A BIMODAL KOREAN ADDRESS ENTRY/RETRIEVAL SYSTEM

Hyun-Yeol Chung, Cheol-Jun Hwang, and Shi-Wook Lee

Department of Information and Communication Engineering,
Yeungnam University
214-1 Dae-Dong, Kyongsan, Kyongpuk 712-749, Korea
{ chy ,hcj, lsw }@speech.yeungnam.ac.kr

ABSTRACT

This paper describes the development of a Korean address entry/retrieval system using bimodal input; speech recognition and touch sensitive display. The system works on a personal computer and employs automatic speech recognition and touch sensitive display techniques as user interface for input Korean address, which consisted with about 40,000 words. To meet the needs that practical speech recognition system should be worked in real time without any degradation of performance of recognition accuracy, 1)speaker and environmental adaptation by Maximum a posteriori (MAP) estimation were adopted for higher recognition and 2)fast search by tree-structured lexicon and frame synchronous beam search technique were employed for real time response. To offer more convenient user interface, touch sensitive display is also implemented. As the results, the system worked in 3 seconds after completion of address utterance with sentence recognition accuracy of above 96%.

1. INTRODUCTION

With the growth of multimedia and high-speed information network technologies, businesses through communication network have been increased remarkably. Furthermore, with the growing demands for the high quality of services in sales and for managing the customer's information, and for the delivery of commodities etc., address inputs to the personal computer are now common. This kinds of data input by hand could be processed fast and effectively by employing multi-modal interfaces such as speech, gesture, handwriting, face tracking and so on. Multi-modal interface technique to make this kind of work possible has been made considerable progress over the last few years and many systems were developed so far; for example, SPHINX-II[1], JANUS-III[2] of Carnegie Mellon university, DRAGON system of Stanford university, SUMMIT and GALAXY of MIT, ASURA of ATR Interpreting Telecommunications Research Labs, Japan, SPOJUS-SYNO[3] of Toyohashi university of Technology, Japan.

But there has been reported very few such a system in Korea. The basic idea of the development of an address entry /retrieval with bimodal input interfaces, speech recognition and touch sensitive display technique, working on a personal computer is based on the techniques resulted from these practical system

2. SYSTEM OVERVIEW

The system works on a personal computer with a built in sound card and mouse. Users can input address using not only bimodal inputs by speech or touch sensitive display but also keyboard or mouse. Figure 1 illustrates the overall block diagram of the system. In this system, if speech recognition for upper-class address component is completed, the system displays lower-class candidates belong to the address set. It can be possible because Korean address is consisted by a hierarchical order.

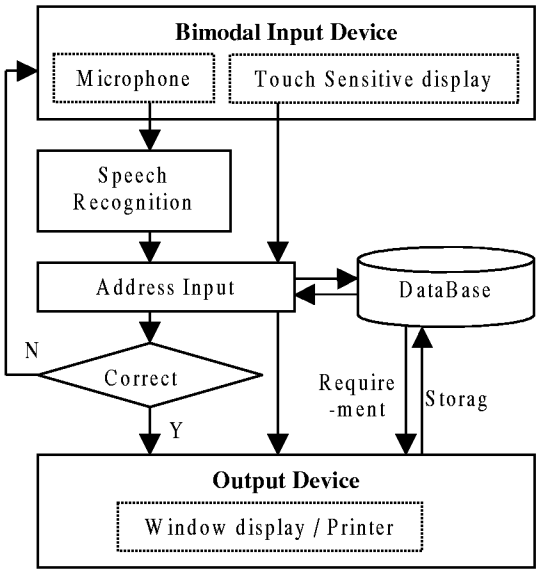

Figure 1: System overview of Bimodal Korean address entry /retrieval system.

3. SPEECH DATA

All speech data are sampled at 16KHz with accuracy 16bits, pre-emphasized with $1-0.98z^{-1}$ and divided into frames of 16msec at a rate of 5msec. Each frame is multiplied by a Hamming window with 16msec(256samples). From these smoothed speech samples, 14 LPC coefficients are extracted by auto-correlation method. A set of 10 MFCCs is then computed from the LPC coefficients and are used as feature parameters. We added 10 regression coefficients (RGC) as dynamic features [4].

445 phoneme balanced Korean words uttered by 14 male in sound proof booth with phoneme-labeled information are used for training initial HMMs. Utterances by 3 male with 2 different microphones in office are used for adaptation and recognition. Table 1 shows the speech data and system environments.

Speaker (Number)	Male(14)	Male(3)	
Utterance type(#)	PBWs(445)	Connected Word(25)	Connected Word(75)
# of utterance	1	1	1
Usage	Training	Adaptation	Recognition
Environ -ment	Sound Proof booth	Office	
Record device	DAT Recorder	PC(Sound Card)	
Micro -phone	Dynamic Headset	Dual-channel, Condenser Desktop Dynamic Headset	

Table 1: Speech data for training, adaptation and recognition.

4. HMM TRAINING AND ENVIRONMENTAL ADAPTATION

Each PLU is characterized by a left-to-right Continuous Hidden Markov(CHMM) model of 4-states 3-output probability with a full covariance matrix without skip.

