THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.25, NO.1E 2006. 3 pp. 20-24

Implementation of HMM Based Speech Recognizer with Medium Vocabulary Size Using TMS320C6201 DSP

Sungyun Jung*, Jongmok Son**, Keunsung Bae***
* Telecom Examination Div., The Korean Intellectual Property Office.
** Application Technology Research Department, National Security Research Institute.
*** School of Electronic and Electrical Engineering, Kyungpook National University.
(Received February 28 2006; accepted March 27 2006)

Abstract

In this paper, we focused on the real time implementation of a speech recognition system with medium size of vocabulary considering its application to a mobile phone. First, we developed the PC based variable vocabulary word recognizer having the size of program memory and total acoustic models as small as possible. To reduce the memory size of acoustic models, linear discriminant analysis and phonetic tied mixture were applied in the feature selection process and training HMMs, respectively. In addition, state based Gaussian selection method with the real time cepstral normalization was used for reduction of computational load and robust recognition. Then, we verified the real-time operation of the implemented recognition system on the TMS320C6201 EVM board. The implemented recognition system uses memory size of about 610 kbytes including both program memory and data memory. The recognition rate was 95.86% for ETRI 445DB, and 96.4%, 97.92%, 87.04% for three kinds of name databases collected through the mobile phones.

Keywords: *HMM, Speech Recognizer, TMS320C6711, Cepstral Normalization*

I. Introduction

When speech recognition technique is used to a mobile terminal having limited size of memory, it is important to reduce total memory size by minimizing program and database (DB) sizes. To reduce the total memory size of the recognizer without severe degradation of recognition performance, linear discriminant analysis (LDA) method is generally used to reduce the order of features[1].

Basic unit of speech recognition is an important factor in reducing the size of parameters of a recognition system. It may consist of word-based unit or phoneme-liked unit (PLU). It is efficient for a mobile terminal to use phoneme-liked units because it needs small size of database and is easy to increase the size of word list to be recognized. By using the phonetic tied mixture (PTM) or other parameter tying techniques, however, it is necessary to consider the coarticulation effect.

In this paper, we implemented a real time speech recognizer based on HMM of PLU using continuous mixture Gaussian densities. We defined 46 Korean PLUs, and used the PTM model to reduce the size of HMM DB and to consider the coarticulation effect. To reduce the computational load in HMM decoding procedure, we also used tree structured search method and state-based Gaussian selection (SBGS) algorithm. After examining the operation of the PC-based system, we verified the real-time implementation of the recognizer by optimizing and porting the program into the TMS320C6201 EVM board. According to the

Corresponding author: Keunsung Bae (ksbae@ee.knu.ac.kr)
School of Electronic and Electrical Engineering, Kyungpook National University, Buk-gu, Daegu

profiling results for the implemented system on the TMS320C6201 EVM board with 160MHZ clock, we found that the total number of cycles to recognize the speech of 39 frames length (1 frame=32ms) was 139,098,387 cycles, which corresponds to about 70% of TMS320C6201 EVM capability and guarantees the real-time operation.

The paper is organized into four sections. In section 2, implemented PC-based speech recognizer is explained with description of recognition techniques applied to it. In section 3, implemented system on the TMS320C6201 EVM is described in the viewpoint of memory size and processing time with results of recognition experiment. And a conclusion is given in section 4.

II. Speech Recognizer with Medium Size of Variable Vocabulary

We used the mel-frequency cepstral coefficients (MFCC) as a feature parameter, LDA method to reduce the order of feature parameters, and a PTM model to include coarticulation effects. And we also used SBGS to reduce the computational time by selecting the subset of Gaussian component that should be computed given a particular input vector.

2.1. Signal Processing

Input speech signal has sampling rate of 8kHz with 16bits resolution. Analysis condition of the speech is shown in Table 1. A frame size of 32ms Hamming window with 16ms overlap, and pre-emphasis factor of 0.97 were used to calculate 19 mel-scale filter bank outputs. Then 19 triangular filters were distributed with uniform interval on a mel-frequency scale, and 12 MFCCs were computed.

