prediction and experimental results.

2. Method of syllable duration prediction

We can imagine easily that if the speed of speech changes the syllable durations also change. So when predicting the target syllable duration, it's effective to use the information of the surrounding syllable durations which are spoken at the same speed as the target syllable. But differences in durations of syllables within a sentence are due to factors such as mora count or position or current syllable type or neighboring syllables etc. At first we should normalize the syllable durations using these factors. For normalization, we define the relation between syllable durations using these factors. These syllables are within the same sentence and spoken at almost the same speed.

$$\frac{d(t)}{d(i)} \cong \sum_j \alpha_j \frac{f_j(t)}{f_j(i)} \qquad (1)$$

$d(t)$: the t-th syllable duration
α_j : weighting value of factor j.
$f_j(t)$: the average duration of all syllables as t-th syllable for factor j.

By formula (1), the predicted duration of t-th syllable is defined as follows.

$$\hat{d}(t) \equiv d(i) \sum_j \alpha_j \frac{f_j(t)}{f_j(i)} \qquad (2)$$

For improving the performance of prediction, it's better to use the average of several syllable durations than to use the only i-th syllable duration. For example in syllable based left-to-right recognition algorithm, formula (3) which uses the average duration of the past m syllables is better than formula (2) which uses the only i-th syllable duration.

$$\hat{d}(t) = \frac{1}{m} \sum_{i=t-m}^{t-1} \left\{ d(i) \sum_j \alpha_j \frac{fj(t)}{fj(i)} \right\} \qquad (3)$$

Before predicting $\hat{d}(t)$ using formula (3), $f_j(t)$ and $f_j(i)$ must be calculated. For $f_j(t)$, at first we must decide the factors j. The necessary conditions of the factors are that they :

[1] Affect the syllable durations strongly

[2] Are independent of the speaker

The factors which influence Japanese phoneme duration have been reported[9),10)]. According to these reports, the duration is

affected mainly by the phoneme type, neighboring phoneme types, and its position in utterance group. In this paper, we select the current and neighboring syllable types as factors.

3. Evaluation of syllable duration prediction

3.1 Relation between syllable duration and syllable type

For confirming whether the syllable type is a good factor for syllable duration prediction, we check whether the syllable type affects the duration strongly (Expt.1) and whether the relation between the type and the duration is speaker independent (Expt.2). In this paper, the matching period is used instead of the syllable duration. This matching period is the result of syllable spotting and selected only in high score results.

Expt. 1 Partial and multiple correlation coefficients

For analyzing the relation between the current syllable duration and the current and neighboring syllable types, we calculate the partial and multiple correlation coefficients using Categorical Factor Analysis. Table 1 describes the database and Table 2 shows the coefficients values. The results are as follows:

[1] The average of the multiple correlation is 0.755 , and they don't vary much.

[2] For partial correlation, the current syllable type has the greatest effect on the current syllable duration .

[3] Current syllable duration is affected more by the following syllable type than the preceding.

[4] The tendency of results of [2] and [3] are consistent among all speakers.

Expt. 2 Correlation between speakers' average syllable duration, correlated by syllable type

For checking whether the relation between the syllable type and the syllable duration is dependent on the speaker , we calculate the correlation coefficients between speakers' average syllable duration, correlated by syllable type. Table 3 shows the coefficients value. All correlation values are high, we conclude the relation is independent of speakers.

From this, we can see that the current syllable type is satisfies the necessary conditions,which are described in section 2.

