한-일 호텔예약 음성번역 시스템
- 한국 프론트데스크 측 -

이영직, 김영섬, 김회린, 류준형, 이정철, 한남용, 안영목, 최운천, 김상훈, 황규웅
한국전자통신연구소 음성언어연구실
305-600 대전직할시 유성구 유성우체국 사서함 106호
ylee@zenith.etri.re.kr

Korean-Japanese Speech Translation System for Hotel Reservation
- Korean front desk side -

Youngjik Lee, Young-Sum Kim, Hoi-Rin Kim, Joon-Hyung Ryoo, Jung-Chul Lee,
Nam-Yong Han, Young-Mok Ahn, Un-Cheon Choi, Sang-Hun Kim and Kyuwoong Hwang
Spoken Language Processing Section, ETRI P.O. Box 106, Yusung, Taejon, 305-600, Korea

ABSTRACT

Recently, ETRI developed a Korean-Japanese speech translation system for Korean front desk side in hotel reservation task. The system consists of three sub-systems each of which is responsible for speech recognition, machine translation, and speech synthesis. This paper introduces the background of the system development and describes the functions of the sub-systems.

1 INTRODUCTION

Three institutions which are Electronics and Telecommunications Research Institute(ETRI, Korea), Korea Telecom(KT, Korea), and Kokusai Denshin Denwa(KDD, Japan) have performed the cooperative research since 1991 to develop a Korean-Japanese speech translation system for hotel reservation task. They planned to perform the cooperative research from 1991 to 1997. Demonstration of the speech translation system was held on May in 1995 as an intermediate result of the seven year project. Each of three institutions has developed its own parts independently which comprises speech recognition, machine translation, speech synthesis. We, in ETRI, are responsible for Korean front desk side, KT for Korean customer side and KDD for both Japanese sides. This paper describes our work about speech recognition, machine translation, and speech synthesis briefly below.

2 SPEECH RECOGNITION

The speech recognition 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.

2.1 Task Domain and Vocabulary Set

The hotel reservation task has following vocabulary set. This set 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[1, 5]. 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 vocabulary set from predefined finite state network grammar. This grammar has 26 types for the data set.

2.2 Preprocessing

We use four codebooks, each with 256 entries, that use (1) 12 LPC cepstral coefficients; (2) 12 Δ LPC cepstral coefficients; (3) 12 Δ-Δ LPC cepstral coefficients; and (4)

Data set		Train	Test
front desk	Words	242 × 40 M	242 × 10 M
side	Sentences	93 × 40 M	93 × 10 M

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

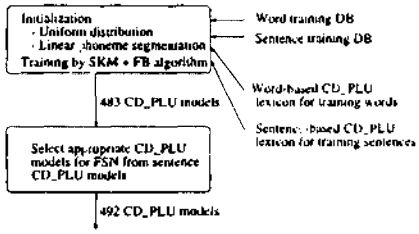

Figure 1: Training procedure for the front-desk side data set to estimate CD_PLU parameters.

normalized log power, Δ power, and Δ-Δ power. For end point detection, we use sequential method which is varied from [2] for demonstration. Speech is sampled at 16kHz and Preprocessings produce 20 dimensional LPC-cepstral coefficients which are then bandpass liftered.

2.3 Training

To train the SCHMM (Semi-continuous hidden Markov model), we used the hybrid algorithm which first segments speech into the unit of CD_PLU (context dependent phone like unit) by the Viterbi decoding and then uses FB (forward backward) 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[3], and our hybrid algorithm considers the best path in the inter-CD_PLU and full paths in the intra-CD_PLU. The training procedure is depicted in Figure 1.