In addition, we also included the first time derivative (delta MFCC), second time derivative (delta-delta MFCC), first time derivative of energy (delta energy), and its second derivate (delta-delta energy), in the feature vector, which makes a 38-dimensional feature vector.

There is some difference between speech data in real mobile environment and training data collected in clean environment. This channel difference is one of the main causes to degrade the recognition performance. Thus cepstral mean normalization (CMN) is generally used to compensate for an acoustic mismatch between them. In this paper, for real-time operation we used the real time cepstral normalization (RTCN) as given in eq.1.

$$\bar{x}_t = \alpha x_t - (1-\alpha)\bar{x}_{t-1} \quad (1)$$

where $\bar{x}_t$ is the recursively estimated mean of cepstrum for t th input vector and x_t is the cepstrum mean of t th input vector itself, and α is the forgetting factor (α=0.125).

2.2. Speech Recognition Engine

Speaker independent continuous HMM is used for each PLU HMM. The topology of a HMM is a 3-state left-to-right Bakis model as shown in Fig 1. Omission of state is permitted to represent short length of basic phonemes. By eliminating phonemes with low occurrence, the total number of Korean PLUs was set to be 46 as given in Table 2. Each PLU HMM in the speech recognizer was trained with 445DB made by Electronics and Telecommunications Research Institute (ETRI) in Korea. Parameters of speech recognizer were estimated using segmental K-Means algorithm and Baum-Welch re-estimation. [2-3]

Table 1. Analysis condition of speech signal.

Pre-emphasis factor	0.97
Analysis window	Hamming
Analysis window size	32 ms (256 samples)
Frame rate	16 ms (128 samples)
Feature parameters	Delta Energy(1) Delta-delta Energy(1) MFCC(12) Delta MFCC(12) Delta-delta MFCC(12)

From the Gaussian distributions for each PLU, the PTM is synthesized for context-independent phoneme models with mixture weights depending on the shared states of triphones. The procedure for building a PTM model is as follows. First, train monophone HMMs that have a large number of Gaussian mixtures for each state. And train shared-state triphone HMMs using the same phone set. Next, assign the mixtures of monophone HMMs to the corresponding states of the triphone HMMs. Then, re-estimate the mixture components and assign the

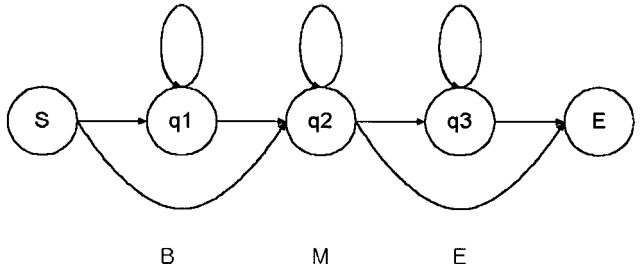

Fig. 1. The topology of a state transition model for a phoneme-liked unit HMM.

Table 2. Phoneme liked units.

b ㅂ 0	n ㄴ 15	ye ㅖ 30	sil 45
b' ㅃ 1	h ㅎ 16	wa ㅘ 31	
p ㅍ 2	r/l ㄹ 17	Weo ㅝ 32	
d ㄷ 3	a ㅏ 18	wi ㅟ 33	
d' ㄸ 4	eo ㅓ 19	wae ㅙ 34	
t ㅌ 5	o ㅗ 20	we ㅞ 35	
g ㄱ 6	u ㅜ 21	oe ㅚ 36	
g' ㄲ 7	eu ㅡ 22	eui ㅢ 37	
k ㅋ 8	i ㅣ 23	Ge ㄱ(end) 38	
j ㅈ 9	ae ㅐ 24	Ne ㄴ(end)39	
j' ㅉ 10	e ㅔ 25	De ㄷ(end)40	
Ch ㅊ 11	ya ㅑ 26	Le ㄹ(end) 41	
s ㅅ 12	yeo ㅕ 27	me ㅁ(end)42	
s' ㅆ 13	yo ㅛ 28	Be ㅂ(end)43	
m ㅁ 14	yu ㅠ 29	Ng ㅇ(end)44	

Table 3. Class classification for adapting PTM.

