Syllable-Based Acoustic Modeling for Japanese Spontaneous Speech Recognition

J.Ogata^{*1} *and Y.Ariki*^{*2}

^{*1}National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, JAPAN
^{*2}Department of Electronics and Informatics, Ryukoku University
Seta, Otsu-shi, Shiga, 520-2194, JAPAN
jun.ogata@aist.go.jp

Abstract

We study on a syllable-based acoustic modeling method for Japanese spontaneous speech recognition. Traditionally, mora-based acoustic models have been adopted for Japanese read speech recognition systems. In this paper, syllable-based unit and mora-based unit are clearly distinguished in their definition, and syllables are shown to be more suitable as an acoustic model for Japanese spontaneous speech recognition. In spontaneous speech, a *vowel lengthening* occurs frequently, and recognition accuracy is greatly affected by this phenomena. From this viewpoint, we propose an acoustic modeling technique that explicitly incorporates the *vowel lengthening* in syllable-based HMMs. Experimental results showed that the proposed model could exceed the performance of conventionally used cross-word triphone model and mora-based model in Japanese spontaneous speech recognition task.

1. Introduction

In spontaneous speech recognition, the improvements of acoustic model is very important because the spontaneous speech is affected by several phenomena such as coarticulation and allophone.

As for a unit of acoustic modeling, a phoneme (triphone) has been widely used in most ASR systems. However, since a phoneme unit takes an extremely short time span, integration of spectral and temporal dependencies is difficult [1]. From this viewpoint, recently, the use of acoustic units with a longer duration has been studied in some reports [1]-[3]. In English, it is difficult to employ *purely* syllable-based acoustic modeling because there are more than ten thousand kinds of syllables. On the other hand, the number of Japanese syllable patterns is quite small compared with English so that it can be said that the syllable is suitable for Japanese acoustic modeling. Nakagawa *et al.* has reported the effectiveness of using the syllable-based acoustic model in Japanese read speech recognition. This paper presents a syllable-based acoustical modeling method for Japanese spontaneous speech recognition. In spontaneous speech, a *vowel lengthening* occurs frequently, and recognition accuracy is greatly affected by this phenomena. From this viewpoint, we propose an acoustical modeling technique that implicitly incorporates the *vowel lengthening* in syllable-based HMMs.

2. Syllable-Based Acoustic Modeling

2.1. Definition of syllable and mora in Japanese

In previous works, e.g. [2][3], 'mora' unit has been used for syllable-based acoustic modeling in Japanese, i.e. the structure of the syllable has been defined in the same way as the mora. In this paper, we distinguish clearly these units by their definition and study the acoustic modeling of them. Hereinafter, we describe the definition of the syllable and mora by referring to [4].

In Japanese, one unit of the syllable consists of 1-3 phonemes including vowels: V, VV, CV, CVV, and CVC. In contrast, it is well known that the mora is a partly decomposed unit of the syllables: e.g. CVV is decomposed to CV and V, CVC is decomposed to CV and C. The mora is one to one correspondence with *kana* in Japanese. Fig.1 shows an example of phonetic structure for each acoustic unit. It should be noted that the mora has special characteristic which treats *long vowel*, *long consonant* and *syllabic nasal* as an independent unit. In this paper, we represent *long vowel* as '-', *long consonant* as 'Q' and *syllabic nasal* as 'N'. It can be said that above three phenomena are clear difference between the syllable and the mora in Japanese.

2.2. Syllable-based modeling

In Japanese spontaneous speech or conversational speech, there is an unique nature that *long vowel* or *vowel lengthening* occurs frequently compared with read speech. The *vowel lengthening* is caused by occurrence of some phonetic phenomena (e.g. filled pause, filler

Table 1: Example of phonetic structure of syllable and mora. The syllable and mora boundaries are marked by ‘/’.

word	syllable	mora
se-ta- ‘*sweater*’	se- / ta-	se / - / ta / -
geNgo ‘*language*’	geN / go	ge / N / go
saQka- ‘*soccer*’	saQ / ka-	sa / Q / ka / -

etc.). In traditional mora models, accurate modeling of the *vowel lengthening* is impossible because it is represented by the concatenation of two or three mono-vowels. From this reason, there is some possibility that recognition accuracy is degraded by using mora models in spontaneous speech. Therefore, an explicit acoustic modeling of the phenomena occurring in spontaneous speech is required.

