KOREAN CONTINUOUS SPEECH RECOGNITION SYSTEM USING CONTEXT-DEPENDENT PHONE SCHMMS

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

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

ABSTRACT - This paper presents the Korean continuous speech recognition system using phone-based semi-continuous hidden Markov model (SCHMM, also known as a tied-mixture model) method. 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 was used. Second, in the HMM parameter estimation, a hybrid estimation method which is composed of the forward-backward algorithm within phoneme boundaries and the Viterbi algorithm to determine those phoneme boundaries, was proposed. Third, a between-word modeling technique in word boundaries was used to solve the strong coarticulation between the Korean postpositional word and its preceding word. Task domain of the system is the query sentences of hotel reservation with 244 words including digits, English alphabet, etc. Speech database for simulation consists of two parts; one is the word data which were pronounced once by 51 male speakers, the other is a set of 5610 different sentences pronounced by the 51 speakers. We have defined 339 context-dependent phone models based on triphone model for pronunciation dictionary of each word. For the phone models, we use both the DHMM and SCHMM methods for performance comparison, and define a model topology with 3 states and 8 transitions including skip transitions. The silence model has an additional null transition. We use four feature vectors: LPC cepstrum with a bandpass lifter, delta cepstrum, delta-delta cepstrum, and energy(logE, delta logE, and delta-delta logE). For HMM training, two training stages were applied to the word and sentence data. In recognition stage, the finite state grammar for language modeling and the Viterbi beam algorithm for search were used. In speaker-independent recognition experiments, the discrete HMM (DHMM) method resulted in 89.7% word accuracy and the SCHMM method in 89.0%.

INTRODUCTION

Recently, speech recognition technology has been improved and shows the possibility to be used in real meaningful domain with continuous speech. To go with this situation, we made a continuous speech recognition system which is near practical use. We selected hotel reservation including telephone exchange service as our task domain (Kim *et al.*, 1993) because it is most efficient in terms of technology and applicability. Minimal required information can be exchanged with small vocabulary in this domain. Brief description of task domain is as follows:

Conversation range: customer's speech in hotel reservation and telephone exchange
Vocabulary: 244 words including numbers and English alphabets
Grammar: finite state grammar with perplexity 4.

We used embedded bootstrapping method to train the recognizer without phoneme segmentation database and hybrid estimation method to reduce computation. These things are described in later sections.

TASK DOMAIN AND PRONUNCIATION DICTIONARY

Task domain for our continuous speech recognition system is hotel reservation and telephone exchange service. This domain sentences consist of those which can be used in hotel reservation, reservation change, reservation cancellation, and telephone exchange service. This excludes the sentences which can be spoken by hotel telephone operator or front desk server. Vocabulary, sentence types, and grammar for recognition and dialog scope are as follows.

This domain consists of 244 words. To reflect real situation, it is good to get data from real situation. But, we chose 244 words from virtual scenario for convenience. This vocabulary consists of room numbers from 10 to 99, 43 words for dates, 7 words for day of week, 26 words for English phoneme, and 59 other words. We made domain sentences from predefined finite state network grammar. This grammar has 74 sentence types. A sentence type is defined by a possible path in a finite state network (FSN) and the nodes in FSN represent word groups. To model the dialogs, we assumed 9 dialog situations. FSN grammar is not proper in general for natural dialogs, especially for Korean which has more flexible word orders than English. But, our task domain is relatively small. So, we used FSN grammar to model task domain best.

Correct pronunciation dictionary for training data is important to make a good speech recognizer especially for our system which is trained on unlabeled speech. Multiple entries in a pronunciation dictionary will provide more accurate description. But, we used single entry for each word for simplicity, except for functional words. To describe Korean speech, we determined 51 phone like units (PLU). Diphthongs are considered as independent units. For consonants, first-in-syllable and last-in-syllable are considered as independent units. Besides, lax consonants are considered as independent units when voiced. If they lies between voiced sounds, we consider them as voiced allophones. We generated context dependent phone like unit (CD_PLU) set by considering left-right context dependency. When the frequency of model is less than needed for training, more general and less context-dependent model is used (Rabiner *et al.*, 1993). We used 100 for the needed frequency of model for training.

database	training	test
244 words	9748	2680
sentences	4367	1203

Table 1: Speech data used in our experiments. speeches of 40 persons is used for training and 11 for test

PREPROCESSING

We used four codebooks, each with 256 entries, that use (1) 12 LPC cepstrum coefficients; (2) 12 delta LPC cepstrum coefficients; (3) 12 delta delta LPC cepstral coefficients; and (4) normalized log power, delta power, and delta delta power. For end point detection, we used sequential method which was varied from (Rabiner *et al.*, 1993) 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

TRAINING

The SCHMM, first proposed by (Huang *et al.*, 1990), is an excellent example of detailed modeling through parameter sharing. The SCHMM has following advantages. First, shared mixtures substantially reduce the number of free parameters and computational complexity in comparison with the continuous mixture HMM, while maintaining its modeling power. Second, the SCHMM integrates quantization accuracy into the HMM, and robustly estimates the discrete output probabilities by considering multiple codeword candidates in the VQ procedure. It optimizes mutually the vector quantization (VQ) codebook and HMM parameters under a unified probabilistic framework. Third, it requires less training data compared with the DHMM (Huang *et al.*, 1992).

