SPEECH RECOGNITION USING SUB-WORD UNITS DEPENDENT ON PHONETIC CONTEXTS OF BOTH TRAINING AND RECOGNITION VOCABULARIES

Hiroaki HATTORI and Eiko YAMADA

Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, JAPAN
E-mail: hattori@hum.cl.nec.co.jp

ABSTRACT

This paper proposes a new speech recognition algorithm using a new context-dependent recognition unit design method for efficient and precise acoustic modeling. This algorithm uses both training and recognition vocabularies to select context-dependent units which precisely represent acoustic variations due to phonetic contexts in a recognition vocabulary. An efficient training algorithm for selected context-dependent units is also proposed. In speaker-independent isolated-word recognition experiments, the proposed algorithm gave a 11% error reduction for 5000 word recognition, and gave a 43% error reduction for 10 digit recognition. These results confirmed the effectiveness of the proposed method.

1. INTRODUCTION

Recently phonetic context-dependent sub-word unit based approaches have been widely used because of their efficiency in representing large vocabulary and because of their ability to cope with acoustic variations due to phonetic contexts. Triphone unit[1] is a well known phonetic context-dependent unit. As is obvious, however, that triphone units can not handle the effects of phonetic contexts larger than sequential three phonemes.

A simple and straightforward way to obtain more precise acoustic modeling is the use of a recognition units with larger phonetic contexts, such as word-dependent sub-word units. In paper [2], the use of a word-dependent phone model for *function words* was proposed. *Function words* consists of prepositions, conjunctions, pronouns, and short verbs, and 42 words were empirically selected. When a task and its vocabulary are fixed, such recognition units can be trained using training data from the task itself. In actual practice, however the task and the vocabulary are likely to be modified, so that all possible context-dependent models have to be trained to cope with the modification. An example of this approach is Context Adaptive Phone(CAP) modeling[3], in which a set of models with different degree of contextual influence up to word levels are used to cover all contextual acoustic variations in a training vocabulary. But, because the units are designed only from the phonetic contexts in a training vocabulary, CAP modelings contain models that are not needed to represent a recognition vocabulary.

In this paper, we propose a new unit design method that uses both training and recognition vocabularies in such a way as to take advantages of the efficiency of sub-word units and the precise acoustic modeling of word-depend units and to keep the model set compact.

The rest of this paper is organized as follows: in the next section, a procedure of the proposed method is described. In Section 3, an algorithm of efficient context-dependent model training is described. Section 4 describes the evaluation experiments and presents the results.

2. ALGORITHM

The our goal is to get a compact set of recognition units that accurately represent acoustic variations due to phonetic contexts. Both training and recognition vocabularies, therefore, have to be taken into account when designing units because what we need is a set of units that can represent the recognition vocabulary and what we can get is a set of units contained in the training vocabulary.

The flowchart of the proposed algorithm is shown in Figure 1. The training and recognition vocabularies are first represented by means of a context-independent recognition unit set. Then the phonetic context selector compares a context of each unit in the recognition vocabulary with that of its

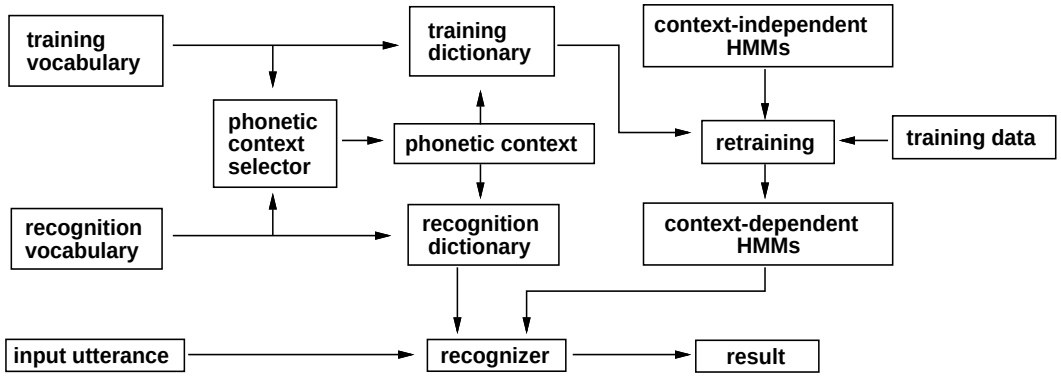

Figure 1: Flowchart of the proposed method.

corresponding unit in the training vocabulary, and selects the unit with the longest matched context as a new context-dependent recognition unit. Finally, the selected units are trained using the training data.

This algorithm assumes that the training data contains enough samples to train word-dependent units. When only a small amount of data is available, the matched context search process can be stopped according to the number of training samples for the selected unit. If the speech data of a recognition vocabulary is available beforehand, a unit selection based on the maximum likelihood can be used[5].

