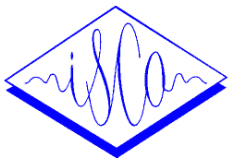

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

WORD SPOTTING USING CONTEXT-DEPENDENT PHONEME-BASED HMMs

Tatsuo Matsuoka

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper proposes a new clustering method for context-dependent phoneme HMMs. This clustering method uses triphone context as far as training samples are sufficient, and automatically selects biphone and uniphone contexts if only a few training samples are given. Using this clustering method, context-dependent models were created and tested in phoneme recognition experiments and word spotting experiments. Compared with the context-independent models, the context-dependent models achieved 7.6% higher phoneme recognition accuracy and 7.0% higher word spotting accuracy.

1. INTRODUCTION

Key-word spotting is useful in reducing the search space required for high-performance recognition in a large-vocabulary continuous speech recognition system. In a hotel reservation task, for example, the situation in the conversation can be predicted by spotting key-words such as date, place, or number of persons. After the situation has been predicted, it is easier to recognize words/sentences because the number of target words is reduced.

A phoneme-based HMM is useful for large-vocabulary word recognition. In large-vocabulary word recognition, it is quite difficult to train word-based HMMs using a few word-utterances. Instead, word models can easily be constructed by concatenating phoneme models [1][2]. This allows large-vocabulary word recognition without word-level training.

Phonemes have many coarticulation variations according to context, and HMM can deal with many of them if sufficient training samples are available. It is difficult, however, to obtain sufficient samples for training such HMMs. When a word model is constructed by concatenating phoneme models, the phoneme contexts in each word are known and limited. So higher recognition performance is expected from more precise phoneme modeling by introducing context-dependent phoneme-based HMMs.

This paper proposes a new clustering method for context-dependent phoneme modeling. This clustering method basically uses triphone context, but if a certain triphone context does not have sufficient training samples, then biphone or uniphone contexts are automatically used. The clustering process chooses triphone, biphone, or uniphone contexts according to the amount of the training samples for each context.

The context-dependent phoneme models created using this clustering method are tested in phoneme-recognition and word-spotting experiments. The context-dependent models performed better than context-independent models in both phoneme-recognition and word-spotting experiments.

2. CONTEXT-DEPENDENT PHONEME MODELING

Several reports have shown that context-dependent phoneme modeling improves recognition accuracy. Schwartz *et al.*[3] and Chow *et al.*[4] combined detailed context-dependent phoneme models and context-independent phoneme models for precise and robust phoneme modeling. Lee *et al.*[5] introduced the 'generalized triphone' model to the SPHINX speech recognition system to deal with coarticulation in continuous speech. Sagayama *et al.*[6][7] introduced Phoneme Environment Clustering to deal with allophonic variation of phonemes due to environmental effects such as coarticulatory or context dependency.

For context-dependent modeling techniques, the most serious problem is a shortage of training samples. Even considering only the left and right context; that is, triphone context, the variety of contexts is too large to obtain a sufficient number of samples for each context from the training data.

Our new clustering method is based on Phoneme Environment Clustering. This new clustering method basically considers triphone contexts; that is, left and right

contexts. Though triphone context is more effective for precise phoneme modeling than biphone or uniphone contexts, it is impossible to obtain a sufficient number of training samples for each triphone context. In the new clustering method, if there are insufficient training samples for certain triphone contexts, clustering can use biphone or uniphone contexts. Moreover, this method does not create a phoneme model for which there are less than the sufficient number of training samples. For the lower limit of the training samples, we use the number 5.

The clustering procedure is as follows.

[1] Initial clusters, C_1 to C_n, are defined for each middle phoneme in a triphone context, where 'n' is the number of phonemes. For C_1 to C_n, any context is allowed.

[2] Centroids S_1 to S_n, and distortions D_1 to D_n, are calculated for each cluster.

[3] (1) In cluster C_i, which has the maximum distortion, sub-centroids $s^{(l)}_{i,1}$ to $s^{(l)}_{i,n}$ for each left (preceding) phoneme context and sub-centroids $s^{(r)}_{i,1}$ to $s^{(r)}_{i,n}$ for each right (following) phoneme context are calculated.

(2) Distortions for each sub-cluster, $d^{(l)}_{i,1}$ to $d^{(l)}_{i,n}$ and $d^{(r)}_{i,1}$ to $d^{(r)}_{i,n}$ are weighted by the number of the samples belonging to each sub-cluster, then added to the respective distortions $Dsum^{(l)}_i$, $Dsum^{(r)}_i$.

(3) If $Dsum^{(l)}_i$ is larger than $Dsum^{(r)}_i$; that is, if left contexts have greater distortion, then cluster C_i is split into two new clusters C_j and C_k using the sub-centroids $s^{(l)}_{i,1}$ to $s^{(l)}_{i,n}$ as the members of the cluster C_i. If right contexts have greater distortion, then sub-centroids $s^{(r)}_{i,1}$ to $s^{(r)}_{i,n}$ are used for splitting.

