COMPARISON OF SYLLABLE-BASED HMMS AND TRIPHONE-BASED HMMS IN JAPANESE SPEECH RECOGNITION

Seiichi Nakagawa, Kengo Hanai, Kazumasa Yamamoto, and Nobuaki Minematsu

Department of Information and Computer Sciences
Toyohashi University of Technology, Toyohashi, 441-8580, Japan
{nakagawa,hanai,kyama,mine}@slp.tutics.tut.ac.jp

ABSTRACT

It is well-known that HMMs only of the basic structure cannot capture the correlations among successive frames adequately. In our previous work, to solve this problem, segmental unit HMMs were introduced and their effectiveness was shown. And the integration of Δ cepstrum and $\Delta\Delta$ cepstrum into the segmental unit HMMs was also found to improve the recognition performance in the work. In this paper, firstly, we compared frame-based models and segment-based models. Results showed the effectiveness of the use of segmental features as input vectors for both of syllable-based HMMs and triphone-based HMMs. Secondly, we compared syllable-based HMMs and triphone-based HMMs. Recognition experiments showed that syllable-based HMMs are suitable for Japanese.

1. INTRODUCTION

Hidden Markov Models(HMMs) are a widely used technique for speech recognition. But it is also well-known that the HMMs only of the basic structure have a defect that they cannot adequately represent the temporal correlations between successive feature vectors. In our previous works[1][2], to solve the problem, segmental unit input HMMs were studied, where a feature vector was derived from several successive frames.

As for a unit of acoustic modeling, we have been using syllables, most of which have more than or equal to 2 phonemes, where almost all syllables are a type of consonant-vowel in Japanese. There exist only 114 syllables. Although context dependent *phoneme* models, e.g. triphone, are widely used to reflect the influences of coarticulation on the features' distribution of a focused phoneme, the above characteristics indicate that the influence is already involved in syllable-size HMMs to some extent.

In Chinese, there are many investigations using syllable-based models as the acoustic units for recognition[3][4]. In Chinese language, every word is composed of from one to several characters, and all of these characters are monosyllabic,and the total number of base syllables is only 408(each hearing 4 different tones).

Unlike the Japanese or Chinese language, English has more than ten thousand kinds of syllables, indicating that it is difficult to built syllable-size acoustic models in English. Some researchers, however, found the effectiveness of using the syllables as a unit of acoustic modeling even in English[9][10][11][12]. The first one stated a combination method of the phone-scale and syllable-scale recognizers by merging and rescoring N-best lists. The second built syllable-based HMMs with syllable-level bigram probabilities, and with word- and syllable-level insertion penalties. The third compared syllable system and triphone system for recognition of continuous alphadigit utterances in English. The forth proposed a hybrid system of 200 monosyllabic word models, 623 syllable models and triphone models for others.

In this paper, comparisons of the recognition performance between the syllable-size models and the triphone models, which are a world-wide standard modeling method, were carried out in Japanese.

2. SEGMENTAL UNIT INPUT HMMS

For an input symbol sequence $y = y_1 y_2 \cdots y_T$ (T is the length of the input sequence) and a state sequence $x = x_1 x_2 \cdots x_T$, the output probability of HMM is given by the following equations[2].

$$\begin{aligned} & P(y_1 \cdots y_T) \\ & = \sum_x \prod_i P(y_i | y_1 y_2 \cdots y_{i-2} y_{i-1}, x_1 x_2 \cdots x_{i-1} x_i) \\ & \qquad \times P(x_i | x_1 x_2 \cdots x_{i-1}) \\ & \simeq \sum_x \prod_i P(y_i | y_{i-3} y_{i-2} y_{i-1}, x_{i-1} x_i) P(x_i | x_{i-1}) \quad (1) \\ & = \sum_x \prod_i \frac{P(y_{i-3} y_{i-2} y_{i-1} y_i | x_{i-1} x_i)}{P(y_{i-3} y_{i-2} y_{i-1} | x_{i-1} x_i)} P(x_i | x_{i-1}) \quad (2) \\ & \simeq \sum_x \prod_i P(y_i | y_{i-1}, x_{i-1} x_i) P(x_i | x_{i-1}) \quad (3) \\ & \simeq \sum_x \prod_i P(y_i | x_{i-1} x_i) P(x_i | x_{i-1}) \quad (4) \end{aligned}$$

Eq.(**1**) or Eq.(**2**) is conditional density HMMs of 4-frame segments; Eq.(**3**) is those of 2-frame segments.

