An Experimental Study on Barging-In Effects for Speech Recognition Using Three Telephone Interface Boards

Sung-Joon Park[*] · Ho-Kyoung Kim[*] · Myoung-Wan Koo[*]

ABSTRACT

In this paper, we make an experiment on speech recognition systems with barging-in and non-barging-in utterances. Barging-in capability, with which we can say voice commands while voice announcement is coming out, is one of the important elements for practical speech recognition systems. Barging-in capability can be realized by echo cancellation techniques based on the LMS (least-mean-square) algorithm. We use three kinds of telephone interface boards with barging-in capability, which are respectively made by Dialogic Company, Natural MicroSystems Company and Korea Telecom. Speech database was made using these three kinds of boards. We make a comparative recognition experiment with this speech database.

Keyword : Speech recognition, Echo cancellation, Barging-in

1. Introduction

In speech recognition systems, barging-in is important for more convenient interaction between users and systems. To allow barging-in, the echo canceller is needed, which is also a critical and central component of a network from the standpoint of voice quality. Echo is caused by impedance mismatches in the local loop. Impedance mismatches can occur when a telephone is connected to the local loop. There can be many impedance mismatches in the local loop, resulting in many echoes. The degree of echo cancellation is various according to the desired performance levels. The method for meeting the standards is left up to the developer.

This paper focuses on how barging-in and echo cancellation affect speech recognition rate. In other words, we study how the speech recognition rate varies depending on whether barging-in is allowed or not and depending on echo cancellation techniques if barging-in is allowed.

We carried out several tests using three kinds of telephone interface boards that perhaps have different echo cancellers respectively. The first board is D/160SC-LS of Dialogic. Another board is AG2000 of Natural MicroSystems. We have no specific information about

[*] Multimedia Technology Laboratory, Korea Telecom

what echo cancellation techniques are implemented on these boards and just use the echo cancellers offered as a library. The last board is made by Korea Telecom. We used TMS320C30 DSP (digital signal processor). Echo canceller is implemented using LMS (least mean square) algorithm [1].

Our goal is to compare and analyze the speech recognition results in three systems for each case of barging-in mode and non-barging-in mode. To this end, we collected speech data on each of the three telephone interface boards for both cases. The system and the details of the database are described in the following sections.

In the first experiment, we perform separate tests on each board for both modes using data collected on the same boards. The next experiment is a cross-checking, which means that data collected on a board are applied to the models that have been created using data collected on other boards.

In the final section, we make our concluding remarks.

2. System Overview

The system with which experiments were made is outlined in Fig.1.

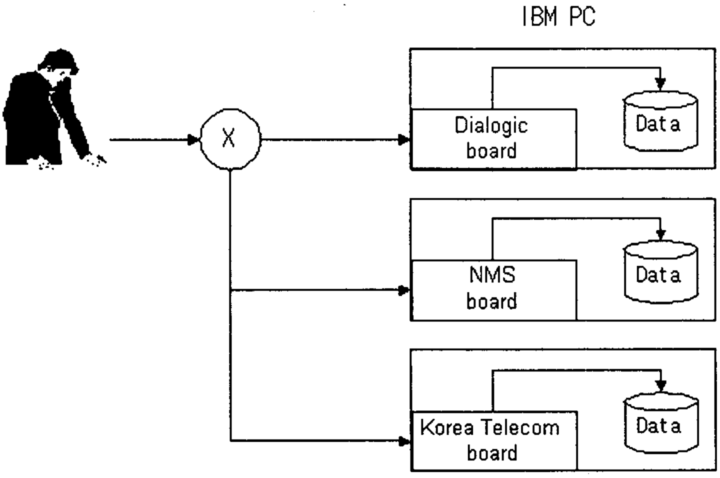

Fig. 1 The baseline system for experiments

The architecture of the three subsystems are basically the same except that each one use different boards and different echo cancellers. In the Dialogic board and the NMS board, echo cancellers always work not depending on whether or not the start point of speech is detected. On the other hand, the echo canceller in KT (Korea Telecom) board works just until the start point is detected. After the start point is detected, the announcement is stopped and the echo canceller does not function any more. The echo canceller and endpoint detector constitute the front-end part of a few speech recognition systems developed by Korea Telecom [2, 3].

The subsystems have different phone numbers respectively and speakers call three times to each subsystem for the same word and each mode. Accordingly, there may be variations of speech even for the same word because speakers may say the same word a little differently depending on the condition of their health and circumstances. This leads to incomplete comparisons of results because speech inputs to the boards become different although they should be the same. To avoid this problem, it is necessary to synchronize the speech inputs and outputs of all the subsystems, and speakers should use just one phone number. When speakers call, all the subsystems should operate in the same way to the eyes of speakers. For example, while each subsystem plays an announcement respectively, it should be guaranteed that speakers think they hear just one announcement. But it is not easy to implement synchronization. While some subsystems are recording speech input, others may play an announcement speech. For this reason, we adopt a simple architecture like the one described in this paper. But this problem should be solved for the more accurate experiments.