[4] Temporal centroids S_j and S_k are calculated using the sub-centroids belonging to each cluster C_j or C_k. Then each sub-centroid is labeled according to the cluster it belongs to. This is continued until $Dsum^{(l)}_j + Dsum^{(l)}_k$ or $Dsum^{(r)}_j + Dsum^{(r)}_k$ converges.

[5] If C_j and C_k have a sufficient number of samples for training HMM, C_j and C_k are defined as new clusters; otherwise, C_i is defined as an unsplitable cluster. If the number of clusters has reached the desired number, clustering is stopped; otherwise, it is repeated from step [3].

Sagayama *et al.*[8] reported on a scheme for selecting the target cluster of splitting. They reported no differences in results between these three schemes: maximum distortion, maximum number of samples, and maximum splitting effect. We used the maximum distortion scheme because of the simplicity of this procedure.

In step [3], either left context or right context is considered when there are insufficient training samples. If there are sufficient training samples, using triphone context will give better results.

3. PHONEME RECOGNITION EXPERIMENT

Twenty-six phonemes are to be recognized. This does not include contracted phonemes, assimilated phonemes, or long vowels. Table 1 lists the experimental conditions for the phoneme-recognition experiments. Several phoneme-recognition experiments were carried out using 26, 64, 128, and 174 phoneme models. In the 174-model case, clustering was carried out until the clusters could no longer be divided. In these experiments, each cluster should have more than 5 training samples.

These HMMs had 6 states and 5 loops. Four-state-3-loop models are often used for phoneme models. In our experiments, because the end points of speech period were sometimes not accurate, 2 transition frames were added to each end of the speech period. Therefore, 2 states corresponding to the transition frames were added to these models .

Fuzzy vector quantization was applied to both the training and recognition stages.

The speech data used were three sets of 216 phoneme-balanced words uttered by one male speaker. Two sets were used for training and one set was used for testing. The evaluation was repeated three times, changing the training and testing sets in turn.

Homma *et al.*[9] reported that Phoneme Environment Clustering produced a smaller dispersion in context-dependent phoneme models than in context-independent models. In our experiment, an increase in the number of phoneme models also decreased the average distortion. Context-dependent modeling has a more concentrated distribution of samples than context-independent modeling. The context-dependent model is therefore more precise than the context-independent model.

Table 2 lists the phoneme-recognition results. Experiment (a) (upper row) used speech data recorded at the same session as the training and testing data. Experiment (b) (lower row) used speech data recorded at different sessions (one year apart) for training and testing data. In experiment (a), the recognition rate was 82.4% for 26 context-independent models, and 90.0% for 174 context-dependent

Table 1 Experimental conditions

Number of models	26, 64, 128, 174
Number of states	6 states, 5 loops
Sampling	12kHz, 16 bits
Analysis	Hamming window, frame length 32 ms, frame shift 8 ms
Feature parameters	cepstrum (16th), Δ-cepstrum (16th)
VQ	fuzzy VQ, fuzziness 1.5, knn 5, codebook size 256
Speech data	one male speaker, 216 phoneme-balanced words × 3 times

Table 2 Phoneme recognition results

Number of models	Recognition rate for training data	Recognition rate for testing data
26 models	(a) 86.6 % (b) 86.8 %	(a) 82.4 % (b) 81.3 %
64 models	(a) 90.3 % (b) 89.3 %	(a) 86.1 % (b) 85.2 %
128 models	(a) 91.4 % (b) 91.4 %	(a) 89.5 % (b) 85.7 %
174 models	(a) 93.1 % (b) 93.0 %	(a) 90.0 % (b) 86.7 %

Experiment (a) (upper row): using speech data recorded at the same time for training and testing.
Experiment (b) (lower row): using speech data recorded at different times for training and testing.

models. There is an improvement of 7.6%. This illustrates that the context-dependent phoneme model is useful for recognizing phonemes in isolated-word utterances. In experiment (b), decreases in recognition rate for 128 or 174 models are larger than that for 26 models. This is because precise models, such as 128 or 174 context-dependent models, are more sensitive to data than context-independent models. One hundred seventy-four models gave a 5.4% higher recognition rate than 26 models. Although this improvement is slightly less effective than in experiment (a), the context-dependent phoneme models were still useful.

4. WORD SPOTTING EXPERIMENT

Table 3 lists the experimental conditions for word spotting. The speech analysis, feature parameters, and vector quantization conditions were the same as in the phoneme-recognition experiment. The 4-state-3-loop models, where the transition frames were not added, were concatenated to the word models with null transitions in the word-spotting experiments.

The key-words of spotting targets were represented by 37 phoneme labels that included contracted phonemes, assimilated phonemes, long vowels, and unvoiced vowels. In the context dependent case, the same context phoneme models were used for creating the word models. The number of phonemes in each word ranged between 4 and 12, and the average was 7.5.

The phoneme models were trained using 3 sets of 216 words and a set of 503 key-words. The testing data were 84 sentences. Each sentence included one to three key-words.

