A STOCK INFORMATION SYSTEM OVER THE TELEPHONE NETWORK

Myoung-Wan Koo, Il-Hyun Sohn and Sang-Kyu Park

Software Research Laboratory, Korea Telecom Research Laboratories, Korea Telecom
17 Umyon-dong, Seocho-gu, Seoul, 137-792, Korea

ABSTRACT

In this paper, we present a large vocabulary, speaker independent speech recognition system(KT-STOCK) and describe its performance over the telephone network. KT-STOCK is a stock information retrieval system with which we can obtain the current price of a stock by saying a stock name among 710 stock names listed on the Korea stock exchange. The system is an HMM(hidden Markov model)-based isolated speech recognizer which uses phoneme-like unit as a basic unit. Four digital signal processors are used for real time. And we also implement echo cancellation function for recognizing speech spoken over the voice announcement. Currently, we have achieved the recognition rate of 78.4% in the real environment.

1. INTRODUCTION

One of major applications in the area of speech recognition technology is to recognize speech over the telephone network. Recently, many progress has been made in this field. Examples of such applications are services that generate new revenues. People will obtain information from computer system by asking for what they want, not by typing commands at a computer keyboard. The key concept of this technology is ease of use.

Nippon Telegraph and Telephone has combined speaker independent speech recognition and speech synthesis technology in a telephone information system called ANSER(Automatic Answer Network System for Electrical Request). This system has provided information services for the banking industry since its introduction in 1981[1]. Beginning in 1985, AT&T had begun investigating the possibility of using limited-vocabulary, speaker independent speech recognition capabilities to automate a portion of calls currently handled by operator. In 1992 after field trials, AT&T announced that it would begin deploying voice recognition call processing[2]. Bell Northern Research(BNR) began deploying Automated Alternate Billing Services through local telephone companies with Ameritech in 1989[3].

BNR has started stock quotation services since mid-1992. Callers to this system can obtain the current price of a stock simply by speaking the name of the stock[4]. This system employs subword-based speech recognition so that vocabularies of hundred or thousands of words can, in principle, be recognized without having to record each word.

Korea Telecom has also developed a large vocabulary, speaker independent speech recognition system(KT-STOCK), which has been put under experimental field trial since 1994[5][6]. Experimental field trial shows that word spotting is one of indispensable elements to cope with real environment[7]. In order to implement the word spotting in KT-STOCK, KT-STOCK was refined to include the algorithm for the continuous speech recognition which had been developed for speech translation system[8].

In this paper, we present a large vocabulary, speaker independent speech recognition system(KT-STOCK) and describe its experimental result over the telephone network. In section 2, the overview of KT-STOCK is presented. In section 3, we introduce speech recognition algorithm including

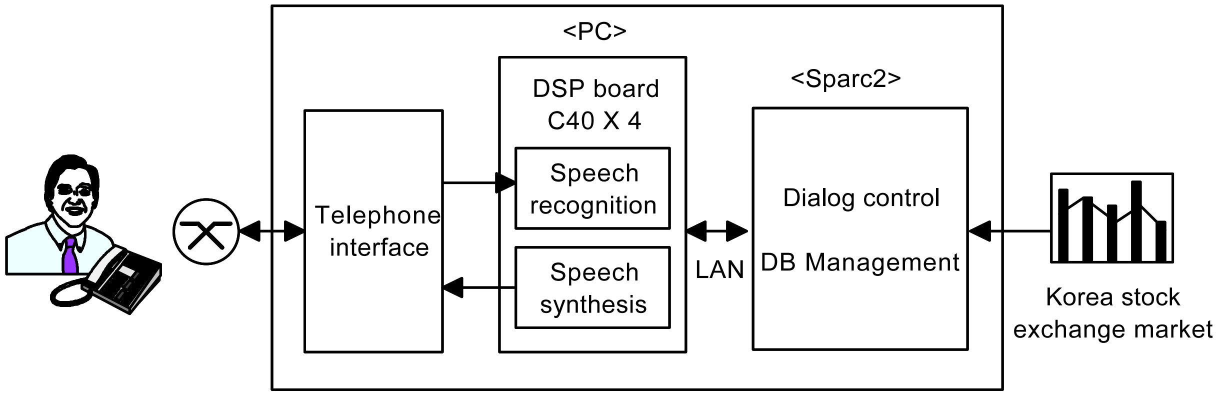


Figure 1. The overview of KT-STOCK

feature extraction, phonetic model and language model. And the experimental result is explained in section 4. The speech database obtained in the real environment is analyzed and the performance of recognizer is also evaluated. Finally, conclusion is made in section 5.

