SPEECH RECOGNITION USING SYLLABLE DURATION RATIO MODEL

Masahide Ariu, Takashi Masuko, Shinichi Tanaka and Akinori Kawamura

Corporate Research & Development Center, Toshiba Corporation, Japan
E-mail: {masahide.ariu, takashi.masuko, shinichi.tanaka, akinori.kawamura}@toshiba.co.jp

ABSTRACT

This paper describes a novel approach to duration information modeling for speech recognition. To eliminate the influence of speaking rate on the duration model, we propose a model utilizing the duration ratios of two successive syllables by log-normal distributions. We refer to this model as a syllable duration ratio model (SDRM), and compare it with a syllable duration model (SDM) that represents the duration of the syllable itself. These duration models are compared in isolated word and connected digit recognition tasks under noisy conditions. Experimental results show that the SDRM outperformed the SDM, and reduced the errors by approximately 30% compared to the baseline system without duration model at 15dB or higher SNR in 10 digits recognition tasks. In addition, we show that the SDRM is robust with respect to the difference in speaking rate between training and test data.

1. INTRODUCTION

Modeling duration information is one of the effective ways to improve the performance of speech recognition systems. Although conventional hidden Markov models (HMMs) represent duration information implicitly by their state transition probabilities, it is well known that they are inappropriate for modeling the actual durations of states, syllables and words. Generally, it is expected that a duration model can reduce the number of deletion and/or insertion errors made by a speech recognition system. In addition, duration information is necessary to discriminate between certain words in some languages. From these points of view, there have been many attempts to incorporate explicit duration models into speech recognition systems [1], [2].

One of the problems with applying duration information to speech recognition is that duration is greatly influenced by the speaking rate. The difference in speaking rate between training and test data causes degradation of the recognition accuracy. To overcome this problem, Zhu and Lee [3], and Qingwei, et al. [4] have proposed construction of duration models separately for different speaking rate classes, and Dong and Zhu [5] have proposed normalization of duration by the estimated speaking rate. However, an error in the speaking rate estimation has a large influence on the results. Alternatively, to reduce the influence of speaking rate, stepwise duration estimation has been considered [3], [6]. In [3], [6], the duration of the current unit was estimated from the previous one by a conditional distribution or a linear regression. In the linear regression, the error distribution is implicitly assumed to be independent of the duration of the previous unit. However, the error distribution depends on the speaking rate and the duration. Therefore, the variance of the error distribution should be different in fast speech and slow speech.

In order to avoid the influence of speaking rate on the duration model, this paper proposes a novel technique for duration modeling based on duration ratio. Our duration model does not need to estimate the speaking rate. The idea and methods of the syllable duration ratio model are introduced in Section 2. In Section 3, experiments to evaluate the proposed model are described. Conclusions are presented and problems to be tackled in the future are indicated in the final section.

2. SYLLABLE DURATION RATIO MODEL

The rhythm of speech is one of the important factors for speech perception. It is known that there are different kinds of rhythm, such as stress-timed (e.g. English), syllable-timed (e.g. French) and mora-timed (e.g. Japanese) rhythms. Although the duration of units composing the rhythm depends on the speaking rate, it is assumed that the relation of successive rhythmic units is less subject to speaking rate than other units such as phones. As a result, the ratio of successive units becomes almost constant, even if the speaking rate varies from training data to test data.

We adopt syllables as the duration unit, and model the duration ratio of the current syllable to the previous one by a log-normal distribution. We refer to this duration model as a syllable duration ratio model (SDRM). Here, we assume that the distribution depends on syllable pairs only, though it may depend on phonetic and linguistic contexts of the pair such as position in the utterance. The model of a syllable to silence and silence to a syllable are undefined. The SDRM is applied within N-best recognition, rescoring N-best results using the SDRM scores.

Figure 1 shows an example of improving discrimination by use of the SDRM. Assume that three candidates were obtained from an input of "6" ("ro-ku" in Japanese): "5 6" ("go:-ro-ku", where ":" denotes a long vowel) which has an insertion error, "5 9" ("go:-kyu:") which has a substitution and an insertion error and "6" ("ro-ku") which is correct. The first candidate has an inserted digit "5" ("go:") with short duration, which has also reduced the duration of "ro". Since the syllable duration ratio of "ro" to "go:" and "ku" to "ro" are outliers of the SDRM, the first candidate yields a low log-likelihood SDRM score. For the third candidate, the syllable duration ratio of "ku" to "ro" is in the middle of the distribution of "ku" to "ro". As a result, the log-likelihood SDRM score is expected be higher than for the first candidate, so the

