STUDIA PHONOLOGICA XXVII (1993)

Phoneme Recognition Improvement in Concatenated HMM Training

Yasuo ARIKI and Shuji DOSHITA

ABSTRACT

Conventional concatenated training of phoneme HMMs can approximate the output probability structure of the HMMs. However, it causes an essential problem that the output probability structure of the HMMs becomes blur, because HMMs are trained over the whole speech data of a given sentence. This paper describes a method to solve this problem and to make the output probability structure sharp, by restricting the training section to the proper section associated with the phoneme, still keeping the advantages of the conventional concatenated HMMs training. Four kinds to experiments were carried out and the proposed method showed the 5.6% improvement of the recognition rate of the phonemes included in the continuously spoken sentences, compared to the conventional concatenated phoneme HMM training.

1. INTRODUCTION

A concatenated HMM training method was proposed by K.F. Lee, to train phoneme HMMs without hand-labels from a large speech database, in the SPHINX project at CMU in 1988 [1]. In the method, initial phoneme HMMs are concatenated into a sentence HMM according to phoneme sequence of a given sentence. After the sentence HMM is trained to the corresponding speech data, the phoneme HMMs are separated. This process is repeated to all the sentences in database until it is converged. This training method has the following two advantages.

(1) Phoneme context is incorporated in the trained HMMs.

(2) Errors and ambiguity in hand-labeling are freed.

It is true that the concatenated HMM training approximates output probability structure of the HMMs because the forward & backward probability becomes high around the proper section associated with the phoneme. However, it causes an essential problem that the output probability structure of the HMMs becomes blur, becauese each HMM is trained over the whole speech data of the sentence. Therefore a little but unnecessary probability information is gathered far from the proper section and then taken into the output probability structure of the HMMs.

This paper proposes a method to solve this problem and to make the output probability structure sharp, by restricting the training section to the proper section

1) Yasuo ARIKI (有木康雄): Professor, Department of Electronics and Informatics, Faculty of Science and Technology, Ryukoku University

2) Shuji DOSHITA (堂下修司): Professor, Department of Information Science, Faculty of Engineering, Kyoto University

associated with the phoneme, still keeping the advantages (1) and (2) of the conventional concatenated HMM training method.

Phoneme HMM training on the isolated phoneme data gathered using the hand-labels (phoneme names and their time sections) can be regarded as supervised learning. On the other hand, the conventional concatenated HMM training is regarded as semi-supervised learning, because only the phoneme sequence included in the speech data is given as the teaching signal. The proposing method is also semi-supervised learning, but the most prospective training section for each phoneme is automatically predicted and used as the teaching signal.

2. CONCATENATED TRAINING (CT)

2.1 Formalization

In the conventional concatenated training, initial or partially trained phoneme HMMs are concatenated into a sentence HMM according to phoneme sequence of a given sentence. Using the sentence HMM, forward and backward probabilities $\alpha_t(i)$, $\beta_t(i)$ are computed all over the time t at each state i as follows.

$$\alpha_t(i) = \sum_j \alpha_{t-1}(j) a_{ji} b_i(o_t) \tag{1}$$

$$\beta_t(i) = \sum_j a_{ij} \beta_{t+1}(j) b_j(o_{t+1}) \tag{2}$$

where a_{ij} and $b_i(o_t)$ are the state transition probability from state i to j, and output probability to produce o_t at the state i respectively. After computing $\alpha_t(i)$ and $\beta_t(i)$, the existence probability $\gamma_t(i)$ of state i at time t is computed as follows:

$$\gamma_t(i) = \alpha_t(i) \beta_t(i) / Pr(O) \tag{3}$$

where $Pr(O)$ is the probability of the speech data. Then the output probability of the quantized vector v_l at the state i is computed based on the following expression.

$$b_i(l) = \frac{\sum_{t \in o_t = v_l} \gamma_t(i)}{\sum_t \gamma_t(i)} \tag{4}$$

The output probability $b_i(l)$ is computed at all the state.

2.2 Problems

In the conventional concatenated training, the probabilities $b_i(l)$ of the phoneme

HMMs are computed over the whole speech data as shown by $\sum_{t}$ or $\sum_{t \in o_t = v_l}$ in the expression (4). This causes the following two problems.