Speech recognition systems tuned to the speech uttered in a specific environment tends to have inevitable degradation of performance when they are used in some other environments due to variations in speech characteristics [5]. To adapt CHMMs to the different environment, MAP estimation is used. It can update all of the acoustic parameters directly. This is called incremental adaptation because the adaptation data is available to the system sequentially, i.e., the adaptation algorithm continues adapting the system to the new environment as long as further adaptation data becomes available. MAP estimation is defined as equation (1). Using equation (1), it is possible to reestimate adaptation parameters even if only one sample is given.

$$\max_{\Theta} P(\Theta \mid X_1,\cdots,X_N) = \max_{\Theta} \frac{P(X_N \mid X_1,\cdots,X_{N-1},\Theta)P(\Theta \mid X_1,\cdots,X_{N-1})}{\int P(X_N \mid X_1,\cdots,X_{N-1},\Theta)P(\Theta \mid X_1,\cdots,X_{N-1})d\Theta} \quad (1)$$

Where, $X_1, X_2,,,\cdots ,X_N$ indicate N samples.

5. FAST SEARCH ALGORITHM

5.1. Tree-structured lexicon

Since the pronunciation of two or more words contain the same n initial phonemes, they share a single sequence of n HMM models representing the initial portion of their pronunciation[6]. It takes an important role for the improvement of performance of fast search technique. And it is better to construct the HMMs to be searched as a phonetic tree than to construct a flat structure of independent linear HMM sequences for each word. An example of tree-structured lexicon for connected word recognition used in our system is shown in Figure 2.

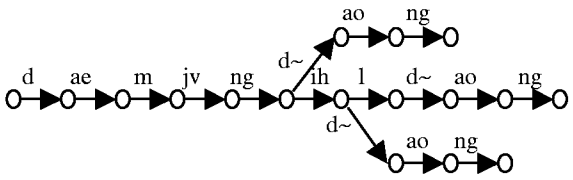

Figure 2: An example of tree-structured lexicon for connected word recognition.

5.2. Beam search and Pruning

Speech recognition, searching for the most likely sequence given the input speech, gives rise to an exponential search space if all possible sequences are considered. Search space can be reduced remarkably by using the beam search technique since beam search evaluate relatively few partial paths for each frame. Therefore, most of paths that their extensions are not guaranteed are pruned [7].

When (i, j^*) is the point with the best path at i, the path to any (i, j) is a candidate for extension at frame $i+1$ only if the cost satisfy.

$$P_{\max}(i,j) \le P_{\max}(i,j^*) + \lambda \qquad (2)$$

The main problem in beam search method is the likelihood accuracy between acoustic model and each candidate. Since applying strict condition of beam width and pruning threshold may prune the path in the middle of process in path selection even if the path could be optimal at the end of frame, the condition need to be simpler for guaranteeing optimal path. However, it enlarges the search space and requires more searching time. To solve this problem, a frame-synchronous variable(FSV) pruning threshold is adopted to get more flexible restricted search space. Frame-synchronous variable pruning threshold is that the threshold varies according to the process of search. That is, the threshold value is relatively simple at the start frame but it becomes more restricted at end frame.

6. EXPERIMENTAL EVALUATION

6.1. Baseline performance

One Pass Dynamic Programming(OPDP) algorithm is used for recognition[8]. The algorithm requires less memory spaces and less searching time. The baseline address(consisted by connected words) recognition tests using MFCC's and RGC's as feature parameters were conducted. The tests were performed by off-line on the workstation. The recognition results for addresses (connected word recognition rate; CWRR) and for each words consisting addresses(word recognition rate; WRR) by 3 male speakers were shown in Table 2. The tests are carried out separately with two different microphones, condenser desktop and dynamic headset.

	Condenser desktop		Dynamic Headset		Total
Speak -er	CWRR (%)	WRR (%)	CWRR (%)	WRR (%)	CWRR (%)
mj	86.7	96.0	88.0	96.0	87.3
ks	92.0	96.9	84.0	93.3	88.0
cj	85.3	94.2	88.0	96.0	86.7
Total	88.0	95.7	86.7	95.1	87.3

Table 2: Average recognition rates for CWRR and WRR for different microphones.

Average connected word recognition rate was 87.3% for the initial HMM training only by 14 male speakers. According to different speakers and microphones, 1.3-8.0% of differences in recognition rates was observed.

6.2. Adaptation using MAP estimation

As shown in Table 2, by changing microphones and speakers the performance of system was degraded. To be robust against the degradation by these environmental variations and to get an acceptable performance for practical use, the initial HMM are re-trained using MAP estimation. The adaptation was carried with 25 connected words for each speaker. Table 3 shows the results.

	Condenser desktop		Dynamic Headset		Total
Speak -er	CWRR (%)	WRR (%)	CWRR (%)	WRR (%)	CWRR (%)
mj	93.3	97.8	93.3	97.8	93.3
ks	94.7	98.2	96.0	98.7	95.4
cj	100.0	100.0	98.7	99.6	99.4
Total	95.9	98.7	96.0	98.7	96.0

Table 3: Results of MAP estimation for different microphones and different speakers.