The spotting algorithm is basically the same as the one Kawabata *et al.*[10] applied to island-driven continuous-speech recognition. Here, island-driven continuous-speech recognition was not tried because sentences in conversation

Table 3 Experimental conditions for word spotting

Number of words in vocabulary	44
Number of phoneme models	37, 64, 128
Number of states in phoneme models	4 states, 3 loops
Training data	one male speaker 216 phoneme-balanced words × 3 503 keywords × 1
Testing data	one male speaker 84 short sentences (Including 156 keywords)

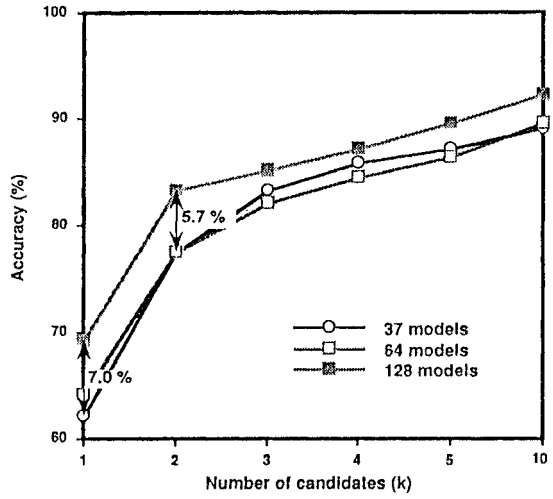

Figure 1 Cumulative detection accuracy

include many unnecessary words. In our word-spotting experiments, several candidates were selected according to the logarithm likelihood.

Figure 1 shows the cumulative detection accuracy. The k-th cumulative detection accuracy is the rate for spotting key-words correctly within the top k choices. Using 64 context-dependent models, the accuracy increased by 2.0% for k=1 compared to the 37-model case. Using 128 context-dependent models, the accuracy increased by 7.0% for k=1 and by 5.7% for k=2.

These results show that the context-dependent phoneme models are also effective for word-spotting.

5. SUMMARY

A new clustering method was proposed for context-dependent phoneme-based HMMs. If only a few training samples are available, this clustering method automatically uses biphone and uniphone contexts instead of triphone contexts .

Using this clustering method, context-dependent phoneme-based HMMs were created. The models were tested in phoneme-recognition experiments and in word-spotting experiments. In the phoneme-recognition experiments, 174 context-dependent models achieved a 7.6% higher recognition rate than context-independent models. In word-spotting experiments, 128 context-dependent models achieved 7.0% higher accuracy than context-independent models. These results demonstrate that context-dependent models are useful for both phoneme recognition and word spotting.

One of the remaining problems is modeling for untrained contexts, which do not appear in the training samples. One idea for resolving this problem is to use a context-independent model instead of the context-dependent model if the context was not in the training sample. A more advanced idea is to interpolate the models for untrained contexts. Another problem is the difference in duration between isolated words and continuous speech due to utterance speed. Some duration-control scheme should be introduced to the phoneme models.

REFERENCES

(1) Chow, Y. L., Dunham, M. O., Kimball, O. A., Krasner, M. A., Kubala, G. F., Makhoul, J., Price, P. J., Roucos, S., Schwartz, R. M., "BYBLOS: The BBN Continuous Speech Recognition System," ICASSP87, pp. 89-92

(2) Lee, K. F., Hon, H. W., "Large-vocabulary Speaker-independent Speech Recognition using HMM," ICASSP88, pp. 123-126

(3) Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., Makhoul, J., "Context-dependent Modeling for Acoustic-phonetic Recognition of Continuous Speech," ICASSP85, pp. 1205-1208

(4) Chow, Y. L., Schwartz, R., Roucos, S., Kimball, O., Price, P., Kubala, F., Dunham, M. O., Krasner, M., Makhoul, J., "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-based Speech Recognition System," ICASSP86, pp. 1593-1596

(5) Lee, K. F., Hon, H. W., Hwang, M. Y., Mahajan, S., Reddy R., "The Sphinx Speech Recognition System," ICASSP89, pp. 445-448

(6) Sagayama, S., "Phoneme Environment Clustering," Proc. of Acoustic Society of Japan Fall Meeting, 1-5-15, pp. 29-30, Oct. 1987 (in Japanese)

(7) Sagayama, S., "Phoneme Environment Clustering for Speech Recognition," ICASSP89, pp. 397-400

(8) Sagayama, S., Homma, S., "Performance of Phoneme Environment Clustering in Phoneme Recognition," Trans. Committee on Speech Research, Acoustic Society of Japan, SP89-78, pp. 17-24, Dec. 1989 (in Japanese)

(9) Homma, S., Sagayama, S., "On the Use of Phoneme Environment Clustering for HMM based Phoneme Recognition," Proc. of Acoustic Society of Japan Fall Meeting Proc., 1-1-18, Oct. 1989 (in Japanese)

(10) Kawabata, T., Shikano, K., "Japanese Phrase Recognition Based on HMM phone Units," Proc. of Acoustic Society of Japan Fall Meeting, 2-P-26, pp. 253-254, Oct. 1988 (in Japanese)