A CONTINUOUS SPEECH RECOGNITION SYSTEM USING FINITE STATE NETWORK AND VITERBI BEAM SEARCH FOR THE AUTOMATIC INTERPRETATION

Nam-Yong Han, Hoi-Rin Kim, Kyu-Woong Hwang, Young-Mok Ahn and Joon-Hyung Ryoo

Automatic Interpretation Section
Electronics and Telecommunications Research Institute (ETRI)
E-mail: nyhan@zenith.etri.re.kr

ABSTRACT

This paper describes a Korean continuous speech recognition system using phone based semi-continuous hidden Markov model (SCHMM) method for the Automatic Interpretation. The task domain is hotel reservation. The system has the following three features. First, an embedded bootstrapping training method that enables us to train each phone model without phoneme segmentation database is used. Second, a hybrid estimation method which is composed of the forward-backward algorithm and the Viterbi algorithm is proposed for the HMM parameter estimation. Third, a between-word modeling technique is used at function word boundaries. The recognition results in speaker independent experiments are as follows. In the case of Version 1, continuous speech recognition result is 89.1% and in Version 2, the result is 97.6%.

1. INTRODUCTION

Recently, speech recognition technology has been improved and shown the possibility to be used in real meaningful domain with continuous speech. To go with this situation, we have made a continuous speech recognition system, which is near practical use, for automatic interpretation [1]. The fundamental structure of an automatic interpretation system is depicted in Figure 1.

As shown in Figure 1, the whole system is composed of speech recognition, machine translation, and speech synthesis. The speech recognition part of Figure 1 is the most critical factor of our system performance. So, we will mainly describe the speech recognition part. The whole interpreting system will be described at the EUROSPEECH95 if possible. We select hotel reservation task to construct a practical automatic interpretation system between Korean and Japanese or between Korean and English.

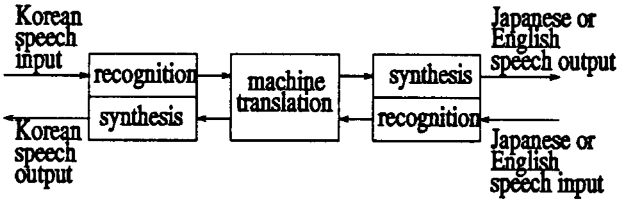

Figure 1: The fundamental structure of the automatic interpretation system between Korean and Japanese/English.

2. TASK DOMAIN AND PREPROCESSING

2.1. Task Domain

The hotel reservation task has two kinds of vocabulary sets. The first one is a word and sentence set related to customer-side speaking such as hotel reservation, reservation change, reservation cancellation, and room exchange. This vocabulary is a set of 244 words and *word-phrases* including digits (room numbers and dates), English alphabet, etc[2]. The meaning of the *word-phrase* will be described in Section 4.

The other one is a word and sentence set related to hotel front-desk-side speaking. This vocabulary is a set of 242 words and *word-phrases* including digits related to dates, 167 Japanese surnames, etc. These data sets are shown in Table 1.

To reflect real situation, it is good to get data from real situation. But, we chose these speech data from virtual scenario for convenience. We made domain sentences for the two vocabulary sets from predefined finite state network grammars. These grammars have 74 sentence types and 26 types for the both data sets, respectively.

Correct pronunciation dictionary for training data is important to make a good speech recognizer especially for our system which is trained on unlabeled speech. The detail informations of the pronunciation

0-7803-2431-5/95 $4.00 © 1995 IEEE

Data sets		Train	Test
customer side	Words	244 × 40 M	244 × 11 M
	Sentences	110 × 40 M	110 × 11 M
front desk side	Words	242 × 40 M	242 × 10 M
	Sentences	93 × 40 M	93 × 10 M

Table 1: Speech databases for the each side. The words pronounced by each speaker are the same ones for both data sets, respectively. But, the sentences pronounced by each speaker are different from other sentences. All speakers pronounce words and sentences only one time. The M means the male speakers.

dictionary are covered in [2].