Table 1. The feature of Database

100 sentences (A-set and B-set in Continuous Speech Corpus for Research edited by the ASJ)
6 males
syllables : 1956 per speaker

Table 2. Partial and multiple correlation coefficient between the current syllable duration and the syllable type

Speaker	Multiple	Partial		
		type of preceding	type of current	type of following
1	0.750	0.401	0.707	0.484
2	0.681	0.415	0.611	0.444
3	0.756	0.473	0.706	0.474
4	0.775	0.517	0.746	0.528
5	0.753	0.548	0.691	0.595
6	0.812	0.584	0.718	0.599
Ave.	0.755	0.490	0.697	0.521

Table 3. Correlation between speakers' average syllable duration, correlated by syllable type

Speaker	1	2	3	4	5	6
1	1.000	0.852	0.876	0.808	0.775	0.884
2		1.000	0.825	0.861	0.794	0.816
3			1.000	0.837	0.773	0.863
4				1.000	0.791	0.824
5					1.000	0.928
6						1.000

3.2 The concrete formula

As the result of section 3.1, the syllable type is suitable as a factor j in the formula (3). In this paper, only one factor, the current syllable type is used , and the prediction formula (3) is reduced to formula (4). In the case of using only one factor, $\alpha_j = 1$,

$$\hat{d}(t) = \frac{1}{m} \sum_{i=t-m}^{t-1} d(i) \frac{f(t)}{f(i)} \qquad (4)$$

f(t) : Average duration of all syllables of same type as the syllable at time t

d(i): the i-th matching period

3.3 The evaluation of prediction accuracy

For evaluating the effectiveness of formula (4) , we evaluate the difference between the predicted value and the true value. We compared our proposed method with the non-weighted-prediction method. The database is same as described in Table 1. Figure 1. shows the distribution of the differences. We compare under three conditions. We find 4 is the most efficient value for "m". In all conditions, all six speakers are used for testing. The f(i) training conditions are:

[Cond. 1] Using the proposed formula (4), for training only one speaker is used [Open Speaker]

[Cond. 2] Using the proposed formula (4), for training all six speakers are used [Closed Speaker]

[Cond. 3] Using non-weighted-prediction , formula (5), which don't use the f(t) [Non -Trained]

$$d(t) = \frac{1}{m} \sum_{i=t-m}^{t-1} d(i) \qquad (5)$$

The results are described as follows:

[1] The average prediction error decreases more using proposed formula (4) than using non-weighted prediction formula (5).

[2] The worst-case error is greatest using non-weighted prediction formula (5).

[3] The tendencies of results [1] and [2] are a little greater by closed speaker than open speaker.

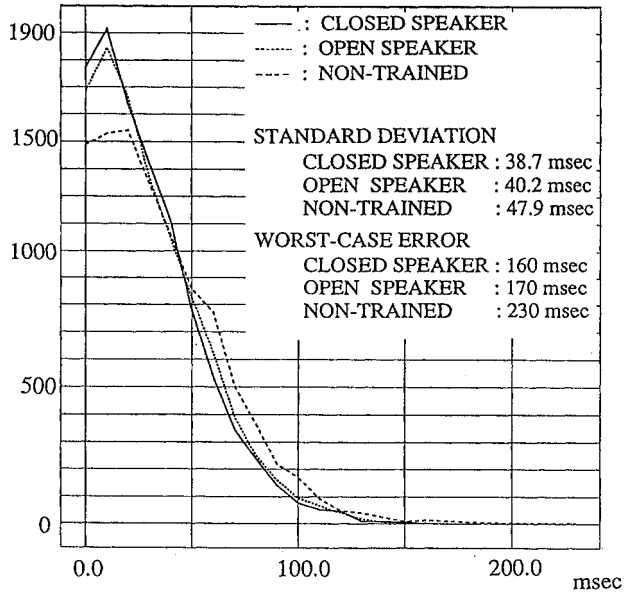

Fig. 1. Prediction Error for Syllable Durations

4. Recognition using syllable duration prediction

4.1. Our standard recognizer using syllable based HMM.

At first a syllable lattice is hypothesized by using a syllable spotting algorithm. Next the word decoding and the sentence decoding are done in order by connecting the syllable lattices and word lattices from left-to-right using the word-dictionary and word bi-gram. The specifications of the recognizer are shown in Table 4.

4.2. Analysis of the matching period of the recognized sentences

By using standard recognizer , we analyzed the syllable matching periods of the recognized sentences and compared the error sentences with the correct sentences.