PLU Class	Default	1st state	3rd state
Plosive 0	b, b' , p, d, d' , t, g, g' , k	-	-
Fricative 1	s, s'	-	-
Affricative 2	j, j' , ch	-	-
Nasal 3	m, n, ng, ne, me, nge	-	-
Whisper 4	H	-	-
Liquid 5	r/l, le		-
Front vowel 6	i, e, ae	ye, wi, wae, we, oe, eui	ya, yeo, yo, yu, ye
mid vowel 7	A	ya, wa	
Back vowel 8	eu, u, o, eo	yeo, yo, yu, weo	wa, weo, wi, wae, we, oe, eui
end 9	ge, de, be	-	-
etc 10	Sil	-	-

weights to discriminate triphones. The mixture itself is also re-trained. [4]. We used the PTM to keep small size of DB and to include coarticulation effect, and class classification table for adapting PTM is shown in Table 3.

The LDA aims at improving discrimination between classes in a vector space, by finding a linear transformation matrix from a M-dimensional vector space to a N-dimensional vector space (M<N). A dimensionality reduction of the vector space can optionally be performed[1]. We reduced the order of feature parameters from 38 to 16 by using the LAD, which resulted in the reduction of computational burden. The SBGS was also chosen because it has less computational load than standard GS and yields a high performance. SBGS limits the number of Gaussian components associated with a state, i.e., the number of Gaussian functions that mapped to one codeword. Much Gaussian components are assigned to the states near the center of the cluster to model them accurately. We used multi-ring method, which assigns more Gaussian components to the states that are often appeared in the training data for that codeword, and assigns fewer to those less appeared[5].

III. Experiment and Discussion

3.1. Recognition experimental results

Recognition experiments were carried out for two kinds of recognizers, a baseline recognition system and a speech recognition system that employed the LDA, PTM, and SBGS. Baseline recognition system uses a 38-dimensional feature vector as an input vector. Conditions described in section 2 except LDA, PTM and SBGS were applied for training of baseline system. For evaluation of both systems we used ETRI 445DB test data (6 males, 4 females) and three name DBs collected from different model of mobile phones (HHP SCH-600, HHP SPH-7000, HF SPH-7000) by Samsung Ltd. The 445DB is recorded in clean environment. But three name DBs are recorded in general mobile terminal environment. Among them, the data denoted by HF SPH-7000 is recorded in hands-free environment

Recognition results of two speech recognition systems evaluated on the 445DB according to the number of Gaussian mixture are shown in Table 4. When the number of Gaussian mixture is set to be 16, recognition results for three name DBs are shown in Table 5. In Table 4, we can see that as the number of Gaussian components representing observation distribution of a state decreases recognition performance degrades a little. But in case of applying SBGS to the recognizer, we can see that the degradation of recognition performance is not so much even though only 20% of computational time of total Gaussian is needed for calculating probabilities in the HMM decoder. Thus we chose the number of Gaussian mixture as 16, which is thought to be reasonable for trade-off between recognition performance and number of Gaussian mixtures. Though speech recognizer is tested on the real mobile DB, as shown in Table 5, we can get good performance except the case of noisy hands-free data, HF SPH-7000.

3.2. Implemented recognizer on the TMS320C6201 EVM

For implementing a speech recognizer to a fixed-point DSP, TMS320C6201, PC-based speech recognition program with a floating-point operation should be transformed into program with a fixed-point operation. In general, Q-format is used in transforming a float-point program into a fixed-point one. Then we transformed all the data used in recognition program into 32-bit, Q-15 format. Next, using intrinsic function supported in TMS320C62xx compiler and loop unrolling method and so on, we optimized the recognition program as efficiently as possible.

Speech recognizer implemented on the TMS320C6201 EVM board handles the input speech signal on the frame basis. It can perform features extraction and decoding processing in a frame of 32ms. Total memory size of 46 acoustic models reduced from 501 kbytes of a baseline system to 370 kbytes with very little degradation in performance. Memory usage for the implemented system on the TMS320C6201 EVM board is shown in Table 6. Total memory size of both program and data is about 610kbytes.