2.2. Preprocessing

We use four codebooks, each with 256 entries, that use (1) 12 LPC cepstral coefficients; (2) 12 delta LPC cepstral coefficients; (3) 12 delta delta LPC cepstral coefficients; and (4) normalized log power, delta power, and delta delta power. For end point detection, we use sequential method which is varied from [3] for demonstration. Preprocessings are done as follows:

sampling rate 16kHz
AD precision 16bit
preemphasis $H(z) = 1 - 0.95z^{-1}$
analysis window 20msec
analysis interval 10msec
Hamming window
LPC-cepstrum 20 order
Band lifter

3. TRAINING

3.1. HMM Topology

The Figure 2 shows the HMM topology used in our HMM-based speech recognizers, Version 1 and Version 2. This HMM topology for each context dependent phone like unit (CD_PLU) is a simple left-to-right model with 3 states and 8 transitions. The transitions are tied into three groups for robust estimation of output probabilities. Transitions in the same group, represented by B, M, and E, share the same output probabilities [4, 5]. This model assumes that there are at most three steady states for a CD_PLU, which are indicated by the self-loops. The HMM model for silence has additional null transition which is not associated with output symbol. This assumption allows optional silence to exist at the start of a sentence, at the end of a sentence, and between phrases.

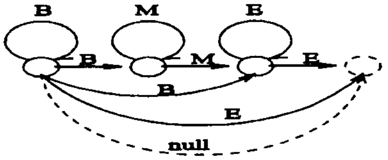

Figure 2: The HMM Topology for each triphone.

3.2. CD_PLU Training

There are two approaches to train CD_PLUs for both side data sets. First, it is a customer-side training method and is the case of the Version 1. In this training procedure, we run training in two stages. The first stage is the training based on the word database and then the second one is the training on sentence database. This method was described [2] in detail. The second approach, however, to train CD_PLUs for front desk-side set is that it runs on the word and sentence data together in only one stage, and this is case of the Version 2. It is a merged training method compared with training one in [2]. In the Figure 3, we used the duration information of 51 context-independent phonetic models for initial segmentation of the word and sentence database like [2]. We then run the forward-backward (FB) algorithm and segmental k-means (SKM) algorithm [6] on the word and sentence database together.

The above training procedures lead to one major problem, namely that the number of occurrences of some of the CD_PLU units is insufficient to generate a statistically reliable model. So, we followed the reduction rule described in [3]. But, the order of triplet, *a_b_c*, is different from [3]. In the triplet, *b* is the left context phone, *c* is the right context phone, and *a* is the current phone. In Figure 4 (B), the # denotes a don't care condition.

To train the SCHMMs, we used the hybrid algorithm which first segments speech into the unit of CD_PLU by the Viterbi decoding and then uses FB algorithm within each CD_PLU boundary. This algorithm requires less computation than full FB algorithm which is applied to the whole utterance. The FB algorithm considers all paths and the SKM considers only the best path, and our hybrid algorithm considers the best path in the inter-CD_PLU and full paths in the intra-CD_PLU.

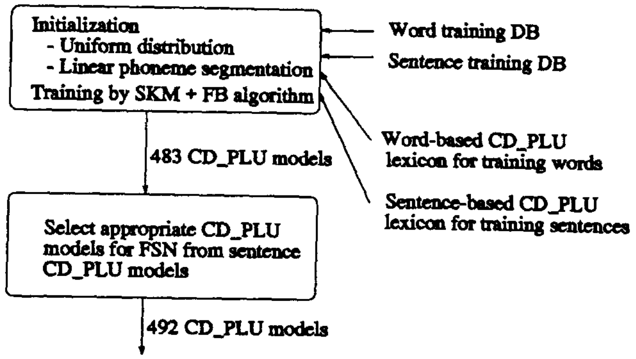
Initialization
- Uniform distribution
- Linear phoneme segmentation
Training by SKM + FB algorithm
Word training DB
Sentence training DB
Word-based CD_PLU lexicon for training words
Sentence-based CD_PLU lexicon for training sentences
483 CD_PLU models
Select appropriate CD_PLU models for FSN from sentence CD_PLU models
492 CD_PLU models

Figure 3: Training procedure for the front-desk side data set to estimate CD_PLU parameters.

