CONTINUOUS SPEECH RECOGNITION USING SEGMENTAL UNIT INPUT HMMs WITH A MIXTURE OF PROBABILITY DENSITY FUNCTIONS AND CONTEXT DEPENDENCY

Kengo HANAI, Kazumasa YAMAMOTO, Nobuaki MINEMATSU, and Seiichi NAKAGAWA
{*hanai, kyama, mine, nakagawa*}*@slp.tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Technology,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441-8580 JAPAN

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, we investigated further refinements of the models by using a mixture of PDFs and/or context dependency, where, for a given syllable, only a preceding vowel was treated as the context information. Recognition experiments showed that the accuracy rate was improved by 23 %, which clearly indicates the effectiveness of the refinements examined in this paper. The proposed syllable-based HMM outperformed a triphone model.

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[3][6], to solve the problem, segmental unit input HMMs were studied, where a feature vector was derived from several successive frames. While the recognition performance was improved by using the segmental HMMs, the integration of Δ cepstrum (ΔC) and $\Delta\Delta$ cepstrum ($\Delta\Delta C$) into the segmental HMMs further increased the recognition rates.

The use of the segmental statistics as acoustic features provides us with new problems, which are specific to the segmental HMMs. When using the features immediately instead of the frame-based parameters, since the dimension of the feature parameters is inevitably incremented, it results in increasing computational cost and decreasing precision in the estimation of covariance matrices. To avoid these, the dimension of the feature parameters should be reduced by some methods without degrading modeling capabilities. Bocchieri et al.[2] and Brown[7] applied the principal component analysis to adjoined 2 frames in DTW-based and HMM-based recognition respectively. The discriminant analysis was also applied to adjoined 2 frames in HMM by Bahl et al.[8]. In our previous studies[3][6], both the methods were used for 4 successive frames and the resulting segmental statistics were introduced into

In our laboratory, speaker independent HMMs with a multivariate Gaussian of a full-covariance matrix characterizing the features' distribution in a state have been studied. However, the variances of some elements were estimated to be so large that clear separation could not be observed among several classes. To cope with these phenomena, in this paper, a mixture of multivariate Gaussians with full-covariance matrices was introduced, where mixtures of 2 and 4 PDFs were examined.

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. There exist only 114 syllables in Japanese. 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 this paper, however, aiming at more precise modeling of the influence, the information of a preceding phoneme, which is a vowel in most of the cases, was introduced as context. And the context dependent syllable models were built using the segmental statistics.

Unlike the Japanese language, English has more than one 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[10][11][12]. Then 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[6].

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

$$\simeq \sum_x \prod_i \frac{P(y_{i-1}y_i|x_{i-1}x_i)}{P(y_{i-1}|x_{i-1}x_i)} P(x_i|x_{i-1}) \quad (3)$$

$$= \sum_x \prod_i P(y_i|y_{i-1}, x_{i-1}x_i) P(x_i|x_{i-1}) \quad (4)$$

$$\simeq \sum_x \prod_i P(y_{i-1}y_i|x_{i-1}x_i) P(x_i|x_{i-1}) \quad (5)$$

$$\simeq \sum_x \prod_i P(y_i|x_{i-1}x_i) P(x_i|x_{i-1}) \quad (6)$$

Eq.(**1**) or Eq.(**2**) is conditional density HMMs of 4-frame segments; Eq.(**3**) or Eq.(**4**) is those of 2-frame segments; and Eq.(**5**) is a segmental unit input HMM of 2-frame segments[7]. Eq.(**4**) was modified as follows :

$$= \sum_x \prod_i P(y_i|y_{i-\tau(i)}, x_{i-1}x_i) P(x_i|x_{i-1}), \quad (7)$$

where $\tau(i)$ is selected to maximize likelihood of the observation sequences and to determine dynamically temporal dependence[4]

The segmental unit input HMM proposed in our previous study[3][6] is obtained by approximating Eq.(**2**), that is, we use only the numerator of Eq.(**2**) :

$$P(y_1 \cdots y_T) \simeq \sum_x \prod_i P(y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i) P(x_i|x_{i-1}) \quad (8)$$

As mentioned in Section **1.**, the immediate use of several successive frames as an input vector inevitably increases the dimension of parameters. Then, 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. These terms are derived in the following formula producing the r-th regression coefficient[9];

$$R_{rk}(t, T, \Delta T, N) = \frac{\sum_{X=1}^{N} P_r(X, N) C_k \left[t + \left[\frac{X-1}{N-1} - \frac{1}{2} \right] (T - \Delta T) \right]}{\sum_{X=1}^{N} P_r^2(X, N)} \quad (9)$$

where $C_k(t)$ is the k-th cepstrum coefficient at time t, T is time width for computing regression coefficients, ΔT is a frame period of speech analysis, and N is the number of frames for computing regression coefficients. Weighting function $P_r(X, N)$ is represented as

$$P_1(X, N) = X \quad (10)$$

$$P_2(X, N) = X^2 - \frac{1}{12}(N^2 - 1). \quad (11)$$

Here, $P_1(X, N)$ and $P_2(X, N)$ correspond to a linear and a quadratic regression coefficient respectively. 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 an addition of M Gaussian distributions.

