A Korean Flight Reservation System Using Continuous Speech Recognition

**Jong Ryong Choi, *Bum Koog Kim, *Hyun Yeol Chung, and **Seiichi Nakagawa

Abstract

This paper describes on the Korean continuous speech recognition system for flight reservation. It adopts a frame-synchronous One-Pass DP search algorithm driven by syntactic constraints of context free grammar(CFG).

For recognition, 48 phoneme-like units(PLU) were defined and used as basic units for acoustic modeling of Korean. This modeling was conducted using a HMM technique, where each model has 4-states 3-continuous output probability distributions and 3-discrete-duration distributions. Language modeling by CFG was also applied to the task domain of flight reservation, which consisted of 346 words and 422 rewriting rules. In the tests, the sentence recognition rate of 62.6% was obtained after speaker adaptation.

Key words : flight reservation, Continuous speech recognition, One-Pass DP, PLU, HMM, CFG, Language Modeling

Ⅰ. Introduction

There have been many studies on continuous speech recognition and as a result of a long period of study some practical speech understanding and recognition systems have been appeared in foreign countries, but only a few can be found in Korea because of lack of a large scale speech database inevitable to the study.

Recently, with the increasing concerns on multi-media communication through man-machine interface, the study for continuous speech recognition became also active in Korea. As a result, some useful systems were developed in the limited filed of application, such as speech dialing system by KAIST[1], stock information retrieval system by KT[2], automatic translation system by ETRI[3].

We are developing a Korean flight reservation system [4] using continuous speech recognition techniques for the task of air-line reservation flying between Korea and Japan. The system is based on the Japanese continuous speech understanding system SPOJUS-SYNO-X[5]. In this paper, we introduce recognition algorithms, speech data, and training of HMM models, which used in the first version of our system.

※This research was supported by the Yeungnam University Research Grants in 1996.

*School of Electrical and Electronics Engineering, Yeungnam University

**Department of Information and Computer Sciences, Toyohashi University of Technology

Manuscript Received July 24, 1996.

Ⅱ. System Overview

To begin with, let us introduce a Korean flight reservation system. The overall block diagram of the system is illustrated in Fig. 1. The recognition part of the first version of our Korean flight reservation system references the version X of SPOJUS-SYNO.

In recognition of sentences, due to the difficulties to get syntax control, many researchers generally adopt a method that get series of candidate words by spotting phoneme or word first and then find syntactically correct sentence candidates by using syntactic analysis[6]. This method is useful because evaluations for recognition and language processing can be carried out separately. But in this method, one more step for mid-output makes have lower recognition rate.

In the version X of SPOJUS-SYNO, to solve above problem, One Pass DP algorithm(OPDP)[7] was added. OPDP algorithm obtains an optimal solution efficiently in a frame-synchronous process, and can be modified to cope with syntactical constraints represented by a finite-state-automation. However, more flexible syntactical constraints, such as context-free grammar(CFG), are necessary for speech recognition process[5]. And thus, the CFG was also adopted in our system.

In language processing of the first version of Korean flight reservation system, we also incorporate a frame-synchronous parsing algorithm into One-Pass search algorithm to integrate language and acoustic processing.