2.4 FSN as a Language Model

In the developed systems, 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

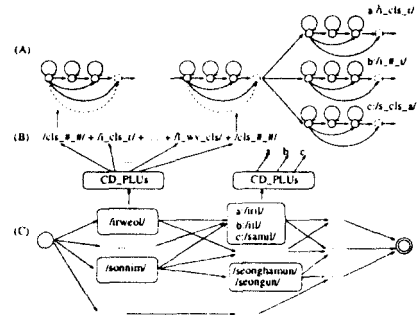

Figure 2: (A) Expanded HMM states according to HMM topology, (B) triphones as a CD_PLU, (C) FSN for a situation.

stage, we used the FSN as a language model of our baseline system. Figure 2 (C) shows one part of the FSN in our task domain. In Figure 2 (C), each node means words which are able to construct all legal sentences in this FSN[1, 5].

2.5 Implementation of a Continuous Speech Recognizer

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 2 (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 2 (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.

In our systems, we used Viterbi beam search [4] as a search algorithm and used partial Viterbi backtracking in demonstration. We considered the threshold value and the number of survived states as some constraints to prune HMM states.

To evaluate our system, we took two kinds of experiments. The first one is an isolated word recognition experiment whose results are 99.3% in the case of close experiment and 97.3% of open experiment. And the second is

a continuous speech recognition experiment. The accuracy are 97.6% in the case when including insertion error, and 97.8% when excluding insertion error. Above accuracy is obtained under speaker independent experiments.

3 MACHINE TRANSLATION

A speech translation system is inherently a conversational system. Inputs to translation system, i.e. outputs from speech recognizer can have various kinds of utterance style or vocabulary selection according to the opposite's response utterance. In some task domains, direct translation for utterance can cause meaningless result due to the difference of culture or pragmatics. If the interface between recognition system and translation system is done in symbolic level, not signal level (for example, N-best or 1-best output of FSN), a dictionary of word or phrase for translation is restricted to the vocabulary used in recognition. For example, some words in Korean recognition task do not have corresponding translation equivalents in Japanese task, and vice versa.

In other words, speech translation system should have translation module and dialogue flow control module. The former analizes recognition results and translates them. The latter analizes translation result and predicts dialogue flow and reflects well opposite's intention. In order to select the proper translation words according to conversion of analysis result, it is necessary to construct domain knowledge in detail.

It is important to construct detailed knowledge of dialogue following the conceptual categorization which cover inherent cultural tradition of each country or pragmatics, e.g. honorific expressions. Korean greeting, "안녕하십니까"(means "How do you do?" or "hello", ...), is not translated to consistent Japanese or English expression.

Interface with recognition system is basically first-best and semantic analysis has an hierarchical conceptual structure due to FSN node number attached to first-best output. The result of re-analysis of the opposite's utterance is assigned to each slot of frame which is classified by conceptual categorization of the corresponding domain.

Translated words are selected through referring to detailed knowledge expression and the slot of dialogue flow control module. Or, the Japanese expression for "안녕하십니까" is selected according to the time of a day.

Translation processes idioms by example based method.

4 SPEECH SYNTHESIS

The advanced Korean text-to-speech conversion system using TD-PSOLA(Time Domain-Pitch Synchronous Overlap and Add) technique, operates in real time on a UNIX workstation without any DSP board.

Speech synthesis system consists of language processing module, prosody processing module, and synthetic speech generation module. Language processing module performs text preprocessing, Korean functional word analysis, parser, prosody marker generation, and grapheme to phoneme conversion. Text preprocessing converts numerals, symbols, English and Chinese characters into Korean. Korean functional word analysis decomposes Korean particles, suffix inflections, adverbs, and conjunctions, and assigns one of 48 attributes eujeol by eujeol. Korean eujeols are composed of [noun+particle], [verb+suffix inflection], adverb, or conjunction. Parser builds phrase level syntax structure using the attributes. Prosody marker generation uses syntax information for each phrase to specify one of 13 prosody markers that influence the spoken output. Exceptional word dictionary provides pronunciations for exceptional words, and 26 Korean phonological rules converts grapheme to phoneme.