(1) It is true that the forward probability $\alpha_t(i)$ and the backward probability $\beta_t(i)$ become high around the proper section associated with the state i of the phoneme, if the initial HMMs are sophisticatedly constructed by using the small amount of hand-lables. Therefore the expression (4) seems to be correct as the output probability. However, if the initial HMMs are weak as in the case where they are all set to be same, $\gamma_t(i)$ is no more high around the proper section, but almost same all over the time. This does not guarantee that the output probability $b_i(l)$ is correctly computed by the expression (4).

(2) Since it sums the $\gamma_t(i)$ all over the time at the state i, a little but unnecessary probability information is gathered far from the proper section and then taken into the output probability $b_i(l)$ of the HMMs. This causes the output probability of the HMMs becomes blur.

To solve these problems, we propose the method to improve the probabilities $\alpha_t(i)$, $\beta_t(i)$ and $\gamma_t(i)$ by restricting the time section for computing their probabilities. This method can be regarded as an equivalent which can make the initial HMMs virtually sophisticated by giving the proper training section as the teaching signal.

3. Section Restricted Concatenated Training (SRCT)

3.1 Concept and Formalization

Fig. 1 shows a concept how to restrict the training section of each phoenme in the concatenated HMM to the proper section on the input speech data. In the figure, the horizontal line corresponds to input speech data and the vertical line to the concatenated HMM. The /r/,/e/,/N/……, /uu/ is phoneme sequence and means "concatenated training" in English. The black stripes are the proper phoneme sections, estimated by Viterbi segmentation algorithm. The training section is produced by putting marginal regions shown as the gray stripes to both side of the proper section. The size of the marginal regions are determined as a certain fixed rate of the proper phoneme section irrespective of phoneme types. The marginal regions serve as to absorb ambiguity of the Viterbi segmentation and to take the phoneme context. The probabilities $\alpha_t(i)$, $\beta_t(i)$ and $\gamma_t(i)$ shown in the expression (1), (2) and (3) are computed within the training section over the whole speech data. This is different from segmenting the input speech data into phoneme training section and carrying out the isolated phoneme HMM training because the probabilities $\alpha_t(i)$ and $\beta_t(i)$ are computed over the whole speech data in our method. This guarantees that each poneme HMM is trained within the sentence to increase the sentence probability, being different from local and individual phoneme training. We call the conventional concatenated

training "CT" and the proposing concatenated training with restricted training section "SRCT" hereafter.

3.2 Altorithm

Fig. 2 shows the block diagram of the proposing method and the procedure flow is summarized as follows:

(1) Conventional concatenated HMM training is carried out using all the speech data.

(2) Viterbi segmentation is carried out for all the speech data, using the phoneme HMM trained in the process (1).

(3) Training sections are produced by putting the marginal regions to the proper sections segmented in the process (2).

(4) Phoneme HMMs are newly trained through concatenated HMM training. In this process, phoneme HMMs are initiated and the forward & backward probability is computed within the training section produced in the process (3).

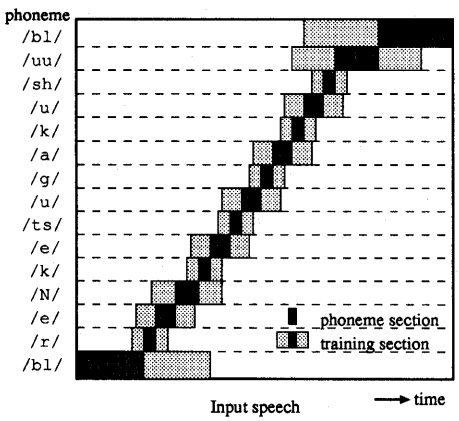

Fig. 1. Concept of concatenated HMM training with restricted training section

4. EXPERIMENTAL RESULTS

4.1 Experimental Condition

Four kinds of phoneme recognition experiments were carred out as well as word recognition experiment for speech data spoken by one person to investigate the effectiveness of the concatenated HMM training within the restricted training section. Conditions for acoustic analysis, vector quantization and HMM are listed in Table 1.

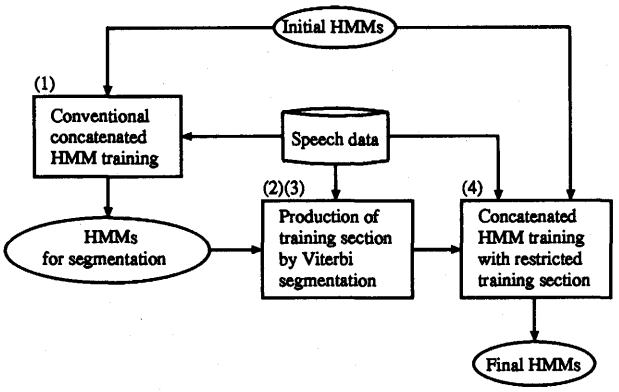

Fig. 2. Block diagram of concatenated HMM training by restricting training section

4.2 Phonmeme HMM Training within Words

[Experiment 1]

The purpose of this experiment is to show effectiveness of restricting the training section. We used hand-labels to restrict the training section at first instead of Viterbi segmentation to investigate the upper limit of SRCT. The speech database used was 5240 ATR important words.