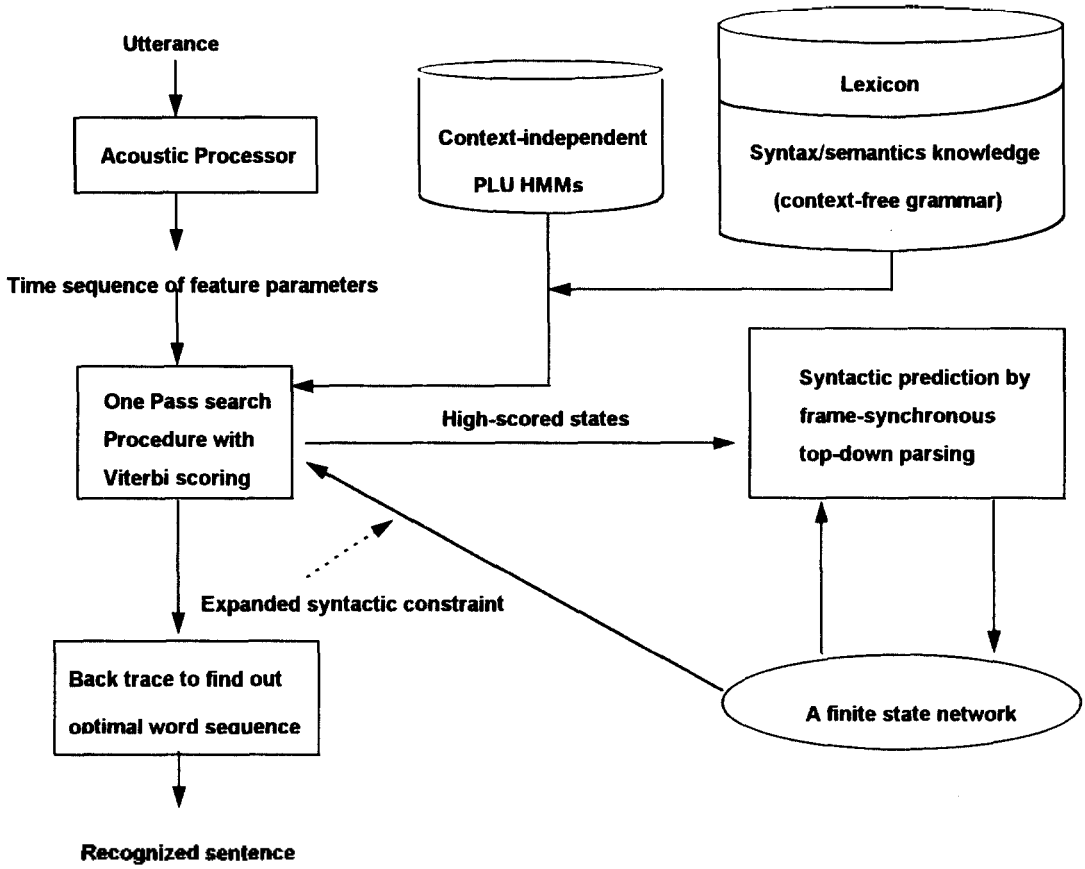

Figure 1. System overview.

Our parsing algorithm is based on top-down strategy which is similar to the Early algorithm. Some parts of this algorithm are modified for efficient left-to-fight parsing, making the computation loss derived from the parsing procedure relatively small. The top-down and breadth-first parsing has an advantage that allows to control extending the search space dynamically in each frame according to the cumulative likelihood score for each grammar state at the time[6].

Ⅲ. Flight Reservation Task

The task of our system is limited to the dialogs between a Korean customer and a flight reservation desk. It is also limited to an airline between Korea and Japan. Each dialog set includes 2-4 sets of talks for reserving, canceling, confirming, changing flight number and so on. Total dialog sets are 15. Fig.2 shows an example dialog for reservation.

In language processing, after grouping sentences along with the type of predicate and the type of sentence structure, the CFG is written with bottom-up method. Thus,

sys: 안녕하세요. 한국 항공입니다.
usr: 유월 오일 서울 토오쿄오간 편도 예약을 하고 싶습니다.
sys: 몇 시경 출발하실 예정이십니까?
usr: 열한 시 출발 KE 704편이 좋겠습니다.
sys: 퍼어스트 클래스, 비지네스 클래스, 이코노미 클래스가 있습니다만, 어느 쪽으로 하시겠습니까?
usr: 비지네스 클래스가 좋습니다.
sys: 성함과 전화번호를 말씀해 주세요.
usr: 이 팔 칠에 삼사구륙, 이 현명입니다.
sys: 영문 성함의 스펠을 불러 주세요.
usr: L E E H Y U N M Y U N G.
sys: 감사합니다. 확인하겠습니다. L E E H Y U N M Y U N G 틀림 없습니까?
usr: 예.
sys: 유월 오일 열한 시 서울 출발 십삼시 토오쿄오 도착 KE 704편 한 분으로 틀림 없습니까?
usr: 예.
sys: 예약 되었습니다. 당일은 출발 한시간 전에 공항에 나오셔서 표를 사 주십시오. 감사합니다.