Figure 2 shows the distributions of the standard deviation of the syllable matching periods of all sentences. We use two conditions as follows :

[cond. 1] Not compensating the matching period d(i)

[cond. 2] Compensating the period d'(i) using formula(6).

For both conditions' the training condition is closed speaker.

$$d'(i) = d(i) \times \frac{f(all)}{f(i)} \qquad (6)$$

f(i) : the average of the matching periods of syllables whose type is the same as the i-th syllable.

f(all) : the average of all matching periods

As a result,

[1] The distribution of the error sentences are broader than the correct sentences, and the average of the distribution is smaller.

[2] The tendency of result [1] is greater with the proposed compensation than without compensation.

Table 4. The feature of recognizer

Analysis
Sampling : 12 kHz sampling
Window length : 21.3 msec.
Shift length : 10.0 msec.
Dimension : 10 order mel-cepstrum + Δpower
Syllable Model
Models : 132 syllables
Distribution : Single Gaussian
States for all models : 4 with loop and 1 for termination
Duration control : Discrete Duration Control[2]
Language Model (Word Dictionary & Bi-gram)
Vocabulary size : 689 words
Word Perplexity : 5.85
Average number of words within one sentence : 14.4
Database
Training : B-set 50 sentences 6 males, A-set 50 sentences 10 males (above A-set 10 males including B-set 6 males)
Recognition : The same database as A-set for training

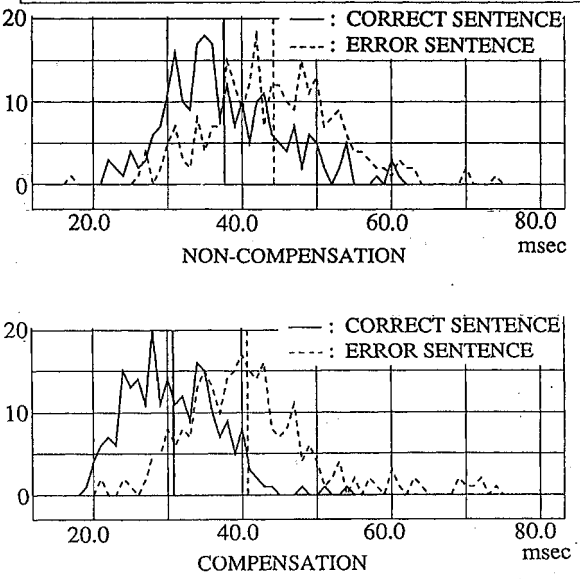

Fig. 2. Distribution of standard deviation of syllable matching period

4.3. The proposed recognition algorithm

For decreasing the number of sentences whose standard

deviations of matching periods are very high, the syllable or word lattice members whose durations are too far from the predicted duration should be eliminated. Now we combine the prediction method described in section 3 with our standard recognizer. When decoding, the durations of syllable and word are predicted and the unsuitable lattice members are eliminated. In the case of the prediction of syllable duration, the formula (4) in section 3 is used; for word , the following formula (7) is used.

$$\hat{dw}(t) = \frac{1}{m} \sum_{j=t-m}^{t-1} dw(j) \times \frac{fw(t)}{fw(j)} \qquad (7)$$

dw(t) : the t-th word matching periods

fw(t) is defined by formula (8)

$$fw(t) = \sum_{i=1}^{K(t)} f(i) \qquad (8)$$

K(t) : Number of syllables in t-th word.

Finally, the upper and lower limits for the constraints are defined by:

$$\hat{d}(t) - \beta < \text{syllable period} < \hat{d}(t) + \beta \qquad (9)$$

$$\hat{dw}(t) - \beta \times K(t) < \text{word period} < \hat{dw}(t) + \beta \times K(t) \qquad (10)$$

In this experiment, "β" is 40 msec which is the average of the standard deviation of the syllable duration in correct sentences. The syllables in the lattice members whose lengths aren't suitable are omitted. As a result, the recognition performance is improved.