3. Experimental Conditions

We experimented with PBW (phonetically balanced words) database. The total number of words is 162 and 10 speakers read one time for each barging-in mode and non-barging-in mode for each board. Speakers consist of 5 males and 5 females. Accordingly, 1,620 utterances were collected for each mode and each board. In barging-in mode, speakers read words while an announcement is being played by the system. In non-barging-in mode, speakers read after a prompting signal. The systems were set up in the laboratory and speakers called at home using wire telephones.

Speech was sampled at 8 kHz and stored in the form of 2 byte linear data. The sampled speech is segmented into frames that span 20 msec. and are overlapped by 10 msec.

Signal processing consists of deriving a 12-th order LPC (linear predictive coding) cepstral coefficients, 12-th order delta coefficients, 12-th order acceleration coefficients, delta energy, and acceleration energy from each frame.

We use continuous HMM (hidden Markov model) and selected PLUs (phonelike units) as basic units. The total number of PLUs is 60 and they are context-independent units. The topology of the phone model follows that of Lees [4] and has 7 states and 12 transitions. 2,592 utterances of 4 males and 4 females were used for training and the rest were used for testing.

4. Results

We performed several tests with the databases described in the previous section.

The first one is a separate test for barging-in model and non-barging-in model. In the barging-in mode, 1,296 barging-in utterances were used in training and 126 barging-in utterances were used in testing for each board. In non-barging-in mode, non-barging-in data were used in training and testing. The result is summarized in Table 1.

Table 1. Recognition accuracy for barging-in and non-barging-in models

Model (train data)	Test Data \ Mixture #	1	2	3	4	5	6	7	8
KT(barging-in)	barging-in	69.14	85.19	87.04	88.89	90.74	87.65	87.65	87.96
KT(non-barging-in)	non-barging-in	57.10	79.94	84.57	88.89	90.74	88.27	88.27	86.42
NMS(barging-in)	barging-in	50.93	76.85	83.64	86.11	87.04	87.65	87.04	86.11
NMS(non-barging-in)	non-barging-in	32.72	67.28	75.62	79.63	80.25	84.88	83.02	83.64
Dialogic(barging-in)	barging-in	77.47	89.51	89.81	89.20	90.43	89.81	88.89	88.27
Dialogic(non-barging-in)	non-barging-in	74.69	83.95	87.04	87.65	86.73	87.35	85.19	83.95

This table shows that KT and Dialogic are better than NMS. Another feature is that NMS's recognition rate changes rapidly with mixture. But it is too hasty that we draw any conclusions just from this table. So we move to the next test.

In the second test, barging-in utterances and non-barging-in utterances were all used in training for each board, but test data are the same as in the first test. The aim of this test is to compensate for probable effects of insufficient training data and to compare the result with that of the previous test.

In table 2, we can find that the recognition rate is higher than those in Table 1. It may be because the training data was doubled. On the other hand, there is no big difference in speech recognition rate for barging-in mode and non-barging-in mode, which is similar to the result of the first test.

Another characteristic is related to NMS board. NMS board shows the best performance in non-barging-in mode and the recognition rate in non-barging-in mode is better than in barging-in mode compared to the result in Table 1. Fig.2 depicts how the speech recognition rate changes as the number of mixture increases.

Here we carefully reach the conclusion that there is actually no effect of echo cancellation in speech recognition rate and NMS board is more sensitive than other boards.

Table 2. Recognition accuracy for models using mixed train data

Model (train data)	Test Data \ Mixture #	1	2	3	4	5	6	7	8
KT (barging-in + non-barging-in)	barging-in	70.37	89.20	91.67	93.21	93.52	93.21	93.21	92.28
	non-barging-in	72.22	84.88	88.27	90.43	89.81	91.05	91.98	92.28
NMS (barging in + non-barging-in)	barging-in	58.64	75.00	82.72	87.96	89.20	90.43	91.98	91.67
	non-barging-in	51.85	74.69	84.57	89.81	91.67	93.52	92.90	93.83
Dialogic (barging-in + non-barging-in)	barging-in	79.01	89.20	91.98	91.36	91.05	92.28	91.05	90.12
	non-barging-in	77.78	88.89	89.81	91.36	92.28	92.28	93.21	92.28