Figure 2. An example dialog set for reservation.

non-terminal symbols or rules become unique and task dependent ones, regardless syntax. For example, ... 하고 싶다(... HAGOSIPTA) is written as S_SIPTA, and a noun clause as NP_EVENT. Fig.3 shows a part of CFG sentences written with these rules. Names, airports, English alphabets of a person's name are regarded as proper noun and registered previously. The CFG sentences consisted in this manner include 346 vocabularies, 197 non-terminal symbols, 126 word classes and 422 rules.

```
START          →  S000
S000           →  SS_SIPTA
SS_SIPTA       →  S_INSA S_SIPTA
S_INSA         →  INSA
S_INSA         →  ε
S_SIPTA        →  OP_EVENT_A VP_SIPTA_A
OP_EVENT_A     →  NP_EVENT YEYAK
                  JOSA_EUL
NP_EVENT       →  NP_EVENT_WAYS
NP_EVENT       →  NP_EVENT_NOWAYS
NP_EVENT_NOWAYS  →  NP_EVENT_NOWAYS2
NP_EVENT_NOWAYS2 →  NP_AP
NP_AP          →  AP FROM AP KAN
NP_AP          →  AP FROM AP
NP_AP          →  ε
VP_SIPTA_A     →  HAGO SIPTA
INSA           →  안녕하세요
YEYAK          →  예약
JOSA_EUL       →  을
HAGO           →  하고
```

Figure 3. Examples of CFG sentences.

Ⅳ. Speech Data

Fig.4 shows the flow of training and test processes of our system. For making an initial phoneme model(INIT-HMM) 611words(ETRI611) which have labeling information for phoneme(or sub-phoneme) boundaries are used. TUT611 and TUT200 uttered by 13 male persons (ages ranging from in their 20's to 40's) are also used for training and test as depicted in Table 1.

Table 1. Speech Data.

No. of speakers	611words (no. of words)	setA (no. of sentences)	setB (no. of sentences)
3	ETRI611 (3666)		
10	TUT611A (6110)	TUT200A (1000)	
3	TUT611B (1883)	TUT200C (300)	TUT200B (300)

Table 2. Analysis condition of speech data.

Sampling frequency	16 kHz
Hamming window	16ms(256 points)
frame rate	5 ms(80 points)
analysis	14 order LPC analysis
feature parameters	10 mel-cepstrum coef. + 10 regression coef.

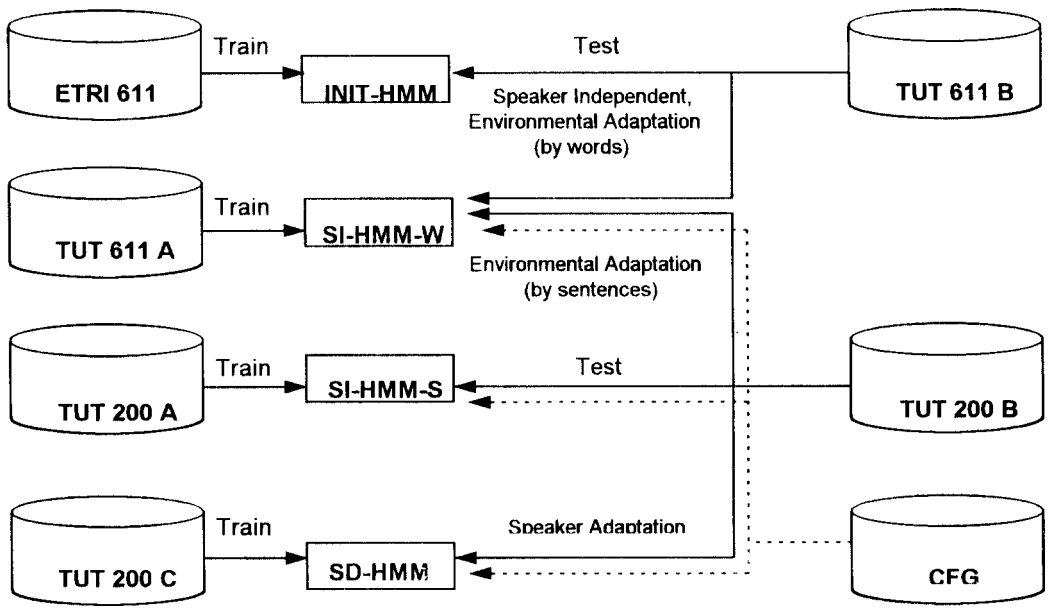

Figure 4. Flow of training and test.

We call a set of 100 sentences consisted for user 'setA' and another set of 100 sentences having nearly the same flow of dialog but different from setA in sentence structure or in vocabulary 'setB'. SetA and setB consist with 166 words and 186 words respectively. Among them 90 words are common. SetA is used for training and setB for test in the experiment.

All speech data were sampled at 16kHz with an accuracy 16bits, and 14 LPC cepstrum coefficients were then extracted from every 5ms. These cepstrum coefficients were transformed to 10 mel-cepstrum coefficients and to 10 regression coefficients. These parameters were used for features for recognition. The analysis conditions are illustrated in Table 2.