Japanese phoneme HMMs were trained using 2620 words taken from even number of 5240 ATR important words. The number of phoneme HMMs to be trained was 46. The phonemes included in the remaining half words were used for recognition. In the recognition, the phoneme data were extracted from the words using the phoneme hand-labels and then recognized using the trained HMMs.

Fig. 3 shows the recognition result of phonemes included in the 2620 ATR important words. The horizontal line is the iteration number of the concatenated HMM training. The dotted line "Conventional" shows the recognition rate of phonemes whose HMMs were trained by the conventional CT. The solid line "Label 1" shows the recognition rate of phonemes whose HMMs were trained by the phoneme data gathered from the words using the hand-labels. Hereafter we call it isolated phoneme training. The highest recognition rate was 77.1%. The dotted line "Lable 2" shows the recognition rate of phonemes whose HMMs were trained by SRCT, where hand-labels were used to restrict the training section instead of Viterbi segmentation in Fig. 2. The highest recognition rate was 76.6%.

Fig. 3 indicates that SRCT is almost equivalent to isolated phoneme training if the restriction of the training section is carred out by hand-labels. Our goal is to give the training system an ability to extract automatically the correct training section as teaching signal.

[Experiment 2]

The purpose of this experiment is to determine the best marginal region of the

Table 1. Conditions for acoustic analysis(A,A), vector quantisation(VQ) and Hidden Markov Modeling(HMM).

A	Sampling frequency	12KHz
A	High-pass filter	$1-0.97z^{-1}$
A	Feature parameter	LPC cepstrum(16th)
A	Frame length	20ms
A	Frame shift	5ms
A	Window type	Hamming window
V	Codebook size	256 codewords
Q	Distance measure	Euclidean distance
H	Number of states	5 states 3 loops
M	Learning method	Concatenated training
M	Type	Left to right discrete HMM

training section in SRCT. Phoneme HMMs are trained by changing the size of the marginal region in SRCT. In the recognition, the phoneme data were extracted from the words using the phoneme hand-labels and then recognized using the trained HMMs. The speech data is same as in the Experiment 1.

The result is shown in Fig. 4. In the figure, the horizontal line is the iteration number of the conventional HMM training in the process (1) described in 3.2 and the vertical is the phoneme recognition rate. The $\times 0$, $\times 1$, $\times 1/2$, $\times 1/3$ and $\times 1/4$ are the ratio of the marginal region to the proper section obtained by Viterbi segmentation in Fig. 2; $\times 0$ and $\times 1/2$ mean no marginal region and the half ratio of the marginal region to the proper section respectively. The dotted line with "origin" indicates the conventional CT. From the figure, it can be concluded that $\times 1/2$ (half

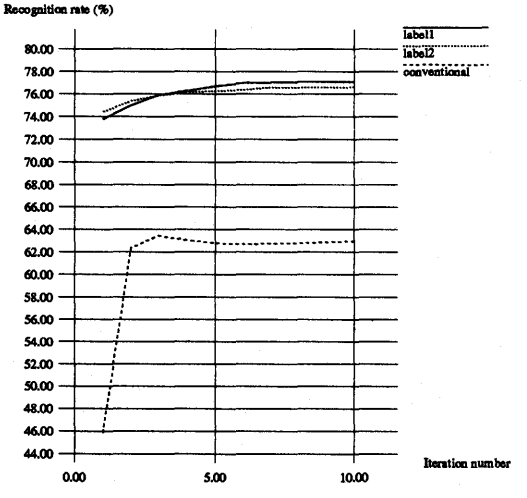

Fig. 3. Recognition result of phonemes in words using hand-labels

size of the proper section) is most effective in this experiment. The ×0 (no marginal region) and the ×1 (same size as the proper section) are not effective. This can be explained that the longer marginal region is required to absorb the phoneme boundary ambiguity caused by hand-labeling, but too long a marginal region takes the unnecessary probability information. The CT is the case where the marginal region is longest, and the recognition rate is worst due to the same reason. As the balance between absorption of boundary ambiguity and exclusion of unnecessary probability information, ×1/2 marginal region achieved the best recognition rate. The highest recognition rate is 68.1% which is 5.3% up compared to the conventional HMM training.