2. SYSTEM OVERVIEW

KT-STOCK is outlined in Figure 1[5]. KT-STOCK is composed of telephone interface, recognizer and database management. Telephone interface makes KT-STOCK accessed from any telephone by dialing the telephone number. Recognizer is implemented in real time on Texas Instruments' TMS320C40 parallel digital signal processors(DSPs) in a IBM-PC. Figure 2. shows the hardware architecture for speech recognizer. Four DSPs are employed, one for feature extraction and vector quantization, three for Viterbi search. One DSP includes an analog daughter module for A/D and D/A conversion through the telephone line and performs endpoint detection, feature extraction and vector quantization frame-synchronously. Three DSPs are connected together using the C40 communication port for parallel running. We also implement echo cancellation function for recognizing speech spoken over the voice announcement. The least-mean-square(LMS) algorithm is used in realization of echo cancellation[9]. Figure 3. shows the effect of echo cancellation. Database manager runs on a workstation, which manages the current stock information from the computer in Korea stock exchange. We use the leased line to connect the workstation to the computer in Korea stock exchange. Since this application requires frequent updating of their vocabulary as the new companies are listed on Korea stock exchange market, subword-based speech recognition is applied so that system manager may simply type the names in Korean. We have also developed a Korean pronunciation generator which automatically converts Korean words into sequences of phonemic transcriptions. The pronunciation generator has the user interface based on X-window, which enables a novice manager to insert and delete the stock name easily.

3. SPEECH RECOGNITION

3.1. Feature extraction

The speech is sampled at 8 kHz over the telephone, and is pre-emphasized with a filter whose transfer function is $1 - 0.95z^{-1}$. The sampled speech is then segmented into frames. Each frame spans 20 msec and is overlapped by 10 msec. LPC(linear predictive coding) analysis is performed with order 14 using the autocorrelation method and a set of LPC driven cepstral coefficients is computed from the LPC coefficients. The LPC driven cepstral coefficient is weighted by a window $W_c(m)$, of the form [10]

$$W_c(m) = 1 + \frac{Q}{2}\sin(\frac{\pi m}{Q}) , \quad 1 \leq m \leq Q$$

In addition to the weighted LPC cepstral coefficients, we also compute their differ-

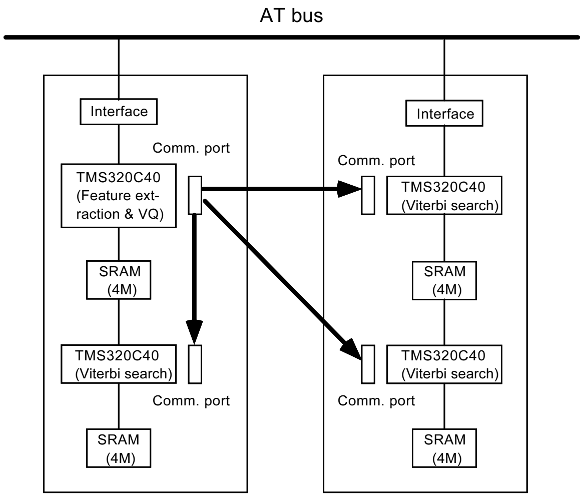
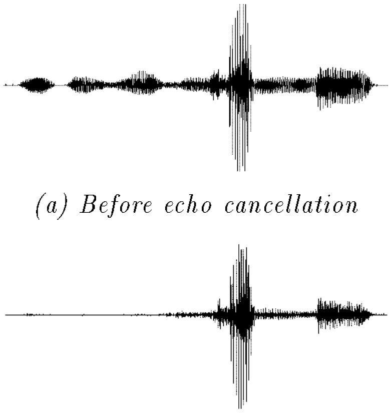