V. HMM Training

HMM acoustical model for each phoneme consists of continuous output discrete-time controlled HMM(DDC-HMM) with 4-state 3-output distribution. DDCHMM is an improved model of a basic HMM. It can present duration information with discrete distribution probability and makes possible to control time length to stay on a state. In our system maximum duration to stay on one state is limited to 200ms. To train differences of microphones, differences of utterance speed in pronouncing words and sentences, and differences of models for speaker independent and for speaker dependent, adaptation algorithm is used.

Initial 48 HMM phoneme models(INIT-HMM) are consisted by Baum-Welch algorithm with labeling information. Some of initial models are modified adequately because labeling information for some phonemes found inaccurate in ETRI611. For smoothing labeling inaccuracies this INIT-HMM is used for auto-segmentation of ETRI611 words and then the model is retrained using the same data. Next, by applying MAP algorithm[5] to these initial HMM models speaker independent model(SI-HMM-S) is constructed. And then, an additional training is conducted with data TUT611A for adapting recording environments(SI-HMM-W) caused by changes of microphone or AD converter and with TUT200A for differences in utterances between words and sentences. Finally, speaker adapted model(SD-HMM) is made from adaptation training of SI-HMM-S by estimated MAP with 100 sentences of set-A(TUT200C).

VI. Experiments and Results

To verify the effects of adaptation resulted from each step of training in Fig. 4, recognition and segmentation experiments for phonemes and for words are carried out. The effectiveness of environmental adaptation models are verified with O(n)DP algorithm for words(TUT611B). Table 3-4 show the results. From the tables we can find environmental adaptation raised word recognition accuracy from 31.0% to 57.5%.

Table 3. Phoneme recognition rate for initial models. (ETRI611 : for train, TUT611B : for test) (%)

Speaker	CJR	KHN	JYC	Average
Correct	27.2	33.1	32.6	31.0
Substitution	59.5	57.4	56.6	57.8
Insertion	31.2	43.5	40.3	38.3
Deletion	13.3	9.5	10.8	11.2
Seg. Rate	55.5	46.9	48.9	50.5

Table 4. Phoneme recognition rate for environmental adapted models(%). (ETRI611+TUT611A : for train, TUT611B : for test).

Speaker	CJR	KHN	JYC	Average
Correct	54.5	61.5	56.5	57.5
Substitution	38.6	33.6	37.2	36.4
Insertion	44.8	46.0	44.4	45.1
Deletion	7.1	5.0	6.3	6.1
Seg. Rate	48.1	49.1	49.4	48.9

But segmentation rate for phonemes appeared to be low due to many insertion errors caused by the phonemes having short duration. Where Seg. rate means that

$$Seg.\ rate = \frac{No.\ of\ input - No.\ of\ insertion - No.\ of\ deletion}{No.\ of\ input}$$

Next, sentence recognition experiments are carried out using frame synchronized One Pass algorithm.

For the test, 100 sentences having average of 5.06 words per sentence are used. The test set perplexity is about 44.4. From the tests 41.0% of sentence recognition rate was obtained by using SI-HMM-W(word adapted model) and 47.3% by SI-HMM-S(sentence adapted model), showing 6.3% of improvement.

Finally using SD-HMM(speaker adapted model) the best sentence recognition rate of 66.3% was obtained.

In error analysis of sentence recognition results, we found a lot of sentences that recognized incorrectly by

only one word as shown in Table 6. Table 6 also shows that the model repeatedly adapted by several kind of environment tends to give the better sentence recognition. Especially, there were much confusions between words such as 은/는, 으로/로, 육/륙, 이에요/예요. These confusions can be solved by applying pronunciation rules that reflect coarticulation. But, the confusions can be disregarded in understanding the meaning of sentences. From this point of view, we calculated the semantic understanding rate. It appeared average of 74%. Beside this problem, there were prominent recognition errors among the same word class, such as in digits. This means that our system needs the better adequate acoustic model and the more training data for recognition.

Table 5. Sentence recognition rate(%).

Environmental adaptation	Speakers			Average
	CJR	KHN	JYC	
by words	44.0	33.0	46.0	41.0
by sentence	49.0	39.0	54.0	47.3
by speaker	67.0	63.0	69.0	66.3

Table 6. Number of sentences recognized incorrectly with only one word.