[Experiment 3]

The purpose of this experiment is to investigate the dependency of SRCT on evaluation methods. Phoneme HMMs were trained in the same way as the Experiment 2. The phoneme recognition was carried out without phoneme hand-labels [2]. The recognition result is shown in Fig. 5. It can be concluded that the ×0 or ×1/4 is most effective and the ×1 is not effective. This can be explained that a very small size of the marginal region is enough because there is no ambiguity on the phoneme boundaries created by Viterbi segmentation for the short speech data like the words. Exclusion of unnecessary probability information is effectively achieved by the very small size of the marginal region. The CT is the case where the marginal region is longest, and the recognition rate is worst due to the same reason. The highest recognition rate is 71.9% which is 2.1% up compared to the conventional HMM training.

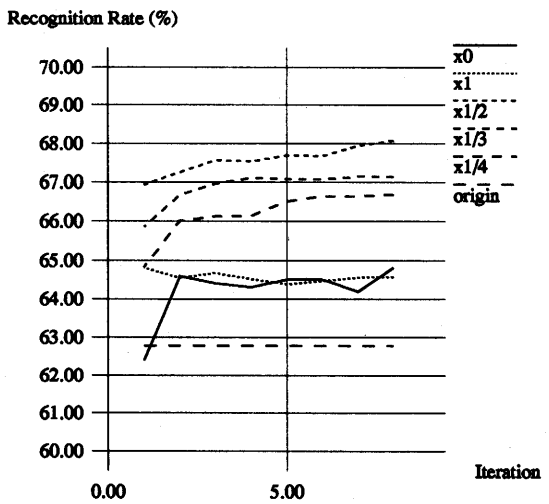

Fig. 4. Recognition result of phonemes in words by the evaluation using hand-labels

The recognition rates by CT are different between two evaluation methods of the experiment 1 and 2. The evaluation without hand-labels shows high recognition rate. This can be explained that CT and the evaluation without hand-labels are both segmentation free so that they are suitable each other. On the other hand the evaluation by hand-labels (Fig. 4) is sensitive to the phoneme segmentation boundary. SRCT in the evaluation by the hand-labels shows the higher improvement (5.3% to 2.1%) than in the evaluation without hand-labels. This can be explained that SRCT adjust the boundary by restricting the training section so that SRCT is more effective in the evaluation using the hand-labels.

4.3 Phoneme HMM Training within Sentences

[Experiment 4]

The purpose of this experiment is to determine the best size of the marginal region included in the training section in SRCT of the phonemes included within sentences.

Japanese phoneme HMMs were trained using 75 sentences taken from an even number of 150 sentences in ASJ speech database. The number of phoneme HMMs to be trained was 46. The phonemes included in the ramaining half sentences were used for recognition without the hand-labels. The recognition result is shown in Fig. 6. It can be concluded that the 1 is most effective and the 0 is not effective. This can be explained that a relatively long size of the marginal region is required because some ambiguity is caused on the phoneme boundaries by Viterbi segmentation for the long speech data like the sentences. Exclusion of unnecessary probability information

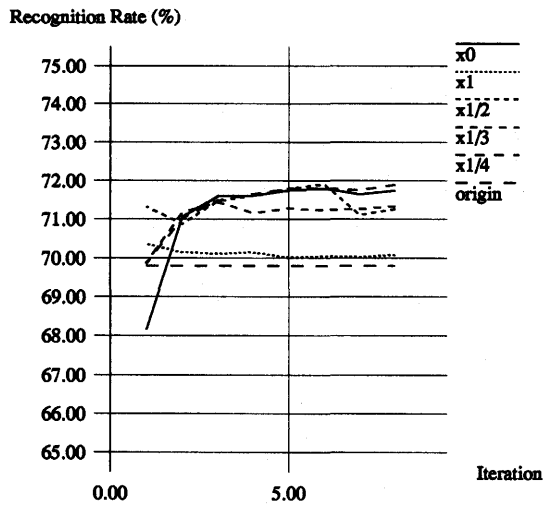

Fig. 5. Recognition result of phonemes in words by the evaluation without hand-labels

is still effectively achieved by the same size of the marginal region as the proper section. The highest recognition rate is 61.4% which is 5.6% up compared to the conventional HMM training. The phoneme recognition improvement for the sentences is 5.6% [this experiment] and better than the improvement for the words (2.1%). This means that the restriction of the training section serves to exclude the unnecessary probability information in long sentences and then to make sharp the output probability structure.