The prosody module calculates the duration of phonemes and the contour of fundamental frequency by rules based on the analysis of a few hundreds of speech data read by an announcer. The phrase level macro prosody uses the syntactic and positional information of prosodic phrases in a sentence. The syllable level micro prosody uses phonetic context and the position of syllable in a prosodic phrase.

The synthesis module selects the synthesis units(total 1226) from phoneme string, modifies and concatenates the synthesis units to generate speech output. The appropriate synthesis units for the phoneme sequence of each prosodic

phrase are selected depending on the phonetic and prosodic context. Each unit contains time domain data, pitch, and segment information necessary for TD-PSOLA application. The synthesis units includes the main variations of Korean phonemes. To decrease the concatenation defects, we use acoustic phonetic knowledge of Korean for spectral match. The prosodic parameters, i.e., duration, F0 and amplitude are directly scaled on the time-domain. We adopts unbalanced short time Hanning window to preserve original values at the pitch boundaries.

The advanced Korean T-t-S conversion system adopting TD-PSOLA technique shows much higher intelligibility and naturalness than the old one based on demisyllable units and Line Spectral Pairs.

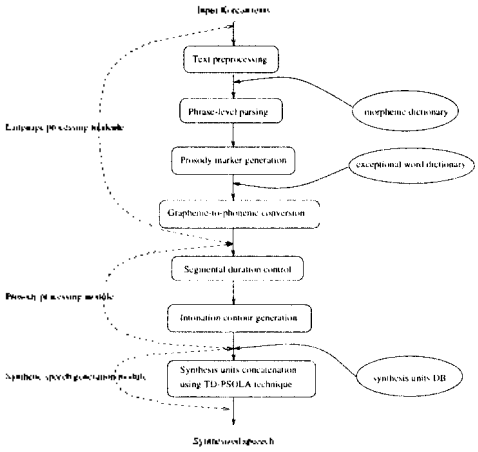

Figure 3: Text-to-speech conversion system: Geulsori-II

5 CONCLUSION

This paper describes a Korean-Japanese speech translation system for Korean front desk side in hotel reservation task. Korean continuous speech recognition system uses phone-based semi-continuous hidden Markov model(SCHMM) method for the automatic interpretation. A finite state network(FSN) grammar is adopted as a language model. Three features of the recognition system are following: an embedded bootstrapping training method, a hybrid parameter estimation method, a between-word modeling technique. The continuous speech recognition rate is 97.6% in the speaker-independent experiment . This system outputs a sentence with the best score to the machine translation system.

For the machine translation, an example-based translation method is used. Translation is performed through simple conversion in phrase level in most cases. But, semantic ambiguities are processed by invoking exceptional procedure according to dialog flow and information about translation equivalents in word dictionary.

The advanced Korean text-to-speech conversion system using TD-PSOLA(Time Domain-Pitch Synchronous Overlap and Add) technique, operates in real time on a UNIX workstation without any DSP board.

6 ACKNOWLEDGMENT

The authors would like to thank Dr. Jae-Woo Yang for valuable suggestions.

References

[1] H.R. Kim, K.W. Hwang, N.Y. Han, and Y.M.Ahn, "Korean Continuous Speech Recognition System Using Context-Dependent Phone SCHMMs," *Proc. SST-94*, Vol.II, pp. 694-699, 1994.

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

[3] 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.

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

[5] N.Y. Han, H.R. Kim, K.W. Hwang, Y.M. Ahn, and J.H. Ryoo, "A Continuous Speech Recognition System using Finite State Network and Viterbi Beam Search for the Automatic Interpretation," *Proc. ICASSP-95*, Vol. 1, pp. 117-120, 1995.

[6] Sang-Hun Kim, Minje Zhi, Un-Cheon Choi, and Hee-Il Han, "Application of TD-PSOLA technique T-t-S conversion," *Proc. IWSP-93*, 1993.9.