4. RECOGNITION

4.1. FSN as a Language Model

In the developed systems, Version 1 and Version 2, we adopted a finite state network (FSN) grammar as a language model. This FSN is usually not a proper grammar for the language whose word order is not important like Korean compared with English. But, because this FSN can represent the syntactic and semantic restrictions and can reduce search space in recognition stage, we used the FSN as a language model of our baseline system. Figure 4 (C) shows one example of the FSNs in the customer-side task. In Figure 4 (C), each node means words which are able to construct all legal sentences in this situation.

The main difference between 1) customer-side experiment (Version 1 recognizer) and 2) hotel front-side experiment (Version 2 recognizer) is that the only Version 1 uses dialog managing. In the case 1), all sentences were classified into 9 categories according to the customer's speaking patterns [2]. Therefore, In recognition stage, one proper network of 9 FSNs is selected by a managing program according to the speaking situations between hotel front desk clerk and guest. By this situation control, recognition time is reduced. The managing program and the Viterbi program are running at the same time. In the case 2), however, because this data set and sentence type are more simple than former, there is only one FSN without categorizing the sentences.

4.2. Between Word Modeling

Korean is an agglutinative language in which the post-positional word (function word) shows syntactic relation of its preceding words. And, the function words always used with preceding words without putting space between two preceding and function words. We call this pair of words as ***word-phrase***. Although the function words form independent word groups in a separate node in Figure 4 (C), the words in a ***word-phrase*** are pronounced like one word. So, the function words shows strong coarticulation and we cannot treat them as separated words. To solve this problem, we used the between-word modeling [2, 4].

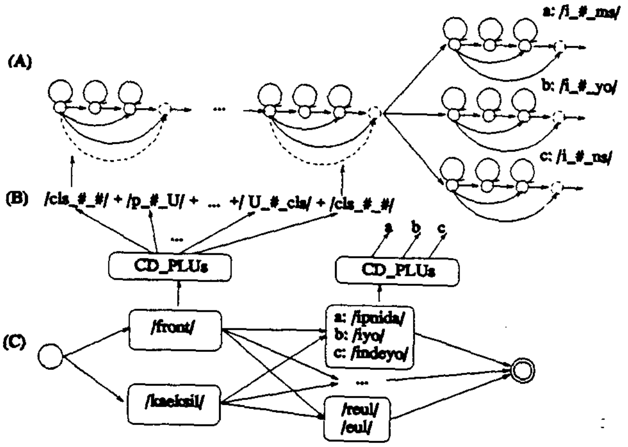

Figure 4: (A) Expanded HMM states according to HMM topology, (B) triphones as a CD_PLU, (C) FSN for a situation.

4.3. Implementation of a Continuous Speech Recognizers

We incorporate word and sentence knowledge into our recognizers in the following manners: Each word is represented as a network of CD_PLUs (in Figure 4 (B)) which encode the way the word can be pronounced. The FSN grammar can be represented as a network whose nodes are words, and the network encodes all legal sentences. We can then take the FSN grammar network, instantiate each word with the network of CD_PLUs, and then instantiate each instance of a CD_PLU with its HMM (in Figure 4 (A)). Then we have a large HMM that encodes all the legal sentences. By placing all the knowledge in the data structures of the HMMs, it is possible to perform a global search that takes all the knowledge into account at every step. This integrated search is implemented in our continuous speech recognition systems, Version 1 and Version 2.