3. MODEL TRAINING

As described in the previous section, the proposed algorithm requires retraining of HMM for every modification of the recognition vocabulary. When the computational cost is high, this algorithm would, therefore, not be feasible for a real application. This section describes a retraining algorithm requiring only a small amount of computation. For simplicity, only the retraining of a transition probability is shown in the following description, but other parameters can be treated in the same manner.

Let's assume that the training word w consists of P_n context-independent units and that the intermediates in HMM training procedure $Na^w(i_p, j_p)$ and $Da^w(i_p, j_p)$ given by the following equations, are calculated before retraining:

$$Na^w(i_p, j_p) = \sum_v \sum_t \gamma_{vw}(i_p, j_p, t)$$

$$Da^w(i_p, j_p) = \sum_v \sum_t \sum_j \gamma_{vw}(i_p, j_p, t)$$

Here, $\gamma_{vw}(i_p, j_p, t)$ is the probability that the transition from state i_p to j_p is made at time t for the given vector sequence $\{O_{vw}\}$ of speaker v's utterance.

Let's also assume that a model q with a certain phonetic context is selected as a new context-dependent unit and that the model appears in the training vocabulary N times. Let $w(i), 1 \leq i \leq N$, be the word which contains the model and let $p(i)$ be the position of the model in the word $w(i)$.

Then the transition probability $a_{i_q j_q}$ from state i_q to j_q of model q is calculated from $Na^{w(i)}(i_{p(i)}, j_{p(i)})$ and $Da^{w(i)}(i_{p(i)}, j_{p(i)})$ as

$$a_{i_q j_q} = \frac{\sum_{i=1}^{N} Na^{w(i)}(i_{p(i)}, j_{p(i)})}{\sum_{i=1}^{N} Da^{w(i)}(i_{p(i)}, j_{p(i)})}.$$

Figure 2 illustrates the retraining process. For the retraining of a parameter, this algorithm requires only $2N$ additions and one division.

4. EVALUATION EXPERIMENTS

In this section, the evaluation experiments of the proposed method is reported. The experiments was carried out on our recognition system based on Japanese demi-syllable units[4]. A demi-syllable is a unit divided at the center of the syllable nucleus. Since Japanese syllables have a combined consonant-vowel structure, there are consonant-vowel(CV) models and vowel-consonant(VC) models. Demi-syllable units, therefore, originally include contextual effects of a two phoneme sequence.

4.1. Experimental Conditions

In the evaluation experiment, demi-syllable was used as a unit for recognition and phonetic context for convenience.

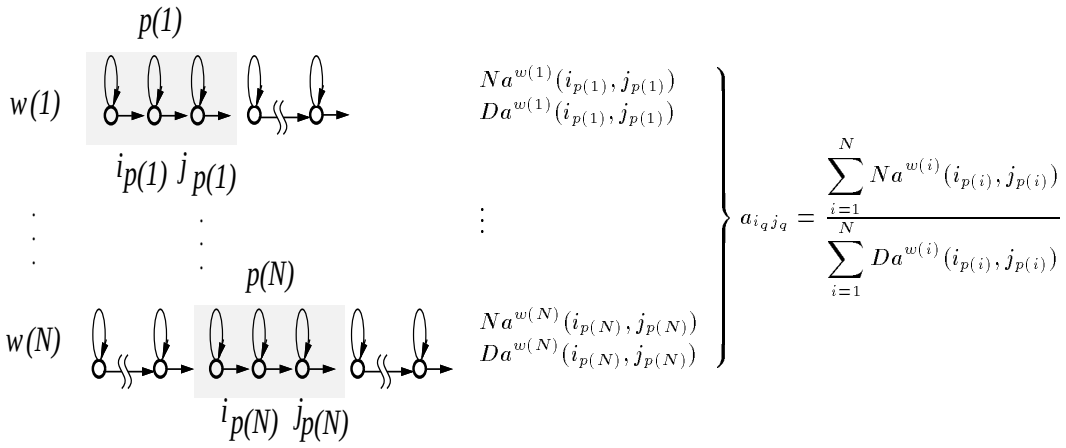

Figure 2: Retraining of a transition probability.

This means that a unit with a preceding context and a succeeding context represents a four phoneme sequence as CV-VC-CV or VC-CV-VC.

The longest matched context was selected under three different phonetic context conditions: (1) independent of context, (2) up to one succeeding and one preceding context, and (3) up to word context. Even under the last condition, there may be a model with no matched context; then a context-independent model is used as a result.

When selecting the longest matched context, there may be several units of the same context length with different succeeding and preceding context lengths; a model which matches one preceding context and two succeeding contexts and a model which matches two preceding contexts and one succeeding context. In such a case, we chose a model which has the following phonetic contexts:

- proceeding and succeeding contexts of the same length.
- a longer proceeding context for CV models.
- a longer succeeding context for VC models.

These are based on three assumptions: (1) a closer context is more influential, (2) influences of preceding and succeeding contexts are the same in degree, and (3) acoustic features of a consonant are strongly affected by the preceding and succeeding vowels.

Acoustic analysis and other experimental conditions are listed in Table. 1.