Table 4. Comparison of recognition performance according to the number of mixture.

Mixture / Recognizer	32	16	8
Baseline system (%)	97.0903	96.4475	95.3536
Implemented system (%)	96.8648	95.8611	93.3574

Table 5. Comparison of recognition rate according to name DB.

	HHP SCH-600	HHP SPH-7000	HF SPH-7000
Recognition rate (%)	96.4	97.92	87.04

Table 6. Memory size of an implemented system on the TMS320C6201 EVM board.

Program memory	10.72 42.88	[kwords] [kbytes]
Data ROM	105.09 420.39	[kwords] [kbytes]
Data RAM (Stack is included)	36.56 146.26	[kwords] [kbytes]

To examine the execution time of the implemented system, TI's TMS320C6x profiler in code composer studio (CCS) was used [6]. According to the profiling result, it used 139,090,387 cycles to recognize a word of 39 frame length. So, average cycle required for recognizing a frame is 3,566,625 cycles that correspond to about 70% of CPU performance of 160MHz clock.

IV. Conclusion

In this paper, we implemented a speech recognition system with medium size of vocabulary considering its application to a mobile phone. First, we developed the PC-based variable vocabulary word recognizer having size of program memory and database of total acoustic models as small as possible. To reduce the memory size of acoustic models, linear discriminant analysis and phonetic tied mixture were adopted to the speech recognizer. In addition, state-based Gaussian selection as well as real time cepstral normalization was used for reduction of computational load and robust recognition.

The implemented system verified real-time operation, and has the memory size of about 610 kbytes including both program memory and data memory.

Acknowledgment

This work was partially supported by grant No. R01-2003-000-10242-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

1. R. Haeb-Umbach, H.Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," International Conference on Acoustic, Speech and Signal Processing, **1**, 13-16, 1992.
2. B.H. Juang, L.R. Rabiner, "The Segmental K-Means Algorithm for Estimation Parameters of Hidden Markov Models," IEEE Trans. on Acoustics, Speech, and Signal Processing, **38** (9) 1639-1641, 1990.
3. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," IEEE Trans. on Acoustics, Speech, and Signal Processing, **1**, No. 3, 345-361, 1993.
4. Akinobu Lee, Tatsuja kawahara, Kiyoshiro Shikano, "A New Phonetic TIED-MIXTURE MODEL For Efficient Decoding," International Conference on Acoustic, Speech and Signal Processing, **3** (2) 1269-1271, 2000.
5. Mark J.F. Gales, Katherine M. Knill "State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMM's," IEEE Trans. on Acoustics, Speech, and Signal Processing, **7** (2) 52-161, 1999.
6. Texas Instrument, TMS320C6000 Programmer's Guide, 2000
7. Markus Lieb, Reinhold Haeb-Umbach, "LDA derived Cepstral Trajectory Filters in Adverse Environmental Conditions," International Conference on Acoustic, Speech and Signal Processing, 2000.
8. Fu-Hua Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero, "Efficient Cepstral Normalization for Robust Speech Recognition,"

Proc. of the Sixth ARPA Workshop on Human Language Technology, 1993.
9. Alejandro Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, (Ph.D. thesis, Carnegie Mellon University, 1990)

[Profile]

•Sungyun Jung

Sungyun Jung received his B.S. and Ph.D. degrees in electrical engineering from Kyungpook National University in 1991 and 2005, respectively. He joined Korean Intellectual Property Office in 2005, where he is working at Telecom Examination Division.

•Jongmok Son

Jongmok Son received his B.S., M.S., and Ph.D. degrees in electrical engineering from Kyungpook National University in 1997, 1999, and 2005, respectively. Since March 2005, he has been with Department of Application Technology Research in National Security Research Institute.

•Keunsung Bae

Keunsung Bae received his B.S., M.S., and Ph.D. degrees in electrical engineering from Seoul National University in 1977, KAIST in 1979, University of Florida in 1989, respectively. He has been working at the School of Electrical Engineering and Computer Science in Kyungpook National University since 1979. His research interests include speech signal processing, adaptive filtering, noise removal, and wavelet transform based signal processing.