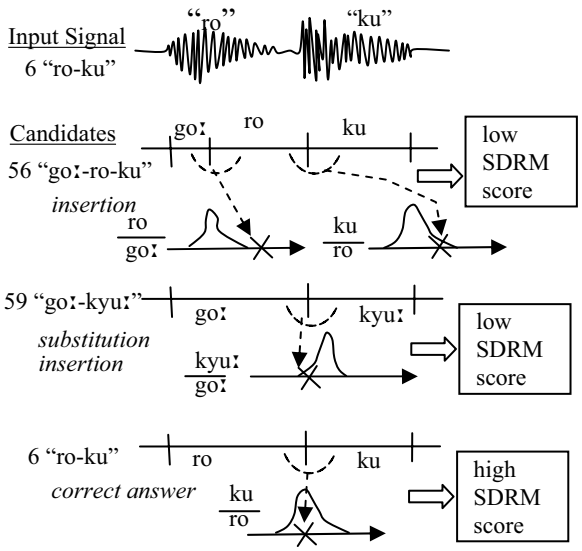

Figure 1. Illustration of improved discrimination by use of the SDRM.

insertion error can be suppressed by the SDRM score. Moreover, the difference of the SDRM distribution between the syllable duration ratio of "kyu:" to "go:" and that of "ku" to "ro" makes it possible to suppress the substitution and insertion errors of the second candidate.

2.1. Model training

Duration statistics for training the SDRM are obtained by running a forced alignment on the training data using a baseline HMM set. The SDRM models each duration ratio for all observable syllable pairs by a log-normal distribution.

Since it is difficult to provide sufficient data for all possible syllable pairs, decision tree-based clustering based on the minimum description length (MDL) criterion [7] is applied to the SDRM in a manner similar to state tying of triphone HMMs.

2.2. N-best rescoring

For the experiments in this paper, the duration models investigated are applied to N-best rescoring. The N-best lists for each utterance were generated by a baseline system without a duration model. The score for each candidate on the list is rescored by the duration model. The new score (S_{NEW}) is defined as the weighted sum:

$$S_{NEW} = w \times S_D + (1-w) \times S_A$$

where w is the weight ($0 \le w \le 1$) for the log-likelihood score of the SDRM (S_D) obtained from the time alignment result for each candidate by the baseline system, and S_A denotes the log-likelihood score of the acoustic model. S_D and S_A are normalized by the number of syllable pairs and the number of frames, respectively. The candidate with the best new score is chosen as the recognition result. We used 10-best candidates in the experiments described below.

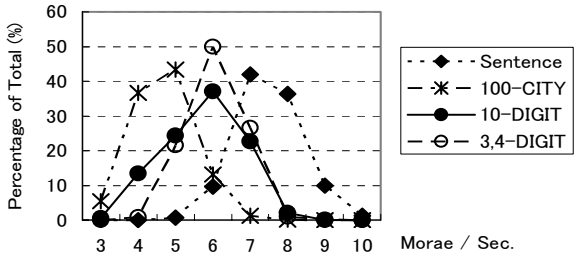

Figure 2. Distribution of speaking rate for each data set.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Database and setup

Recognition experiments were conducted on isolated word and connected digit tasks. We compared the SDRM with a syllable duration model (SDM), which modeled the distribution of syllable duration by a log-normal distribution.

Training data for the acoustic model, SDM and SDRM consisted of read Japanese short sentences of about 75 hours long by around 300 speakers. The speech data was sampled at 11025 Hz. To simulate a noisy environment, in the training of the acoustic model, stationary car noise was artificially added to the equally split training data at signal-to-noise ratio (SNR) of ∞ (clean speech), 18, 15, 12, 7, 5 and 0dB. Speech signals were windowed by a 256-point (about 23ms) Hamming window with 88-point (about 8ms) shift. The feature vector consisted of 12 MFCCs, log energy and their first and second derivatives. The acoustic model was a standard 3-state left-to-right triphone model using six Gaussian mixtures with diagonal covariance matrices per state. State tying was performed using decision-tree based clustering, so that the number of HMM states was reduced to about 1000. The SDM and SDRM were trained from the original clean data.

Three test sets were used for the experiments: one hundred Japanese city names by 40 speakers (100-CITY), 3 or 4 digits by 20 speakers (3, 4-DIGIT) and 10 digits by 80 speakers (10-DIGIT). Stationary car noise was added to the test data to evaluate the robustness of the SDRM to noise. The added noise was different from the noise used in the training. The start and the end of each utterance were assumed to be known in all the experiments.

There were two reasons for selecting the three test sets. First, these data simulated elementary in-car navigation tasks, which are major applications of speech recognition. Second, there was a difference in speaking rate between the training data and the test data.

Figure 2 shows the speaking rate distribution for each data set. The speaking rate is measured by the number of morae per second. In the figure, "Sentence" (the dotted line) is the distribution of the training data, and can be seen to have a faster speaking rate than the other data used in the experiment. The 10-DIGIT data varies more in speaking rate than the 3, 4-DIGIT data. The distribution of 100-CITY has the slowest speaking rate among these data sets.