Environmental adaptation	Speakers			Average	Rates (%)
	CJR	KHN	JYC		
by words	11	16	12	13.0	22.0
by sentence	14	13	11	12.6	24.0
by speaker	13	13	16	14.0	41.6

Ⅶ. Conclusions

A Korean continuous speech recognition system for flight reservation was consisted based upon Japanese continuous speech understanding system SPOJUS-SYNO-X. This system are using a frame synchronous One-Pass DP search algorithm driven with syntactic constraints by context free grammar.

For the recognition, 48 phoneme-like units(PLU) were defined and used as basic units of acoustic modeling of Korean. This modeling was conducted using a HMM technique, where each model has 4-states, 3-continuous output probability distributions, and 3-discrete duration distributions. Language modeling by CFG was also applied to the task domain of flight reservation, which consisted 346 words and 422 rewriting rules. The perplexity for 100 test sentences was calculated to be 44.4. In the tests 62.6% of best recognition rate was obtained after speaker adaptation.

In error analysis of sentence recognition results, we found a lot of sentences had recognized incorrectly just one word. semantic understanding rate was average of 74%. There were prominent recognition errors among the same word class, such as digits. This means that our system needs a more adequate acoustic model and a more training data for recognition.

References

1. S.H.Choi et al, "Continuous digit recognition for a real-time voice dialing system using discrete Hidden Markov models," Proc. 5th WESTPRAC, pp.1027-1031, 1994.
2. Koo et al., "KT-Stock : a speaker-independent large vocabulary speech recognition system over the telephone," Proc. ICSLP, pp.1387-1390, 1994.
3. Y.J.Lee et al., "Korean-Japanese speech translation system for hotel reservation -Korean front desk side-", Proc. 12th Speech Communication and Signal Processing Workshop, pp.204-207, 1995.
4. Jong Ryong Choi, Hyun Yeol Chung and Seiichi Nakagawa, "A Continuous Speech Recognition System for Air Line Reservation," Proc. 1995 IEEE Asia-Pacific Workshop on Mobile Telecommunication, pp.119-122, 1995.
5. A.Kai and S.Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," Proc. ICSLP 92, pp.119-121, 1992.
6. A.Kai, "A study on frame-synchronous continuous speech recognition algorithms using context-free grammar," Master's thesis, Toyohashi Univ., Japan, 1993.
7. H.Ney, "Dynamic programming parsing for context free grammars in continuous speech recognition," IEEE Trans., Vol.3, No.3, pp.336-340, 1990.

▲Jong-Ryon Choi

Jong-Ryon Choi was born in Taegu, Korea, on January 1, 1964. He received his B.C. and M.S degree in the department of electronics engineering from Yeungnam University, in 1990 and 1992, respectively. He recently working on his Dr. degree in the Department of Information and Computer Sciences at Toyohashi University of Technology, Japan.

His research interests are speech analysis, speech recognition and digital signal processing.

▲Bum-Koog Kim

Bum-Koog Kim was born in Kyungnam, Korea, on December 26, 1964. He received his B.C. degree in the department of mathematics, in 1990, D.C. degree in the department of electronics engineering from Yeungnam University, 1992. He currently working on his Dr. degree in the school of electrical and electronic engineering at Yeungnam University.

His research interests are speech analysis, speech recognition and digital signal processing.

▲Hyun-Yeol Chung

Hyun-Yeol Chung was born in Kyungnam, Korea, on November 26, 1951. He received his B.C. and M.S. degree in the department of electronics engineering from Yeungnam University, in 1975 and 1981, respectively, and the Ph.D of engineering degree in Information Sciences from Tohoku University, Japan, in 1989. Since 1989 he has been with Yeungnam University, where he is a professor in the school of electrical and electronic engineering. During 1992 to 1993, he was a visiting scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, USA. He was a visiting scientist in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan in 1994.

His research interests are speech analysis, speech recognition and synthesis and digatal signal processing.

▲Seiichi Nakagawa

Seiichi Nakagawa was born in Kyoto, Japan, in 1948. He received his B.E. and M.E. degree from Kyoto Institute of Technology, in 1971 and 1973, respectively, and a Dr. of engineering degree from Kyoto University in 1977. He joined the reserach associate in the Department of Information and Sciences. From 1980 to 1983 he was an assistant professor, from 1983 to 1990 he has been a professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. During 1985 to 1986, he was a visiting scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, USA. He is the author of the book *"Speech recognition Based on Stochastic Model"* (Inst. Elect. Infor. Comm. Engrs. Japan, 1988)

Dr. Nakagawa was a corecipient of the 1977 paper award from the IEICE and the 1988 JC bose memorial award from the Institution of Electro. Telecomm.