4.2. 5000 Word Recognition

A speaker-independent 5000 word recognition experiment was carried out in order to evaluate the effectiveness of the proposed method.

For the training of HMMs, 291 phonetically balanced words uttered by 22 male speakers and 20 female speakers were used. As a recognition vocabulary, 5000 words independent of the training vocabulary were selected; 250 of these words uttered by another 5 male and 5 female speakers were used for the evaluation.

The number of selected units, the average number of training samples, and the word recognition rate for each select condition are listed in Table 2.

In the case of conventional context-independent modeling, 264 models were trained from 8.83 samples and gave the recognition rate of 79.2%. When the longest matched context up to a preceding and a succeeding contexts was selected, 2350 models were trained from 2.3 samples and the recognition rate was 81.5%. When the longest matched context up to a word context was selected, 4695 models were trained from 1.7 training samples and the rate of 81.4% was achieved. These are about a 11% reduction in error rate compared to the conventional context-independent modeling.

Table 1: Experimental conditions

Sampling freq.	11.025 kHz
Analysis period	16.0 ms
Freq. bandwidth	150 $\sim$ 5000 Hz
Feature parameters	power derivative(1) mel-cepstrum(10) mel-cepstrum derivatives (10)
Acoustic models	semi-continuous HMM [6] 3 state left-to-right model
No. of PDFs	32 for power derivative, 512 for mel-cepstrum 256 for mel-cepstrum derivatives
Retraining of models	mixture weights and transition probabilities

Table 2: Word recognition results.

select condition	no. of units	no. of samples	recognition rate (%)
independent	264	8.8	79.2
longest matched+	2350	2.3	81.5
longest matched	4695	1.7	81.4

+: up to a preceding and a succeeding contexts.

In addition, the number of selected models was kept to a manageable size. The proposed method yielded 2350 models when phonetic context up to a proceeding and a succeeding demi-syllables were taken into account, where the number of units trainable from the training data is more than 4,900 in the same condition. These results clarified the effectiveness of the proposed method.

The select condition up to a preceding and a succeeding contexts and the select condition up to a word-context gave almost the same improvement over the conventional context-independent modeling. This is because the training vocabulary was limited so that a preceding and a succeeding contexts were sufficient to reduce the number of training samples. The average of 8.8 of training samples for the conventional context-independent models was reduced to 2.3 by taking into account a preceding and a succeeding contexts. Taking into account phonetic contexts up to word-context, however, reduces the average number of training samples by only another 0.6 samples. Further study with a larger training vocabulary would be needed to evaluate the difference between these select conditions.

4.3. Digit Recognition

In the previous experiment, the proposed method reduced the error rate by 11% even when the recognition vocabulary was completely different from the training data. For a vocabulary included in the training vocabulary, a more improvement would be expected. An experiment involving the recognition of 10 digits included in the training vocabulary was, therefore, carried out to investigate the potential of the proposed method.

As testing data, 10 digits uttered 2 times by different 8 male and 8 female speakers from the speakers of the training data were used. Other conditions are the same as in the previous experiments, and the results are shown in Table 3. The 98.4% recognition rate of the conventional context-independent modeling was improved to 99.1% by using the proposed method. This is a 43% error reduction.

Table 3: Digit recognition results.

select condition	no. of units	no. of samples	recognition rate (%)
independent	28	27.8	98.4
longest matched+	35	3.9	99.1
longest matched	36	1.0	99.1

+: up to a preceding and a succeeding contexts.

5. CONCLUSIONS

A new context-dependent recognition unit design algorithm that uses both training and recognition vocabularies was proposed and its effectiveness was confirmed in speaker-independent word and digit recognition experiments.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Watanabe for his valuable comments and Mr. Shinoda for his programming support. The authors also thank other members of the speech recognition group in the Human Language Research Laboratory for their support and discussions.

REFERENCES

1. R. Schwartz et al., "Improved Hidden Markov Modeling Of Phonemes For Continuous Speech Recognition," Proc. of ICASSP84, 35.6 (1984.5).
2. K.F. Lee, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX system," PhD thesis, Computer Science Department, Carnegie Mellon University (1988.4).
3. R. Moore et al., "A Comparison Of Phoneme Decision Tree(PDT) And Context Adaptive Phone Based Approaches To Vocabulary-Independent Speech Recognition," Proc. of ICASSP94, Vol. 1, pp. 541-544 (1994.5).
4. K. Yoshida et al., "Large Vocabulary Word Recognition Based On Demi-Syllable Hidden Markov Model Using Small Amount Of Training Data," Proc. of ICASSP89, S1.1 (1989.4).
5. T. Matsumura and S. Matsunaga, "Non-Uniform Unit HMMs For Speech Recognition," Proc. of EUROSPEECH, pp. 499-503 (1995.9).
6. X.D. Huang and M.A. Jack, "Semi-Continuous Hidden Markov Models for Speech Signals," Computer Speech and Language, Vol. 3, pp. 239-251 (1989).