Utterance accuracy (word and sentence accuracy) and relative error rate reduction (RERR) were used to measure the change in performance from the baseline system for each task. The RERR is defined by

Table 1. Utterance accuracy on 100-CITY task.

SNR	Baseline	SDM	SDRM
0dB	74.8%	75.6%	77.2%
5dB	87.7%	87.8%	89.2%
15dB	94.7%	94.7%	95.6%
25dB	96.5%	96.5%	97.3%
Clean	96.9%	96.9%	97.4%

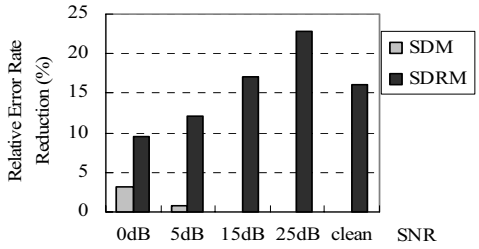

Figure 3. Relative error rate reduction on 100-CITY task.

$$RERR = 100 \times \frac{U_D - U_A}{100 - U_A},$$

where U_A is the utterance accuracy of baseline, and U_D is the utterance accuracy with the SDM or SDRM.

3.2. Isolated word recognition test

First, isolated word recognition results are shown. The test data was 100-CITY. The N-best rescoring weight was selected as the average of leave-one-out cross-validation run at each SNR. Table 1 presents the utterance accuracy of each model. Shown in the "Baseline" column are the results of the acoustic model only. Figure 3 illustrates the relative error rate reduction at each SNR. Table 1 and Fig. 3 show that the SDRM is effective under all SNR conditions, and the SDRM achieves a relative error rate reduction more than 15% at 15dB or higher. However, the SDM has almost no effect under all SNR conditions. The SDRM can improve 100-CITY results in spite of the difference from the training data in speaking rate.

From the recognition results, we found that the utterances "Hon-Joː" mistaken for "Hon-do" in the baseline model were corrected by the SDRM because of the difference in the vowel duration. However, the SDRM could not correct some confusion such as "ki-ryuː" with "chi-ryuː", because the syllable duration similarity made classification by duration difficult. These results indicate that (as expected) the SDRM is more effective when the difference in duration is important for distinguishing among candidates.

3.3. Connected digit recognitions tests

Connected digit recognition tests were conducted to show the validity of the SDRM in continuous speech recognition.

For both the 3, 4-DIGIT and 10-DIGIT tests, we used an open digit loop grammar which allowed any digit sequence with optional silence between digits. The silence segment between syllables from the time alignment result was ignored for the SDM and SDRM. The insertion penalty was optimized for utterance accuracy in 15dB by preliminary experiments, and the N-best rescoring weight was selected as the average of leave-one-out cross-validation run at 15dB. These parameters were used in all SNR conditions.

Table 2. Utterance accuracy for 3, 4-DIGIT task.

SNR	Baseline	SDM	SDRM
0dB	38.6%	50.4%	45.1%
5dB	74.0%	81.9%	79.4%
15dB	95.1%	96.5%	96.6%
25dB	96.7%	97.4%	97.6%
Clean	96.5%	97.0%	97.4%

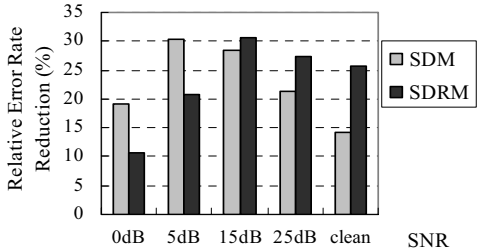

Figure 4. Relative error rate reduction on 3, 4-DIGIT task.

Table 3. Number of Errors for 3, 4-DIGIT task.

Model	SNR	Del.	Sub.	Ins.
Baseline	5dB	160	200	13
	25dB	14	7	14
SDM	5dB	55	157	23
	25dB	2	6	19
SDRM	5dB	115	163	12
	25dB	6	7	12

Table 2 shows the utterance accuracy for 3, 4-DIGIT, and the relative error rate reduction is illustrated in Fig. 4. Table 2 and Fig. 4 indicate that the SDM outperforms the SDRM when the SNR goes below 15dB. However, it also shows that the SDRM yields more than 25% relative error rate reduction and has higher performance than the SDM at 15dB or higher. Table 3 shows the number of errors (deletion, substitution and insertion) for 3, 4-DIGIT at 5 and 25dB. The SDM score for short duration units tends to be higher than that for long units, since the SDM was trained on training data with a faster speaking rate than in 3, 4-DIGIT. Accordingly, the SDM reduces the deletion errors at the cost of increasing insertion errors. The SDRM, however, can improve the utterance accuracy by decreasing deletion errors without increasing other errors.