From the result, connected word recognition rates after adaptation increased 2.7 to 14.7% according to the microphones and speakers.
It can be reduced average recognition difference gap by 0.1% for three speakers in connected word recognition, presenting the effectiveness of an environmental adaptation using Maximum a posterior estimation in our task.

6.3. Results of Fast search

To prove the effectiveness of FSV pruning threshold, search and acoustic probability computation time were measured for the utterances by 3 speakers for 75 addresses(connected words). This experiment was carried out on the workstation(Ultra-sparc1:167MHz) to get reliable CPU time.
We measured computation time for several pruning thresholds with a fixed beam width (BW=10). The results were then compared with those from FSV pruning thresholds. Table 4 shows computing time and recognition results according to the initial pruning threshold(Init. Pr_th), variation step(frame threshold), and variation value.

Microphone		Desktop			Headset		
	Pr_th	CWRR (%)	WRR (%)	Time (sec)	CWRR (%)	WRR (%)	Time (sec)
Full search		96.0	98.5	557.3	96.0	98.7	498.5
Ada-ptati-on	-500	96.0	98.7	6.28	96.0	98.9	5.43
	-400	96.0	98.5	5.21	96.0	98.7	4.65
	-300	95.7	98.5	4.5	96.0	98.7	4.06
	-200	94.8	98.4	3.76	96.0	98.7	3.4
	-100	87.6	95.4	3.23	88.0	96.0	2.96
	FSV1	93.3	97.8	4.13	94.7	98.2	3.81
	FSV2	96.0	98.7	5.06	96.0	98.7	4.53
	FSV3	88.8	96.3	3.35	88.4	96.1	3.09
	FSV4	95.7	98.5	4.47	96.4	98.8	4.0
	FSV5	92.9	97.6	3.61	93.3	97.7	3.39

Table 4: Computation time and recognition rates.
FSV1: Init. Pr_th -500, variation step 50, variation value 50
FSV2: Init. Pr_th -500, variation step 100, variation value 50
FSV3: Init. Pr_th -400, variation step 50, variation value 50
FSV4: Init. Pr_th -400, variation step 100, variation value 50
FSV5: Init. Pr_th -300, variation step 100, variation value 5

By using a tree-structured lexicon mentioned in 5.1 and beam search with pruning (BW=10, Pr_th=400) in 5.2, recognition speed is increased to a factor of 100 without any degradation of recognition rate.
From table 4, although FSV pruning threshold is not so effective in saving time, but it was found helpful in keeping the optimal path reliable.

6.4. Correction of recognition errors

To correct word errors in spoken address recognition, touch sensitive display is adapted. With top-down menu on the touch sensitive display, word errors could be corrected with 100% of accuracy in 100 test experiments, showing the effectiveness of touch sensitive display.

7. CONCLUSION

We have introduced the Korean address entry/retrieval system having bimodal input; automatic speech recognition and touch sensitive display. To make our system to be robust to the degradation of performance caused by the variation of environments and to get acceptable performance in practical use, MAP adaptation technique was adopted. After environmental adaptation, we achieved average recognition rate of 96.0% for addresses and 98.7% for words.

To make the system working in real time, beam search with pruning and tree-structured lexicon were adopted. With beam search with pruning threshold(BW=10, Pr_th=400) and tree-structured lexicon the system performed a 100-fold speedup over full search algorithm without any loss of word recognition rate.
Although FSV pruning threshold is not so effective in saving time, but it was found helpful in keeping the optimal path reliable.
To correct word errors in spoken address recognition, touch sensitive display is adapted. With top-down menu on the touch sensitive display, word errors could be corrected with 100% of accuracy in 100 tests.

8. PREFERENCES

1. Alleva. F., et al., "Applying SPHINX-II to the DARPA Wall Street Journal CSR task," *Proc. of Speech and Natural Language Workshop*, 393-398, 1992.
2. Alon Lavie, et al., "JANUS-III: Speech-to-speech translation in multiple languages," *Proc. IEEE ICASSP97*, Vol.1, 99-102, 1997.
3. A. Kai, S. Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," *Proc. ICSLP92*, 257-260, 1992.
4. X.D. Huang, Y. Ariki, M.A. Jack, "Hidden Markov Models for Speech Recognition," *Edinburgh Univ.*, 1990.
5. J.C. Junqua, J.P. Haton, "Robustness in automatic speech recognition," *Kluwer Academic Publishers*, 1996.
6. P.S. Gopalakrishnan, L.R. Bahl, and R.L. Mercer, "A tree search strategy for large-vocabulary continuous speech recognition," *Proc. IEEE ICASSP95*, Vol.1, 572-575, 1995.
7. John R. Deller, Jr., John G. Proakis, and John H.L. Hansen, "Discrete-Time Processing of Speech Signals," *Macmillan Publishing Company*, 1993.
8. S. Nakagawa, "Speech recognition Based on Stochastic Model," *IEICE*, Japan, 1988.