$$b_{ij}(y) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(y) \quad (12)$$

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.

$$\sum_{m=1}^{M} \lambda_{ijm} = 1, \qquad \int b_{ijm}(y) dy = 1 \quad (13)$$

Even with a mixture of PDFs, the equations for re-estimating the parameters are obtained by the Baum-Welch algorithm as in the case of the single Gaussian HMM, which are shown in the below.

$$\hat{\lambda}_{ijm} = \frac{\sum_t \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t)}{\sum_t \alpha a_{ij} b_{ij}(y_t) \beta(j, t)} \quad (14)$$

$$\hat{\mu}_{ijm} = \frac{\sum_t \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t) y_t}{\sum_t \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t)} \quad (15)$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_t \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t) A}{\sum_t \alpha(i, t-1) a_{ij} \lambda_{ijm} b_{ijm}(y_t) \beta(j, t)} \quad (16)$$

$$A = (y_t - \mu_{ijm})(y_t - \mu_{ijm})^t$$

In the experiments, full covariance matrices were used. 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 matrices[5].

3.3. Context Dependent Models

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.

In English, to make context dependent models, all the kinds of phoneme should be considered as right/left context irrespective of a unit of acoustic modeling, i.e. phonemes/syllables. And this is the case with *phoneme* models in Japanese. However, if syllables are allowed to be used as a unit of acoustic modeling in Japanese, we can find a great 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, 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 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.

4. EXPERIMENTS AND RESULTS

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. Comparison among these models were done based on continuous syllable recognition tests in a speaker-independent mode.

In these experiments, continuous HMMs (having full covariance matirices) with 5 states (4 output distributions) having duration control were used. They 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 one of following three 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)
- both ASJ and JNAS(ASJ+JNAS)

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. The number of syllables used as a unit of acoustic modeling was 114.

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

Table 1: Comparison of 3 database[%]
speaker-independent mode, average rate over 9 male speakers

METHOD	DATABASE	ACC.	COR.	SEG.
C+ΔC+ΔΔC Mix1	ASJ	48.7	63.1	81.7
	JNAS	53.9	67.7	82.7
	ASJ+JNAS	53.1	67.1	82.4
C(K-L)+ΔC+ΔΔC +(ΔE+ΔΔE) Mix4	ASJ	61.8	71.5	87.1
	JNAS	69.9	77.4	89.3
	ASJ+JNAS	70.1	78.3	89.1

First, we compared the 3 databases in terms of the recognition performance separately indicated by each databases's models. Table **1** shows the result of experiments. "ACC." is the accuracy rate of recognition, "COR." is the correct rate of recognition, and "SEG." is the segmentation rate defined as follows:

$$SEG = \frac{N_{total} - N_{ins} - N_{del}}{N_{total}}[\%] \qquad (17)$$

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.

In comparison of ASJ with JNAS, the models built with JNAS give us 5 to 8% improvement in accuracy and correct rates. This is considered due to JNAS has efficient data for the parameter estimation. As shown in Section 4, JNAS has three times as many sentences as ASJ has.

Table 2: Continuous Syllable Recognition [%]
speaker-independent mode, average rate over 9 male speakers (ASJ+JNAS)

METHOD	#mix	context	ACC.	COR.	SEG.
(1) C+ΔC+ΔΔC	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
(2) C+ΔC+ΔΔC +(ΔE+ΔΔE)	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
(3) C(K–L)+ΔC +ΔΔC	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
(4) C(K–L)+ΔC +ΔΔC +(ΔE+ΔΔE)	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

Next, we investigated the effectiveness of energy, a mixture of PDFs, and context dependent models. In the experiments, we used ASJ+JNAS as training database. Table **2** shows the experimental results. Throughout the experiments, method (4) with a mixture of 4 PDFs gave us the best recognition rate, i.e. 70.1 % in accuracy rate and 78.3 % in correct rate. Comparison between (3) and (4) irrespective of the number of PDFs shows that the integration of Δ and ΔΔ E improves 3 to 4 % in accuracy rate and 2 to 3 % in correct rate. This clearly indicates the validity of the introduction of the energy-related parameters.