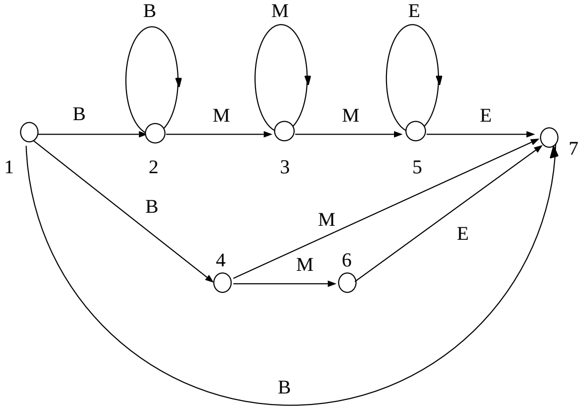
Figure 2. The hardware architecture of speech recognizer



(a) Before echo cancellation

(b) After echo cancellation

Figure 3. The effect of echo cancellation



B,M,E : output pdf's

Figure 4. Topology of phone model

ences, second order differences, differenced log power and second order differenced log power for each frame. These coefficients are then vector quantized. We use four VQ codebooks, three with 256 codewords and one with 64 codewords, using (1) 12 weighted LPC cepstral coefficients, (2) their differences, (3) their second order differences, (4) differenced log power and its second order difference. Our VQ algorithm is based on the Linde-Buzo-Gray (LBG) algorithm.

3.2. Phonetic model

It is necessary to choose basic units for the speech recognition system based on HMM. We choose phoneme-like phones as basic units. We start from 61 context-independent phone models. Each phone is modeled to be independent of context. The topology of our model is shown in Figure 4, which is similar to that of Lee's[11].

This model has 7 states and 12 transitions. The transitions are tied into three groups. Transitions in the same group share the same output probabilities. We also expand our context-independent phone models to context-dependent phone models. We choose 300 context-dependent phone models after considering the size of memory available in each DSP. The unit reduction rule is used for generating a statistically reliable model[12].

3.3. Language model

The continuous speech recognition algorithm has been modified to be used in the isolated word recognition system having the capability of word spotting. Since our continuous speech recognition system uses the bigram model based on the statistical first-order class grammar, the change of parameters for bigram is enough for implementing the word spotting for the isolated word recognition system.

Figure 5. shows the language model for the isolated word recognition system. Only node 1 is set to be the terminal node. We use the silence model to match any background noise that is contained within the endpointed ut-

Table 1. Characteristics of evaluation database

	Training		Test	
Ages	Male (Talkers)	Female (Talkers)	Male (Talkers)	Female (Talkers)
10	18	8	2	2
20	63	17	7	3
30	57	18	7	2
40	0	18	7	2
50	18	10	2	0
Total	227 Talkers		27 Talkers	
	20,700 Tokens		2,457 Tokens	

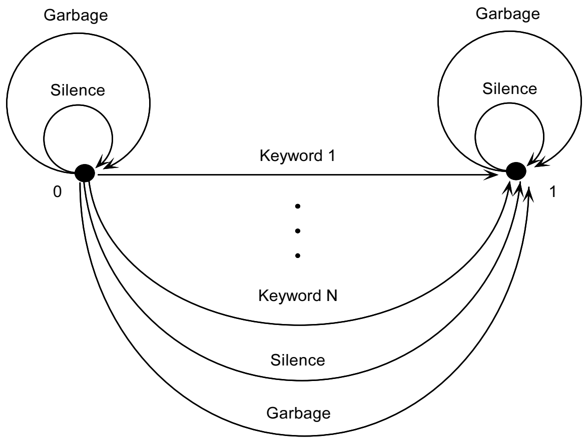


Figure 5. HMM network

terance. And the garbage model is used to model extraneous speech that may occur on either side of a valid keyword.

4. EXPERIMENTAL RESULT

In order to obtain the speech database in the real environment, we use the data collection facilities we have developed[6]. A total of 254 people participated in gathering the speech database. They were willing to use KT-STOCK. All the speech were recorded and edited for performance evaluation. Their ages range from 10's to 50's. 227 of the 254 talkers were designates as the training set talkers and the remaining 27 were designated as test set talkers. Considering sexes and ages, we divided the corpora into two groups as shown in Table 1.