In this study, we apply context-dependent phone HMMs.

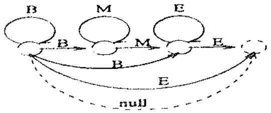

Figure 1: Each CD_PLU HMM is a simple left-to-right model with 3 states and 8 transitions. The transitions are tied into three groups for robust estimation of output probabilities. Transitions in the same group, represented by B, M, and E, share the same output probabilities. This model assumes that there are at most three steady states for a CD_PLU, which are indicated by the self-loops. The HMM 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.

Without phonetically segmented speech database, it is easier to train HMMs on the word database because utterances in it have smaller number of phonemes and is less coarticulated than those in the sentence database. So, our training procedure operates in two stages as in Figure 2.

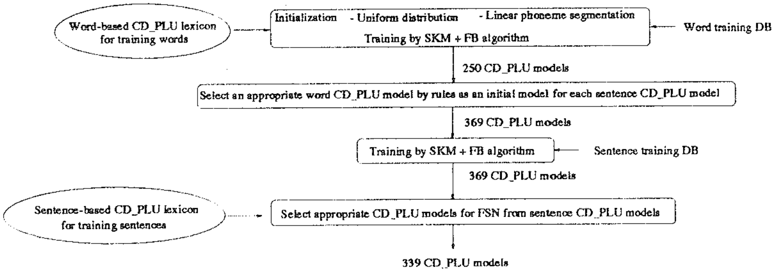
Word-based CD_PLU lexicon for training words
Initialization - Uniform distribution - Linear phoneme segmentation
Training by SKM + FB algorithm
Word training DB
250 CD_PLU models
Select an appropriate word CD_PLU model by rules as an initial model for each sentence CD_PLU model
369 CD_PLU models
Training by SKM + FB algorithm
Sentence training DB
369 CD_PLU models
Sentence-based CD_PLU lexicon for training sentences
Select appropriate CD_PLU models for FSN from sentence CD_PLU models
339 CD_PLU models

Figure 2: Two stage training procedure to estimate CD_PLU parameters. We used the duration information of 51 context-independent phonetic models for initial segmentation of the word database. We run the forward-backward (FB) algorithm and segmental k-means (SKM) algorithm(Rabiner *et al.*, 1986) on the word database. In the second stage, we run the above algorithms on the sentence database.

To train the SCHMMs, we used the hybrid algorithm which first segments speech in 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 whole utterance in the unit of CD_PLU and full paths within each CD_PLU.

RECOGNITION

Language Model

In our system, we adopted a 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 for some relations of the between-words, and can reduce search space in recognition stage, we used the FSN as a language model of our baseline system. Figure 3 (C) shows a part of the FSN, which consists of 9 finite state networks, for our task domain.

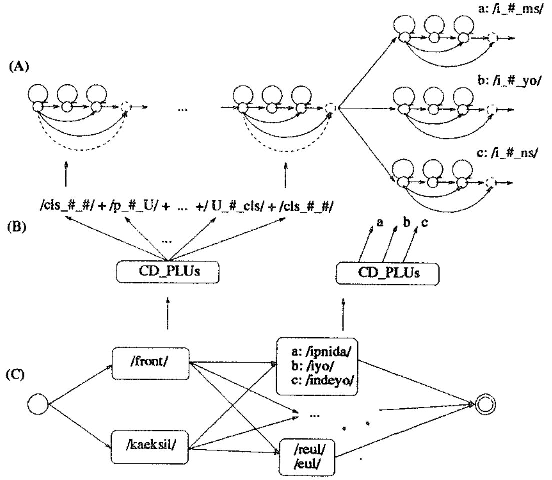

Figure 3: FSN for a situation, triphones as a CD_PLU, and expanded HMM states according to HMM topology. In (C), each node(box) means words which are used to construct all legal sentences in this situation, and the double circle means the final state.

Korean is an agglutinative language in which the postpositional word shows syntactic relation of its preceding words. We call this pair of words as *word-phrase* (ujul in Korean). Words in a *word-phrase* are pronounced like one word, and show strong coarticulation and we cannot treat them as separated words. To solve this problem, we used the inner-connection which is similar the between-word modeling (Lee *et al.*, 1990) as in Figure 4 c).