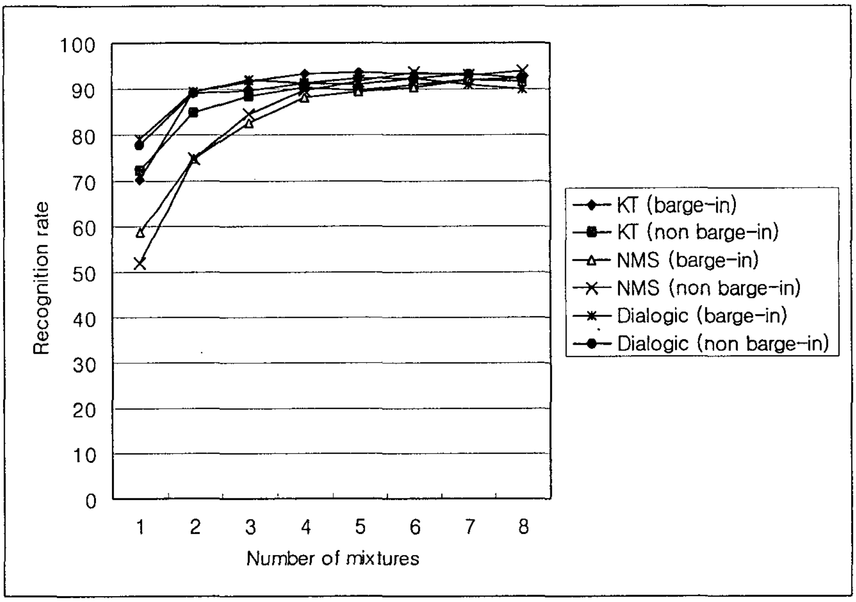

Fig.2 Variation of values in Table 2 according to the mixture number

We perform more tests. Data collected on a board are applied to the models that have been created using data collected on other boards. For example, if data are collected on KT board, they are applied to the models based on the data of NMS and Dialogic.

Table 3 is the result of using separate train data for barging-in mode and non-barging-in mode. Models which show the best result were selected from the Table 1.

Table 3. Recognition accuracy of barging-in and non-barging-in models using test data from all boards

Model (train data)	Test Data \ Board of Test Data	KT	NMS	Dialogic
KT(barging)	barging-in	90.74	83.02	85.80
KT(non-barging)	non-barging-in	90.74	82.10	85.80
NMS(barging)	barging-in	84.88	87.65	87.35
NMS(non-barging)	non-barging-in	84.57	84.88	80.56
Dialogic(barging)	barging-in	89.20	87.04	90.43
Dialogic(non-barging)	non-barging-in	85.19	86.73	87.65

Table 4 is the case when mixed train data were used.

Table 4. Recognition accuracy of models using mixed train data for test data from all boards

Model (train data)	Test Data \ Board of Test Data	KT	NMS	Dialogic
KT (barging + non-barging)	barging-in	93.52	86.42	89.20
	non-barging-in	92.28	91.36	89.81
NMS (barging + non-barging)	barging-in	89.20	91.98	89.20
	non-barging-in	91.98	93.83	91.98
Dialogic (barging + non-barging)	barging-in	92.90	91.67	92.28
	non-barging-in	88.89	89.51	93.21

From Table 3 and 4, we can see that HMM for a board should be created using the data collected on the very board for better speech recognition rate.

5. Conclusions

We performed several experiments on a speech recognition system for barging-in and non-barging-in utterances. From the experiments, we came to a few conclusions.

There is no actual difference of speech recognition rate between barging-in model and non-barging-in model. Models using mixed training data show better recognition accruracy. This is probably because of data quantity. The NMS board is more sensitive than the other boards. A board should use a model created using the data collected on the very board for better recognition rate.

References

[1] B. Widrow and S. D. Stearns. 1985. *Adaptive signal processing,* Prentice Hall, Inc. Englewood Cliffs, New York.

[2] M. W. Koo et al., 1995. "A stock information system over the telephone network," Proceedings of the 6th International Conference on Signal Processing Applications & Technology, pp. 2039-2043.

[3] Sung-Joon Park, Myoung-Wan Koo, Chu-Shik Jhon. 1999. "An implementation of continuous speech recognition for a stock information retrieval system," Proceedings of International Conference on Speech Processing, pp. 461-464.

[4] K.-F. Lee. 1989. *Automatic Speech Recognition: the development of the SPHINX system,* Kluwer Academic Publishers, Norwell, Mass.

Received: Jan. 31, 2001.
Accepted: Feb. 25, 2001.

▲ Sung-Joon Park
Multimedia Technology Laboratory, Korea Telecom
17 Woomyeon-dong, Seocho-gu, Seoul 137-792, Korea
Tel: +82-2-526-6771
Fax: +82-2-526-5909
E-mail: sjpak@kt.co.kr

▲ Ho-Kyoung Kim
Multimedia Technology Laboratory, Korea Telecom
17 Woomyeon-dong, Seocho-gu, Seoul 137-792, Korea
Tel: +82-2-526-6774
Fax: +82-2-526-5909
E-mail: hokyoug@kt.co.kr

▲ Myoung-Wan Koo
Multimedia Technology Laboratory, Korea Telecom
17 Woomyeon-dong, Seocho-gu, Seoul 137-792, Korea
Tel: +82-2-526-5090
Fax: +82-2-526-5909
E-mail: mwkoo@kt.co.kr