4.4 Evaluation of recognition result

We compare the proposed recognition method using duration prediction to our standard method without prediction. We evaluate the proposed method under 2 conditions. One condition is "open speaker", another condition is "closed speaker" - the same conditions as cond.1 and cond. 2 in section 3.3. The recognition rates are shown in Table 5. The results are

[1] By using the proposed method, the average recognition rates improve by an amount independent of the training speaker.

[2] The rate for the closed speaker condition is a little greater than for ope speaker.

[3] The effectiveness of the proposed method is shown for most speakers. (except speaker no.5). but the rate of effectiveness does depend on speaker.

5. Discussion

[1] The proposed duration prediction method using both of the syllable matching periods and syllable type is effective in decreasing the prediction error, where the matching periods have already been spotting in the same sentence.

[2] From the recognition results, the distribution of the matching periods of the error sentences Vary more broadly than the correct sentences. When the matching periods are compensated by considering the syllable type, the above tendency is greater. The results show it's necessary to constrain the matching periods. And the syllable type is an effective factor.

[3] The proposed recognition method which constrains the matching periods of syllables and words by the prediction method is effective to improving the recognition rate. And this improvement is independent on training speakers .

6. Conclusion

We propose the method of word and syllable duration prediction and a recognition algorithm using this method. And we confirm it is effective improving Japanese sentence recognition rate. Future work will be :

[1] Improving the accuracy of duration prediction using other factors which affect the syllable duration strongly.

[2] Applying this prediction method to other recognition algorithms.

[3] Applying this prediction method to phonemes.

References

1) J.D.Ferguson "Variable Duration Models for Speech," Proc. Symposium on the Application on Hidden Markov Models to Text and Speech, pp. 143-179, Oct.1980.

2) Nakagawa.S., Hashimoto.Y. " Segmentation of Continuous Speech by HMM and Bayesian Probability," EIC (D-II) Vol.J-D-II, No.1, pp. 1-10, Jan.1989.

3) L.R.Rabiner " A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition " Proc. of IEEE Vol.77, No.2, pp.257-286, Feb. 1989.

4) S.E.Levinson " Continuously Valuable Duration Hidden Markov models for automatic speech recognition," Computer,Speech and Lang., Vol 1, No.1,pp.29-45, Mar.1986.

5) M.J.Russel, R.K.Moore,"Explicit modeling of state occupancy in hidden Markov Models for automatic speech recognition," Proc. ICASSP85,pp.5-8,1985.

6) Takada Y., Tsuboka E., Wakita H." A Comparative Study on Duration Models - Poisson, Gamma, and Logarithmic Normal Distribution - "Paper of Spring Meeting of J. Acoust. Soc. Japan, Oct.1991.

7) K.F.Lee, *"Automatic Speech Recognition"* (Kluwer Academic Publishers)

8) L.R.Rabiner, J.G.Wilpon, F.K.Soong " High Performance Connected Digit Recognition using Hidden Markov Models," Proc. ICASSP89, pp. 1214-1225

9) Kaiki N., Mimura K., Sagisaka Y. " Statistical Modeling of Segmental Duration and Power Control for Japanese " Proc. Eurospeech 91 , pp.625-628,Sep.1991

10)Takeda K, Sagisaka Y, Kuwabara H "On sentence-level factors governing segmental duration in Japanese," J.Acoust.Soc.Am. 86-(6) Dec.1989.

Table 5. The sentence recognition rate

no. speaker	Standard	Proposal (open)	Proposal (closed)
1	60.0%	70.0%	70.0%
2	34.0%	38.0%	42.2%
3	50.0%	52.0%	52.0%
4	42.0%	48.0%	52.0%
5	56.0%	50.0%	52.0%
6	34.0%	36.0%	38.0%
7	60.0%	64.0%	68.0%
8	46.0%	50.0%	48.0%
9	50.0%	58.0%	60.0%
10	40.0%	42.0%	42.0%
Ave.	47.2%	51.0%	52.4%