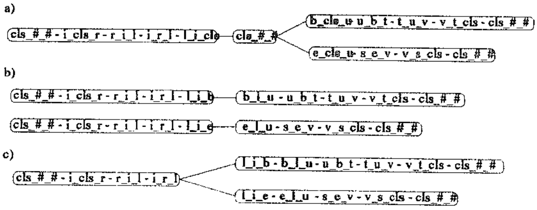
a)
cls_#_# - i_cls_r - r_i_l - i_r_l - l_i_cls
cls_#_#
b_cls_u u_b_t - t_u_v - v_t_cls - cls_#_#
e_cls_u s_e_v - v_s_cls - cls_#_#
b)
cls_#_# - i_cls_r - r_i_l - i_r_l - l_i_b
b_l_u - u_b_t - t_u_v - v_t_cls - cls_#_#
cls_#_# - i_cls_r - r_i_l - i_r_l - l_i_e
e_l_u - s_e_v - v_s_cls - cls_#_#
c)
cls_#_# - i_cls_r - r_i_l - i_r_l
l_i_b - b_l_u - u_b_t - t_u_v - v_t_cls - cls_#_#
l_i_e - e_l_u - s_e_v - v_s_cls - cls_#_#

Figure 4: The inner-connection. a) A part of FSN without between-word modeling. 'cls_#_#' is the silence model with null transition which lies between words except within *word-phrase*. b) With between-word modeling. c) The common part is merged by modifying the lexicon.

In recognition stage, one proper network of 9 FSNs is selected by a manage program according to the speaking situations between hotel front desk clerk and guest. By this situation control, recognition time is reduced.

Implementation of a continuous speech recognition system

situation	speaking pattern	perplexity
0	front desk or telephone exchange	3
1	yes or no	7
2	room number	7
3	reservation, reservation change, or reservation cancel	3
4	reservation period	4
5	surname	3
6	single, double, or twin room	3
7	room type	3
8	customer requests for situation 3	3

Table 2: Situations and perplexities according to customer's speaking patterns

We incorporate word and sentence knowledge into our recognizer in the following manner: Each word is represented as a network of CD_PLUs which encode the way the word is 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. Then we have a large HMM that encodes all the legal sentences.

By placing all the knowledge in the data structure 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 system(Han *et al.*, 1994).

In addition to 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 gives us CD_PLU, Word, FSN node information immediately. Therefore, it is convenient to picture an overall view in recognition stage. The field Fsn_node means the node code of a FSN. The Word_id means the real word code. The CD_PLU_id means the code of CD_PLUs which are every path through which the word can be pronounced. The State_id means the HMM state number of the large HMM network, and the In and the Out fields have some HMM state informations which are the incoming-to and outgoing-from present HMM state with linked list structure respectively.

In our system, we used Viterbi beam search as a search algorithm (Furui *et al.*, 1991) and used partial Viterbi backtrace in demonstration.

RESULTS AND ANALYSIS

To evaluate our system, we took two kinds of experiments. The first one is for isolated word recognition experiment shown in Table 3 , and the second is for continuous speech recognition experiment shown in Table . Both tables show the results of the SCHMM and the DHMM methods, respectively.

In Table 3, The SCHMM method is better than the DHMM method of speaker-independent experiments. We can also see the normalized logarithmic energy features take a good role in elevating the recognition accuracy. In speaker-independent continuous speech recognition experiments, the recognition results have various accuracy, from 86.2% to 98.1%, according to the 9 categories. The Table shows average recognition results.

The main reason for the obtained low recognition rate is due to the ambiguous pronunciations. For example, *aiemiyo and aiemyo* were sometimes replaced with *aiemyo and aiemiyo*, respectively.

method	Close		Open	
	DHMM FB	SCHMM SKM+FB	DHMM FB	SCHMM SKM+FB
Training 3, Recognition 3	1.8	5.1	16.2	13.4
Training 4, Recognition 3	2.3	5.3	15.7	13.5
Training 4, Recognition 4	1.5	4.4	11.1	9.5

Table 3: Speaker dependent and independent isolated-word recognition error rates in % on some feature combination. 4 means full 4 feature set, and 3 means feature set excluding normalized log power features.

method	WA (PC) with power feature	WA (PC) without power feature
SCHMM	89.0(91.2)	88.4(90.8)
DHMM	89.7(91.1)	88.7(90.5)

Table 4: Word accuracy (WA) and percent correct (PC) of our system. (unit: %)

CONCLUSION

We made a Korean continuous speech recognition system 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. Although it needs recognition performance enhancement, it will be a prototype system for a small domain continuous speech application.

REFERENCES

Kim, H.R., Hwang, K.W., Han, N.Y., and Lee, Y.J. (1993) '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* (In Korean).

Han, N.Y., Hwang, K.W., Kim, H.R., and Ahn, Y.M. (1994) 'An Implementation of Continuous Speech Recognition Using Finite State Network and Viterbi Beam Search', *Proceedings of the 1994 Korean Signal Processing Conference*, To be published (In Korean).

Labiner, L., and Juang, B.H. (1993) *Fundamentals of Speech recognition*, pp. 460-461, Prentice Hall: New Jersey.

Huang, X., and Jack, M. (1990) *Semi-Continuous Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U.K.

Huang, X., Alleva, F., Hon, H.W., Hwang, M., and Rosenfeld, R. (1992) *The SPHINX-II Speech Recognition System: An Overview*, CMU-CS-92-112, Jan 15.

Rabiner, L.R., Wilpon, J.G., and Juang, B.H. (1986), 'A segmental k-means training procedure for connected word recognition', *AT&T Technical Journal*, 65(3), 21-31.

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

Furui, S., and Sondhi, M.M. (1991) *Advances in Speech Signal Processing*, Marcel Dekker, Inc., pp. 623-650.