The segmental unit input HMM proposed in our previous study[1][2] is obtained by approximating Eq.(**2**), that is, we use only the numerator of Eq.(**2**) :

$$\begin{aligned} & P(y_1 \cdots y_T) \\ & \simeq \sum_x \prod_i P(y_{i-3} y_{i-2} y_{i-1} y_i | x_{i-1} x_i) P(x_i | x_{i-1}) \quad (5) \end{aligned}$$

However, the immediate use of several successive frames as an input vector inevitably increases the dimension of parameters. Therefore, the K–L expansion was used to reduce the dimension in the experiments.

3. REFINEMENTS OF THE MODELS

3.1. Energy

In this study, the term of energy was defined as the 0-th mel-cepstrum coefficient. And its regression coefficients, ΔE and $\Delta\Delta E$, were used as the dynamic feature of the energy. In the experiments, while ΔC and $\Delta\Delta C$ were used assuming no correlation between them, the correlation between ΔE and $\Delta\Delta E$ was estimated to make a covariance matrix.

3.2. Mixture of PDFs

We assumed that the output probability density function (PDF) of $b_{ij}(y)$ could be represented by a mixture of M Gaussian distributions.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \qquad (6)$$

Here, λ_{ijm} is the m-th branching factor at transition from state i to state j . And b_{ijm} is m-th PDF at the transition. They satisfy the following conditions.

$$\sum_{m=1}^{M} \lambda_{ijm} = 1, \qquad \int b_{ijm}(y)dy = 1 \qquad (7)$$

One of our previous works showed that the use of a single Gaussian with a full covariance matrix gave us almost the same performance as that with a mixture of 10 to 16 Gaussians of diagonal covariance matrices[5].

3.3. Context Dependent Models –syllable versus triphone–

(a)syllable

In recent works, to reflect the influences of coarticulation on the features' distribution of a focused phoneme, context dependency is widely introduced into phoneme HMMs. However, since the number of the models is drastically incremented, the context dependency will often result in undesirable phenomena, such as the increase of computational cost and the decrease of precision in estimating the parameters. To avoid these phenomena, we have been using context independent models by a unit of syllables. Almost all Japanese syllables consist of a consonant and a vowel. The above characteristics indicate that the influence between a consonant and a vowel is already involved in syllable-size HMMs to some extent. The number of syllables is 114. In Table **1**, the HMMs (having full covariance matrices) consists of 5 states with 4 output distributions having duration control were used. The duration control is modeled by a discrete probability distribution to improve recognition performance[14]. In Tables **2** and **3**, the HMMs (having diagonal covariance matrices) consisting of 6 states with 5 output distributions having no duration control were used for comparison with the triphone models.

However, if syllables are allowed to be used as a unit of acoustic modeling in Japanese, we can still find a difference between a set of entries of right context and that of left context. Namely, with syllable-size models, only several kinds of phonemes(almost cases are vowels) can be found in the left context. As for the right context, all the phoneme can appear as in English. Therefore, the left-context dependency with syllable-size models is expected to increase the recognition performance without the above mentioned undesirable phenomena.

Based on these considerations, we investigated the syllable-size HMMs with the left-context dependency. In this case, all the entries of left-context was only the following 7 phonemes; /a/, /i/, /u/, /e/, /o/, /N/, and /@/(silence). And the total number of left-context dependent syllable-size models was 908.