Based on the fact mentioned above, we investigated the frequency rate of three phonetic phenomena occurring in 200 lecture speech syllabic-transcriptions of the CSJ (Corpus of Spontaneous Japanese) monitor version. CSJ is a spontaneous Japanese speech corpus collected under the Japanese national project [5].

As shown in Table 2, the *long vowel* is the most frequent phonetic phenomenon and accounts for about 11% of all syllables in this data. This observation tells us that explicit acoustic modeling of the *vowel lengthened-syllable* (i.e. CVV, VV) is possible in terms of its enough amount of training data.

Table 2: Occurrence rate of the three phonetic phenomena.

	rate (%)
long vowel (vowel lengthening)	11.4%
long consonant	2.9%
syllabic nasal	6.2%

From these viewpoints, we propose a syllable-based acoustic modeling incorporating the *vowel lengthening* for Japanese spontaneous speech recognition. In our syllable model, the *vowel lengthening* itself is included in the immediately previous CV or V and they are regarded as one whole sub-word unit for acoustic modeling. However, we did not treat the modeling of *long consonant* and *syllabic nasal* because of insufficiency of the training data.

In this paper, we evaluate our proposed syllable models through the comparison with traditionally used two models: phoneme (triphone) and mora. Table 3 shows the number of types in our defined phoneme and mora model. The phoneme set is almost same as the model generally used in Japanese LVCSR systems. Also, an example of phonetic representations is shown in Table 4. As can be seen in this example, the *vowel lengthening* is directly modeled in the previous CV or V in our model while the traditionally used mora model cannot treat strictly this phenomenon. In our syllable model, it consisted of totally 244 entries (124 for CV, V and silence to the mora set and 120 for CVV and VV to *vowel-lengthened* CV and V).

Table 3: Number of types in each model.

	phoneme	mora
V	5	5
C	27	-
CV	-	115
long vowel	5	-
long consonant	1	1
syllabic nasal	1	1
silence	2	2
total	41	124

Table 4: Example of phonetic representations. ‘/’ indicates a boundary between models and ‘:’ indicates a *vowel lengthening*.

	se-ta- (‘*sweater*’)
phoneme	s / e: / t / a:
mora	se / e / ta / a
syllable (proposed)	se: / ta:

2.3. State-tying in syllable model

In the proposed syllable model, while the *lengthening vowel* can be modeled strictly compared with the mora model, there is some possibility that the recognition accuracy is degraded by the sparseness of the training data by increasing the model parameters. Therefore, we introduce a state-tying approach for each HMMs [6]. In syllable model, *vowel-lengthened* CV and sole-CV are trained independently as different models. However, the part of the beginning is almost same in terms of acoustic characteristics between the two units. In this paper, heuristically tied-state method is conducted between the CV and the *vowel-lengthened* CV. As shown in Figure 1, 3 states from the beginning are tied in 5 states HMMs between the two units.

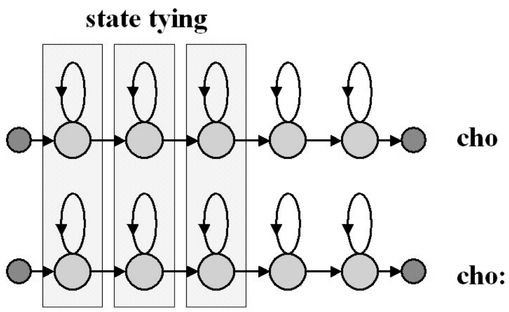

Figure 1: State tying in vowel lengthening model.