Table 4 shows the utterance accuracy for 10-DIGIT. The SDRM achieves about 30% relative error rate reduction at 15dB or higher as shown in Fig. 5. Table 5 shows the number of errors for 10-DIGIT at 5 and 25dB. While the SDM can reduce deletion errors, insertion errors increase and cancel out the effect of the reduction in deletion errors. As a result, the SDM yields slight improvements in all SNR conditions. However, the SDRM can reduce all kinds of errors, and gives better discrimination despite the difference from the training data in the speaking rate.

Table 4. Utterance accuracy for 10-DIGIT task.

SNR	Baseline	SDM	SDRM
0dB	49.8%	51.0%	57.6%
5dB	72.5%	73.2%	78.5%
15dB	86.6%	86.9%	90.9%
25dB	87.1%	87.2%	91.1%
Clean	82.8%	82.9%	88.1%

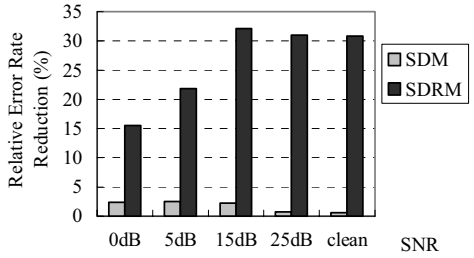

Figure 5. Relative error rate reduction on 10-DIGIT task.

Table 5. Number of Errors for 10-DIGIT task.

Model	SNR	Del.	Sub.	Ins.
Baseline	5dB	572	639	195
	25dB	122	150	371
SDM	5dB	513	640	222
	25dB	107	148	399
SDRM	5dB	307	639	163
	25dB	69	145	230

3.3. Robustness against speaking rate

Figure 6 shows the log-likelihood scores of the SDM and SDRM vs. speaking rate on the isolated word task (100-CITY). The points in the figure represent the median of the score distribution, and error bars correspond to 25% and 75% quantiles. In the figures, the score of the SDM is shown on the left and that of the SDRM on the right. Figure 6 shows that the score of the SDRM is less subject to speaking rate, whereas the score of the SDM varies with the change in speaking rate.

This result indicates that the SDRM is less sensitive to the difference in speaking rate between the training and the test data. As a result, the SDRM could improve all the evaluated tasks, even though the speaking rate was different from the training data. In contrast, the difference in speaking rate is considered to influence the score of the SDM. This accounted for the previous experimental results in which the SDM scarcely improved the utterance accuracy of the 100-CITY and 10-DIGIT tasks.

4. CONCLUSIONS

In this paper, the syllable duration ratio model (SDRM) which models the duration ratio of successive syllables was proposed to model duration information with robustness against speaking rate.

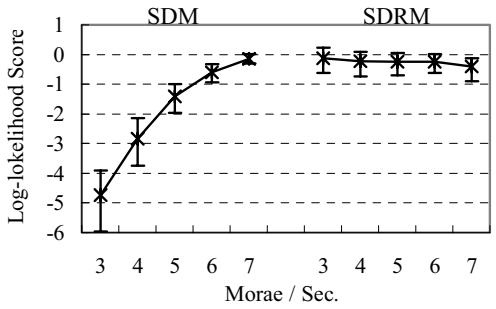

Figure 6. Log-likelihood score distributions of the SDM and SDRM on 100-CITY task.

Experimental results showed that the SDRM improved the utterance accuracy at all SNR and on all tasks compared to a baseline HMM system and a syllable duration model (SDM), with a relative error rate reduction of about 30% at 15dB or higher SNR for 10 digits recognition. The SDRM was shown to be able to distinguish between candidates with different syllable durations and reduce deletion and/or insertion errors. It was also shown that the log-likelihood score and the recognition results achieved with the SDRM were robust against speaking rate variation between the training and the test data.

Future work will include evaluation of the technique on languages other than Japanese and the influence of real noisy conditions such as the Lombard effect.

5. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[2] J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," *Proc. ICSLP-2004*, pp. 385-388, 2004.

[3] Y. Zhu and T. Lee, "Explicit duration modeling for Cantonese connected-digit recognition," *Proc. ICSLP-2004*, pp. 685-688, 2004.

[4] Z. Qingwei, W. Zuoying and L. Dajin, "A study of duration in continuous speech recognition based on DDBHMM," *Proc. Eurospeech'99*, pp. 1511-1514, 1999.

[5] R. Dong and J. Zhu, "On use of duration modeling for continuous digits speech recognition," *Proc. ICSLP-2002*, pp. 385-388, 2002.

[6] Y. Osaka and S. Makino, "Phoneme recognition using phoneme duration estimation based on speaking rate," *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP96-20, pp. 1-6, 1996. (in Japanese)

[7] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," *Proc. Eurospeech'97*, pp. 99-102, 1997.