(b)triphone

In recent works, to reflect the influences of coarticulation on a focused phoneme, context dependent phoneme models are often used. Triphones have been proposed to take into account both contexts of phoneme, the previous and posterior. In Japanese, almost all successful continuous speech recognition systems also adopt triphone-based HMMs[6][7]. We make the triphone models in order to compare them and syllable models. The number of models is 7921. Training of triphone models was done with HTK, where the mixture number of PDFs is 16 or 32 with diagonal covariance matrices and the model's topology is 5 states with 3 output distributions. The number of states are 3013 or 570, which are reduced by tying. In Japanese, a standard syllable consists of a consonant and a vowel, therefore, a syllable model corresponds to 10 states with 6 output distributions by the concatenation of triphone models.

4. EXPERIMENTS AND RESULTS

4.1. Database and Speech Analysis

Recognition experiments were carried out using the proposed HMMs, i.e. the segmental input models with a mixture of PDFs and/or with context dependency, and conventional HMMs with single Gaussians or those without context dependency. The comparison among these models were done based on continuous syllable recognition tests in a speaker-independent mode.

HMMs were trained by syllable-segmented data from A–J sets (50 sentences each) of ATR speech database (uttered 6 male speakers). For syllable categories which have a small number of data in the database, 216 word data sets were additionally used for the categories. After that, they were retrained with MAP estimation[13] by using following two databases.

- Acoustic Society of Japan database uttered by 30 male speakers(ASJ)(4518 sentences)
- Japan Newspaper Article Sentences database uttered by 125 male speakers(JNAS)(12703 sentences)

The test data consisted of 939 newspaper article sentences spoken by 9 other male speakers.

The analysis conditions are as follows: sampling frequency is 12kHz; Hamming window size is 21.33ms ; frame period is 8ms; and LPC analysis is of the 14th order, feature parameters are LPC 10 mel cepstrum coefficients and energy.

4.2. Syllable-based Recognition Result

We performed the evaluation using each of the following methods of parameter configuration:

(1) C+ΔC+ΔΔC LPC mel-cepstrum coefficients calculated frame by frame are used in addition to their first and second derivations.

(2) C+ΔC+ΔΔC+(ΔE+ΔΔE) (1) + the first and second derivatives of energy

(3) C(K-L)+ΔC+ΔΔC Segmental statistics calculated from 4 successive frames (40 dimensions) with the dimension reduction into 20 by the K-L expansion are used in addition to the first and second derivatives of the LPC mel-cepstrum coefficients.

(4) C(K-L)+ΔC+ΔΔC+(ΔE+ΔΔE) (3) + the first and second derivatives of energy

We investigated the effectiveness of energy, a mixture of PDFs, and context dependent models. The number of syllables used as a unit of acoustic modeling was 114. Table **1** shows the experimental results[8]. "ACC." is the accuracy, "COR." is the correct rate of recognition and "SEG." is the segmentation rate defined as follows:

$$SEG = \frac{N_{total} - N_{ins} - N_{del}}{N_{total}}[\%] \qquad (8)$$

where N_{total} is the number of syllables in the correct syllable sequence, N_{ins} is the number of inserted syllables and N_{del} is the number of deleted syllables.

Throughout the experiments, method (4) with a mixture of 4 PDFs gave us the best recognition rate, i.e. 70.1 % in the accuracy and 78.3 % in the correct rate. The comparison between (3) and (4) shows that the integration of Δ and ΔΔ E improves 3 to 4 % in the accuracy and 2 to 3 % in the correct rate irrespective of the number of PDFs. This clearly indicates the validity of the introduction of the energy-related parameters. In the case of no duration control, the recognition performace is decreased about 2.5% in the accracy.

Difference in the number of PDFs shows the following findings regardless of the method. Accuracy and correct rate in 2-mixture models is increased by approximately 9 % and 2 % respectively from those in 1-mixture models.

Further improvements by about 3 % are found in both the rates in 4-mixture models from those in 2-mixture models. Larger improvements in the accuracy than in the correct rate means that the increment of the number of PDFs especially helps avoiding syllables' being inserted.

Context dependent models in comparison with context independent models with 1-mixture shows 5 to 6 % improvement in the correct rate and only 1 to 4 % improvement in the accuracy. And this is the case of every method. The observed improvements are not as high as expected, which is considered due to lack of training data caused by the increase of the number of model parameters.