3. Experiments

3.1. Experimental set-up

All the experiments were performed on Japanese lecture speech task using CSJ monitor version. As for training data of acoustic modeling, we used 200 male speaker's presentation speech in CSJ. Table 5 shows the experimental conditions for acoustic analysis and HMM. Language model trained by 612 lecture speech transcriptions is available from CSJ monitor version. Here, all the experiments were conducted using the bigram model. Vocabulary size of language model is 19K. As an evaluation data, 400 utterances in 4 male speaker's presentation were used.

Table 5: Acoustic analysis.

Sampling frequency	16kHz
Feature parameter	MFCC (39 dimensions)
Analysis frame length	30ms
Analysis frame shift	10ms
Analysis window	Hamming window

3.2. Recognition results

3.2.1. *Cross-word triphone model*

In the first experiment, we investigated the performance of the baseline triphone models. In triphone-based systems, cross-word context dependency was fully incorporated: i.e. we constructed *cross-word triphone models*. To construct tied-state triphone models, a top-down clustering method based on the phonetic decision tree was used [6]. Figure 2 shows the word error rate (WER) for some number of total states (800, 1500, 2500). As can be seen in this table, the 1500 states model obtained the best performance. In this case, the optimum number of mixtures per state was 40. Hereafter, we used the 1500 states model as a triphone in the following experiments.

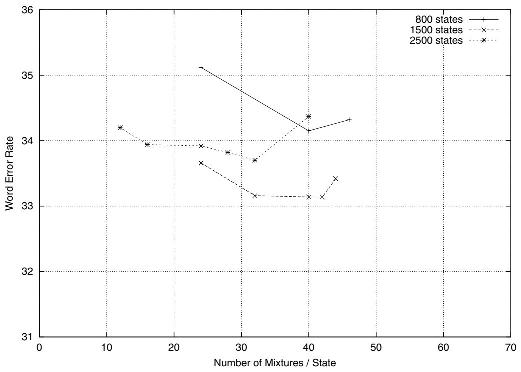

Figure 2: Recognition performance of triphone model.

3.2.2. *Syllable model*

In the next experiment, we carried out the recognition using our proposed syllable models, and compared with triphone and mora. Figure 3 shows the recognition performance for each model. In this figure, ”syllable-m244” and ”syllable-m244-tied” indicate the proposed syllable model and tied-state syllable model respectively. Also, the number of models and states for each model are shown in Table 6.

The proposed syllable model (syllable-m244) obtained almost same performance as the triphone model and outperformed the mora model significantly. According to this result, it can be said that explicit acoustic modeling of the *vowel lengthening* is important in Japanese spontaneous speech recognition. In addition, our tied-state syllable model (syllable-m244-tied) obtained further improvements. As a result, the proposed model outperformed the cross-word triphone model. The fact that the state-tying technique was effective in the syllable model suggests that there was insufficiency of training data at the beginning of CV models.

In Table 7, the WER of the optimum number of mixtures summarized. As can be seen, our tied-state syllable model reduced the WER by 3.3% absolutely compared with the mora model and by 1.8% compared with the cross-word triphone model.

Table 6: Number of models and states.

	#logical	#physical	#states
triphone	27659	2357	1517
mora	124	124	607
syllable	244	244	1192
tied-state syllable	244	244	**847**

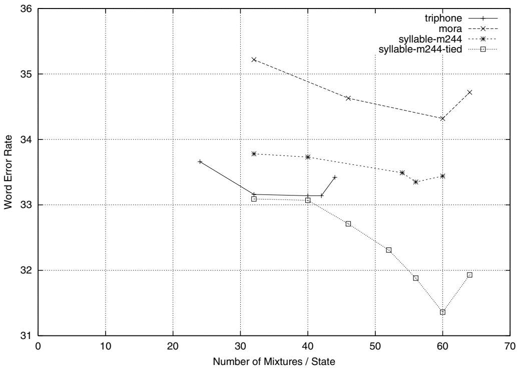

Figure 3: Recognition performance of proposed model.