In addition to the above data structure, we added following data structure to obtain another information in recognition stage. The structure consists of 6 fields: Fsn_node, Word_id, CD_PLU_id, State_id, In, and Out. It gives us easy method to search best path and once we know present HMM state number at any step, it

Data sets	Close experiment	Open experiment
Version 1	95.6	90.1
Version 2	99.3	97.3

Table 2: Speaker independent isolated-word recognition rates (in the Top 1 case) for Version 1 and Version 2 experiments (unit: %).

Data sets	Including insertion err.	Excluding insertion err.
Version 1	89.1	91.2
Version 2	97.6	97.8

Table 3: Continuous speech recognition rates for Version 1 and Version 2 experiments. (unit: %)

gives immediately us CD_PLU, Word, FSN node information, and etc. Therefore, it is convenient to picture an overall view in recognition stage. The detail descriptions of the fields are covered in [2].

In our systems, we used Viterbi beam search ([7]) as a search algorithm and used partial Viterbi backtracking in demonstration. Especially, In the case of Version 1, we considered beam search strategy using a fixed beam width as a pruning threshold and in the case of Version 2, we considered the threshold value and the number of survived states as some constraints.

5. RESULTS AND ANALYSIS

To evaluate our system, we took two kinds of experiments. The first one is an isolated word recognition experiments shown in Table 2, and the second is a continuous speech recognition experiments shown in Table 3. The recognition results of the SCHMM and the DHMM methods are covered in [2].

In speaker-independent continuous speech recognition experiments, the recognition results have various accuracy, from 86.2% to 98.1%, according to the 9 categories in the case of Version 1. The Table 3 shows the average recognition results for Version 1 and shows the one result for Version 2 which is using only one FSN.

The main reasons for the obtaining of the results like Table 3 as follows. First, the data set and the sentence type for Version 2 (although the branching factor is 167 for the case of Japanese surnames at a decision point, most of the branching factor is 1. As a result, the perplexity is 1.4) are more simple than that of Version 1 (the perplexity is 4). Second, training approach of Version 2 is different from Version 1. Namely, to train CD_PLUs for Version 2, we used both word and sentence data together in only one stage.

6. CONCLUSIONS

We made a Korean continuous speech recognition system for automatic interpretation for the hotel reservation domain using SCHMM with finite state network grammar. In the system, embedded bootstrapping method, hybrid reestimation method, and between word modeling are used for training. We especially adopted a merged training approach compared with [2] to recognize a sentence in the case of Version 2. We also constructed demonstration system for the both side Versions. The recognition time of the Version 2 is faster 7 times than Version 1.

7. REFERENCES

[1] H.R. Kim, K.W. Hwang, N.Y. Han, and Y.J. Lee, "A Preliminary Study on Continuous Speech Recognition of Hotel Reservation Task for Automatic Interpretation," Proceedings of the 10th Workshop on Speech Communication and Signal Processing, pp. 256-261, 1993. (In Korean).

[2] H.R. Kim, K.W. Hwang, N.Y. Han, and Y.M.Ahn, "Korean Continuous Speech Recognition System Using Context-Dependent Phone SCHMMs," Proceedings of the Fifth Australian International Conference on SPEECH SCIENCE AND TECHNOLOGY, Vol.II, pp. 694-699, 1994.

[3] L.R. Labiner and B.H. Juang, Fundamentals of Speech recognition, Prentice Hall: New Jersey, pp. 460-461, 1993.

[4] C.H. Lee, L.R. Rabiner, R. Pieraccini, and J.G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," Computer Speech and Language, Vol.4, pp. 127-165, 1990.

[5] K.F.Lee, "Context-dependent phonetic hidden Markov models for speaker independent continuous speech recognition," IEEE Trans. on ASSP, vol.38, no.4, pp. 599-609, 1990.

[6] L.R. Rabiner, J.G. Wilpon, and B.H. Juang, "A segmental k-means training procedure for connected word recognition," AT&T Technical Journal, 65(3), pp. 21-31, 1986.

[7] S. Furui and M.M. Sondhi, Advances in Speech Signal Processing, Marcel Dekker, Inc., pp. 623-650, 1991.