4.3. Triphone-based Recognition Results

Next, we compared triphone models with syllable models in continuous syllable recognition experiments. Frame-based and segment-based triphone models were used for continuous syllable recognition by using syllable constraint of phoneme sequences (e.g. a language model of phoneme pairs). Frame-based triphone model was trained with parameters of (2) in Section **4.2**, and segment-based triphone model was with (4). The mixture number of PDFs is 16 with diagonal covariance matrices. The triphone-based HMM doesn't have a duration distribution for each state, that is, only using conventional state transition probabilities. So we also performed experiments of the syllable-based HMMs with no duration control and with diagonal covariance matrices, that is, under the same conditions.

Table 1. Syllable recognition rate for syllable models with duration control[%]

(full covariance matrix)

METHOD	#mix	context	ACC.	COR.	SEG.
C+ΔC+ΔΔC (frame)	1	no	53.1	67.1	82.4
	2	no	58.1	69.7	85.1
	4	no	61.6	72.1	86.3
	1	yes	55.0	73.1	79.2
C+ΔC+ΔΔC +(ΔE+ΔΔE) (frame)	1	no	54.8	70.3	81.6
	2	no	64.7	73.8	87.8
	4	no	67.3	75.7	88.7
	1	yes	59.0	76.4	80.7
C(K–L)+ΔC +ΔΔC (segment)	1	no	57.0	71.0	83.3
	2	no	63.3	73.5	86.5
	4	no	66.2	76.0	88.5
	1	yes	57.8	76.0	79.7
C(K–L)+ΔC +ΔΔC +(ΔE+ΔΔE) (segment)	1	no	57.6	72.3	82.8
	2	no	67.0	75.6	88.4
	4	no	70.1	78.3	89.1
	1	yes	59.9	77.4	80.6

Table 2. Syllable recognition rate for triphone models and syllable models without duration control[%]

(diagonal covariance matrix)

METHOD	#mix	#states	ACC.	COR.	SEG.
triphone (frame)	16	3013	64.2	79.9	82.0
	16	570	65.2	77.7	84.2
	32	570	66.6	78.5	84.2
triphone (segment)	16	3013	64.9	81.3	81.3
	16	570	65.8	78.8	83.7
	32	570	67.6	79.9	84.3
syllable (frame)	16	570	66.5	72.9	89.8
	32	570	68.9	75.6	89.3
syllable (segment)	16	570	68.5	75.8	89.5
	32	570	70.5	78.2	89.3

Table **2** shows results of the experiments. The comparison between segment-based models and frame-based models shows that the former improved 0.7 ∼ 1.0% in the accuracy and 1.1 ∼ 1.4% in the correct rate, in average. This result indicates the effectiveness of the use of segmental unit input HMMs.

4.4. Discussions

Although frame-based syllable models are worse in the correct rate than frame-based triphone models, they are better in the accuracy. The reason is caused by that the insertion error rate is larger for triphone models than for syllable models. The insertion error rate for triphone models is approximately 13%. On the other hand, that for syllable models is approximately 7%. This may be caused that the triphone model is a shorter model unit than syllable models.

Table 3. Number of free-parameters for triphone and syllable models

METHOD	Num. of models	Num. of states	Num. of mix	Num. of estimated parameters
triphone (frame)	7921	3013	16	3133520
	7921	570	16	592800
	7921	570	32	1185600
triphone (segment)	7921	3013	16	4097680
	7921	570	16	775200
	7921	570	32	1550400
syllable (frame)	114	570	16	593370
	114	570	32	1186170
syllable (segment)	114	570	16	783180
	114	570	32	1550970

And, segment-based syllable models also show better results in the accuracy than segment-based triphone models, but they are worse in the correct rate. This is the same reason. The comparison between segment-based models and frame-base models shows that the former improves about 2% in the accuracy and approximately about 3% in the correct rate in syllable models, but only about 1% in both the accuracy and correct rate in triphone models. This clearly indicates that segmental unit input HMMs are more effective for syllable models than for triphone models. This is considered so that segment-based syllable models deal with the context dependency by means of using CV(syllable) and segmental unit input HMMs. These results led us to confirm that syllables are an appropriate unit of acoustic modeling in speech recognition, especially, for Japanese.