4.4 Word Recognition based on SRCT HMM

The purpose of this experiment is to evaluate how the SRCT of phoneme HMMs is effective in word recognition. In the same way as the experiment 1,2,3, phoneme HMMs were trained by SRCT using the 2620 words taken from an even number of 5240 ATR important words. The remaining half words were used for word recognition. The results is shown in Table 2.

From the table, the SRCT showed the highest recognition rate 88.1% and the isolated phoneme training showed the lowest rcognition rate 85.4%. This can be

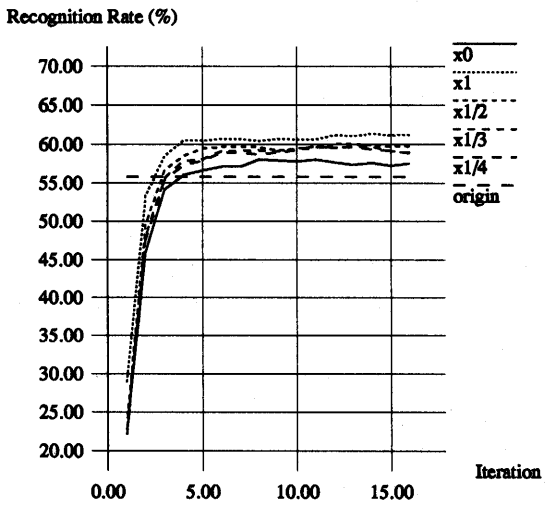

Fig. 6. Recognition result of phonemes in sentences by the evaluation without hand-labels

Table 2. Word recognition result (%)

Training method	Word recognition
Isolated phoneme training	84.1
CT	86.3
SRCT	88.1

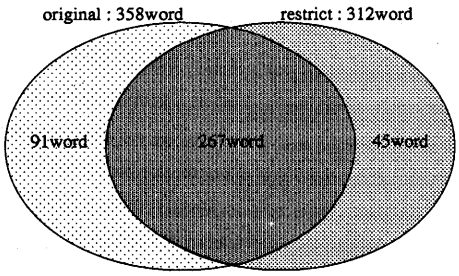

Fig. 7. Relation of misrecognized words between CT and SRCT

explained that pure phoneme information is required in the phoneme HMM training for phoneme recognition, but phoneme information with context is required for word recognition, because the phoneme HMMs are concatenated into the words in the various context.

Fig. 7 shows the relation of misrecognized words between the CT and SRCT. The 267 words are commonly misrecognized.

5. VARIATION OF THE SRCT

The idea of the proposed method lies in the separation of The CT for Viterbi segmentation (process 1) from the SRCT (process 4).

Four other methods can be derived as follows.

(a) In the process (4), phoneme HMMs trained in the process (1) are again used as the initial model, instead of new phoneme HMMs. We call this process (4'). In this case, the phoneme recognition is not improved because the output probability structure of the phoneme HMMs trained in the process (1) is already blurred.

(b) After the process (4) is completed, the final HMMs are used for the Viterbi segmentation and this process (4) is carried out again. This improves the phoneme recognition rate a little.

(c) After each iteration of the conventional concatenated HMM training in the process (1), the processes (2),(3) and (4') are carried out. This method is regarded as combined one in that the restriction of the training section is carried out in the conventional HMM training. This method improves the phoneme recognition for short sentences like words. However it sometimes makes computation errors of the forward & backward probability for long sentences because the training section is restricted in a too early stage.

(d) In the process (4), the computation of the forward and backward probabilities $\alpha_t(i),\beta_t(i)$ is not restricted within the training section, but $\gamma_t(i)$ is only computed within the training section. This does not improve the recognition rate.

6. Conclusion

New concatenated HMM training algorithm was proposed. The key idea is to restrict the training section to exclude the unnecessary probability information and to make sharp the output probability structure. The phoneme recognition was improved by 5.6% in continuously spoken sentences.

References

[1] K.F. Lee, H.W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", ICASSP' 88, pp. 123–126, 1988.

[2]Y. Minami, T. Matsuoka, K. Shikano, "Phoneme HMM Evaluation Algorithm without Phoneme Labeling", ICSLP'92, pp. 1535–1538, 1992.