For training, we used an iterative procedure, Baum-Welch algorithm. And Viterbi beam search algorithm was used in the stage of recognition. Our criterion of beam search was based on the threshold.

Table 2 shows the recognition result when 300 context-dependent phone models are used in the real environment.

Table 2. The recognition result in the real environment

Top 1	Top 2
78.4(%)	89.3(%)

We can obtain the recognition rate of 78.4 % for the first candidate (Top 1), and 89.3 % for the first and second candidates (Top 2) in the simulation environment.

Table 3. The comparative performance evaluation
with regard to various kinds of telephone handsets

Telephone handsets	The number of tokens	Recognition rate	
		Top1	Top2
Ordinary telephone	963	85.3	94.5
Wireless telephone	552	77.7	88.0
Public telephone	529	82.4	91.3
Speaker phone	413	58.1	76.0

Table 3 shows the comparative performance evaluation with regards to various kinds of telephone handsets.

We can get the recognition rate of 85.3 % when we use the ordinary telephone and the worst recognition rate of 58.1 % when we use the speaker phone.

5. CONCLUSION

In this paper, we described KT-STOCK which was a large vocabulary, speaker independent speech recognition system and introduced the experimental result in real environment. The KT-STOCK is an HMM-based isolated speech recognizer which can recognize 710 word vocabulary over the telephone. 300 context-dependent phone models of Korean speech were used as basic recognition units. We can obtain the current price of a stock from this system by just saying a stock name. We use four digital signal processors for real time. And we also implement echo cancellation function for recognizing speech spoken over the voice announcement.

The performance has been evaluated in the real environment. we could obtain the recognition rate of 78.4 %. The comparative performance analysis with regard to various kinds of telephone handsets has been done. We could obtain the best recognition rate of 85.3 % with the ordinary telephone and the worst recognition rate of 58.1 % with the speaker phone.

6. ACKNOWLEDGEMENT

The authors are grateful to all members in the spoken language research team, software research laboratory for their continuous support.

REFERENCES

[1] R. Nakatsu, "Anser: an application of speech technology to the Japanese banking industry," *IEEE computer,* Vol. 23, No. 8, pp. 43-48, Aug. 1990

[2] D. B. Roe and J. G. Wilpon, "Whither speech recognition: the next 25 years," *IEEE computer,* Vol. 31, No. 11, pp. 54-62, Nov. 1993

[3] M. Lennig, "Putting speech recognition to work in the telephone network," *IEEE computer,* Vol. 23, No. 8, pp. 35-41, Aug. 1990

[4] M. Lennig et al., "Flexible vocabulary recognition of speech," in *Proc. 1992 Int. Conf. on Spoken Lang. Processing,* pp. 93-96, Oct. 1992

[5] M.-W. Koo et al., "KT-STOCK:-A speaker-independent, large-vocabulary speech recognition system over the telephone, " in *Proc. 1994 Int. Conf. on Spoken Lang. Processing,* pp. 1387-1390, Sep., 1994.

[6] M.-W. Koo et al., "An experimental field trial of a large vocabulary, speaker independent recognition system, " in *Proc. Second IEEE Workshop on Interactive Voice Technology for telecomm. Applications,* pp. 33-36, Sep., 1994

[7] D. J. Krasinski et al., "Automatic speech recognition for network call routing," in *Proc. Second IEEE Workshop on Interactive Voice Technology for telecomm. Applications,* pp. 157-160, Sep., 1994

[8] M.-W. Koo et al., "KT-STS: a speech translation for hotel reservation and a continuous speech recognition system for speech translation," To be appeared in *Proc. 4-th European Conf. on Speech and Comm. and Technology,* Sep., 1995

[9] B. Widrow and S. D. Stearns,*Adaptive signal processing.* Prentice-Hall, Inc. Englewood Cliffs, N.J., 1985.

[10] C. H. Lee et al., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language,* No. 4, pp. 127-165, 1990

[11] K. -F. Lee, *Automatic speech recognition: the development of the SPHINX system.* Kluwer Academic Publisher, Norwell, Mass., 1989.

[12] C. H. Lee et al., "Acoustic modeling of subword units for speech recognition," in *Proc. 1990 IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 721-724, April 1990.