Table **3** shows the number of the free-parameters for triphone and syllable models, respectively (except for state transition probabilities). In these experiments, the triphone models reduced parameters using a tying technique, but not for the syllable models. Although two models have almost same recognition performance, the number of models of the syllable-based model is quite smaller than of the triphone-based models. It shows that the syllable model is a compact representation of the acoustic model for Japanese.

5. CONCLUSION

In this paper, we compared syllable-based and triphone-based HMMs. Results of continuous syllable recognition experiments show that, while the syllable models are worse in the correct rate than triphone models for both frame-based and segment-based HMMs, they are better in the accuracy. We believe that syllables are an appropriate unit of acoustic modeling in speech recognition for Japanese. As future plans, we should compare them for a large vocabulary continuous speech recognition systems, i.e., word recognition rates.

Acknowledgement

This work has been supported by CREST(Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation(JST).

References

[1] S.Nakagawa, Y.Hirata and Y. Ono, "Syllable recognition by hidden Markov models using fixed-length segmental statistics," Trans. Inst. Elect. Inform. Comm., Vol. J75-DII , No.5, pp.843-851 (1992, in Japanese).

[2] K.Yamamoto, S.Nakagawa, "Comparative evaluation of segmental unit input HMM and conditional density HMM," Proc. EUROSPEECH'95, pp.1615–1618 (1995)

[3] L.S. Lee, C.Y. Tseng, K.J. Chen, I.J. Hung, M.Y. Lee, L.F.Chien, Y. Lee, R. Lyu, H.M.Wang, Y.C. Wu, T.S.Lin, H,Y, Gu, C.P. nee, C.Y. Liao, Y.J. Yang, Y.C. Chang, R.C. Yang, "Golden Mandarin(II)-An Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," Proc. ICASSP-93, pp.503–506 (1993)

[4] Z.Wang, J.Wu, J.Guo, "Methods towards the very large vocabulary Chinese speech recognition," Proc. EUROSPEECH'95, pp.215–218 (1995)

[5] S.Nakagawa, L.Zhao, H.Suzuki, "A Comparative Study of Output Probability Functions in HMMs," IEICE Trans., Vol.E78-D, No.6, pp.698–704 (1995)

[6] J.Takami, S.Sagayama, "A successive state splitting algorithm for efficient allophone modeling," Proc. ICASSP-92, pp.573–576 (1992)

[7] T.Matsuoka, K.Ohtsuki, T.Mori, S.Furui, K.Shirai, "Japanese Large-Vocabulary Continuous Speech Recognition using a Business-Newspaper Corpus," Proc. ICASSP-97, pp.1803–1806 (1997)

[8] K.Hanai, K.Yamamoto, N.Minematsu, S.Nakagawa, "Continuous speech recognition using segmental unit input HMMs with a mixture of probability density functions and context dependency," Proc. ICSLP-98, pp.2935–2938 (1998)

[9] S.Wu, B.Kingsbarg, N.Morgan, S.Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," Proc. ICASSP-98, pp.721–724 (1998)

[10] R.James, S.Downey, J.S.Mason, "Continuous speech recognition using syllables," Proc. EUROSPEECH'97, pp.1171–1174 (1997)

[11] J.Hamaker, A.Ganapathiraju, J.Picone, J.J.Godfrey, "Advances in alphadigit recognition using syllables," Proc. ICASSP-98, pp.421–424 (1998)

[12] A.Ganapathiraju, V.Goel, J.Picone, A.Corrada, G.Doddington, K.Kirchhoff, M.Ordowski and B.Wheatley, "Syllable – a promising recognition unit for LVCSR," Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp.207–214 (1997)

[13] Y.Tsurumi, S.Nakagawa, "An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation," Proc. ICSLP-94, pp.431–434 (1994)

[14] Y.Hashimoto, Y.Hirata, S.Nakagawa, "A study on Japanese phoneme recognition using continuous observation hidden Markov models," Tech. Report of IEICE, SP89-48 (1989, in Japanese)