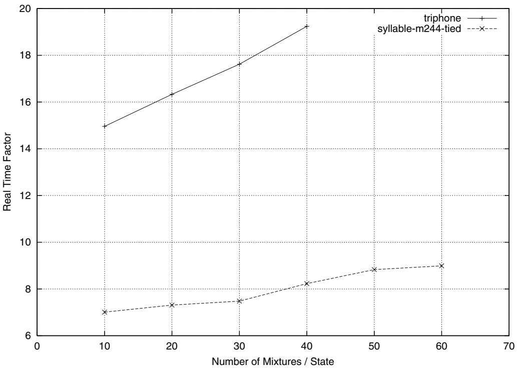

Figure 4: Real time factor of recognition process.

Table 7: WER for optimum number of mixtures.

	#mixtures/state	WER(%)
triphone	40	33.16
mora	60	34.72
syllable	56	33.35
tied-state syllable	60	31.36

3.3. Comparison of computational costs

Finally, we discuss computational costs in the recognition process. Here, we compare the tied-state syllable model with the cross-word triphone model. In this case, especially, we have to consider two factors for efficient decoding: one is a cross-word context dependency, and the other is the number of mixtures per state.

It can be said that computational costs of cross-word context dependency depend heavily on the decoding method. In this paper, a generally used *word-conditioned lexical tree search* [7] was used for decoding algorithm. Figure 4 shows the Real Time Factor (RTF) of recognition for each model. RTF was measured on machines equipped with Intel Pentium III 1GHz CPU, 500MByte main memory. As can be seen, the computational costs of the tied-state syllable model was quite inexpensive compared with the case of the cross-word triphone model. This reason can be explained that cross-word dependency was not incorporated in our syllable or mora model. Consequently, our syllable model is effective in terms of recognition accuracy and efficiency for Japanese spontaneous speech recognition.

4. Conclusion

We have presented a technique of syllable-based acoustic modeling for Japanese spontaneous speech recognition. Based on the analysis of the distribution of three acoustic phenomena, we constructed syllable-based acoustic models that explicitly incorporates the *vowel lengthening*. Experimental results showed that the proposed model could exceed the performance of conventionally used cross-word triphone model and mora-based model in Japanese spontaneous speech recognition task. Furthermore, we confirmed the efficiency of computational costs through the comparison of the cross-word triphone model.

5. References

[1] A.Ganapathiraju *et al.*, “ Syllable-Based Large Vocabulary Continuous Speech Recognition”, *IEEE Trans. on Speech and Audio Processing*,Vol.9, no.4, pp.358-366(2001).

[2] S.Nakagawa *et al.*, “Comparison of Syllable-Based HMMs and Triphone-Based HMMs in Japanese Speech Recognition”, Proc. Int. Workshop on ASRU, pp.197-200 (1999.12)

[3] N.Takahashi, S.Nakagawa, “Syllable Recognition Using Syllable-Segment Statistics and Syllable-Based HMM”, Proc. ICSLP’2002, pp.2633-2636, (2002).

[4] H.Kubozono, “The Mora and Syllable Structure in Japanese: Evidence from Speech Errors”, Language and Speech 32, vol.3, 249-278.

[5] S.Furui *et al.*, “Toward the Realization of Spontaneous Speech Recognition –Introduction of a Japanese Priority Program and Preliminary Results–”, Proc. ICSLP’2000, pp.518-521, (2000).

[6] J.Odell: “The Use of Context in Large Vocabulary Speech Recognition”, Ph.D thesis, University of Cambridge, UK 1995.

[7] S.Ortmanns, H.Ney, X.Aubert: “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition”, Computer Speech and Language,Vol.11, No.1, pp.43-72, 1997.