Difference in the number of PDFs shows the following findings regardless of the method. Accuracy rate 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 accuracy rate than in correct rate means that the increment of the number of PDFs especially helps avoiding syllables' being inserted.

Comparison between context dependent models and context independent models with 1-mixture shows 5 to 6 % improvement in correct rate and only 1 % improvement in accuracy rate. 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.

Finally, we compared triphone models with syllable models in continuous syllable recognition experiments. Although the analysis condition was the same between the above two models, frame-based and segment-based models were separately built as syllable-size models, i.e. syllable constraint of phoneme sequences. Therefore, we examined three modeling configurations listed in Table **3** and Table **4**. While triphone models and frame-based syllable models were trained with parameters of (2), segment-based syllable models were with (4). And the mixture number of PDFs is 4 in both the syllable models. Training of triphone models was done with HTK, where the mixture number of PDFs is 16 with diagonal covariance matrices and the model's topology is 5 states with 3 output distributions.

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

METHOD	Num. of models	Num. of states	Num. of estimated parameters	duration control
triphone (frame)	7921	3013	3133520	no
syllable (frame)	114	570	1927968	yes
syllable (segment)	114	570	3295968	yes

Table 4: Syllable recognition rate for triphone and syllable moels[%]
speaker-independent mode, average rate over 9 male speakers

METHOD	ACC.	COR.	substitution	insertion	deletion	SEG.
triphone (frame)	64.2	79.9	17.7	15.7	2.4	82.0
syllable (frame)	67.3	75.7	21.4	8.3	3.0	88.7
syllable (segment)	70.1	78.3	19.1	8.2	2.7	89.1

Table **4** shows results of the experiments. Although frame-based syllable models are worse in correct rate than triphone models, they are better in accuracy rate. On the other hand, segment-based syllable models also show better results in accuracy rate than triphone models with almost the same correct rate as that of triphone models. However, we should notice that the syllable-based HMM has a duration distribution for each state, but not for the triphone-based HMM. These results led us to confirm that syllables are an appropriate unit of acoustic modeling in speech recognition.

5. CONCLUSION

In this paper, we examined several refinements of the syllable-size acoustic models by using a mixture of PDFs and/or context dependency, where only a preceding vowel is used as context. Results of continuous syllable recognition experiments show that, while the use of energy in the form of its derivatives and that of a mixture of PDFs improves accuracy and correct rates, context dependent modeling unexpectedly shows only a little improvement of accuracy rate. This is considered due to lack of training data for a increased number of models. As future plans, we will introduce other training methods, ex. MCE, to the segmental HMMs and verify their validity.

ACKNOWLEDGMENT

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

REFERENCES

1. E.Tsuboka, Y.Takada, H.Wakita: Neural predictive hidden Markov model, Proc. ICSLP, Vol.2, pp.1341–1344 (1990)
2. E.L.Bocchieri, G.R.Doddington: Speaker-independent digit recognition with reference frame-specific distance measures, ICASSP, pp.2699–2672, (1986).
3. 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).
4. J.Ming, P.O'Boyle, J.Smith: An HMM with optimized segment-dependent observations for speech recognition,EUROSPEECH'95, vol.II, pp.1475–1478, (1995).
5. Seiichi Nakagawa, Li Zhao, Hideyuki Suzuki:A Comparative Study of Output Probability Functions in HMMs, IEICE Trans., Vol.E78-D, No.6, pp.698–704, (1995)
6. K.Yamamoto, S.Nakagawa: Comparative evaluation of segmental unit input HMM and conditional density HMM, Proc. Eurospeech, pp.1615–1618 (1995)
7. P.F.Brown: The acoustic-modeling problem in automatic speech recognition, Ph.D. thesis, Carnegie-Mellon University (1987).
8. L.R.Bahl, P.F.Brown, P.V.de Souza, R.L.Mercer: Speech recognition with continuous parameter hidden Markov models, ICASSP, vol.I, pp.40–43 (1988).
9. T.H.Applebaum, B.A.Hanson: Tradeoffs in the design of regression features for word recognition, Proc. EUROSPEECH, pp.1203–1206 (1991)
10. S.Wu, B.Kingsbarg, N.Morgan, S.Greenberg: Incorporating information from syllable-length time scales into automatic speech recognition,ICASSP, pp.721–724 (1998).
11. R.James, S.Downey, J.S.Mason: Continuous speech recognition using syllables, EUROSPEECH,pp.1171–1174 (1997)
12. J.Hamaker, A.Ganapathiraju, J.Picone, J.J.Godfrey: Advances in alphadigit recognition using syllables, ICASSP, pp.421–424 (1998)
13. Y.Tsurumi, S.Nakagawa: An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation, Proc. ICSLP, pp.431–434 (1994)