Implementation of HMM-Based Speech Recognizer Using TMS320C6711 DSP

Hyojoon Bae*, Sungyun Jung**, Jongmok Son**, Hongseok Kwon**, Siho Kim**, Keunsung Bae**
* Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea
** School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Korea
Tel : +81-53-950-5527 Fax : +81-53-950-5505 E-mail: si5@mir.knu.ac.kr, ksbae@mir.knu.ac.kr

Abstract: This paper focuses on the DSP implementation of an HMM-based speech recognizer that can handle several hundred words of vocabulary size as well as speaker independency. First, we develop an HMM-based speech recognition system on the PC that operates on the frame basis with parallel processing of feature extraction and Viterbi decoding to make the processing delay as small as possible. Many techniques such as linear discriminant analysis, state-based Gaussian selection, and phonetic tied mixture model are employed for reduction of computational burden and memory size. The system is then properly optimized and compiled on the TMS320C6711 DSP for real-time operation. The implemented system uses 486kbytes of memory for data and acoustic models, and 24.5kbytes for program code. Maximum required time of 29.2ms for processing a frame of 32ms of speech validates real-time operation of the implemented system.

Real-time implementation, HMM-based speech recognizer, TMS320C6711

1. INTRODUCTION

Nowadays speech interface is becoming increasingly important as the size of hand-held terminals are getting smaller and smaller. For the application of speech recognition technologies to hand-held terminals such as mobile phones and PDA systems, it is important to reduce both computational burden and required memory size in a speech recognition system.

Basic unit of speech recognition is an important factor determining memory size of data for acoustic models and recognition performance. It may consist of word-based unit or phoneme-liked unit (PLU). The PLU could be efficient for hand-held terminals since it needs small size of database and is easy to increase the size of word list to be recognized. However it is necessary to consider the coarticulation effect using a phonetic tied mixture (PTM) or other parameter tying methods [1]. Linear discriminant analysis (LDA) can be used to reduce the dimension of the feature parameters without severe degradation of recognition performance [2,3].

In this paper, we implemented an isolated word speech recognizer based on HMMs of PLU with continuous mixture Gaussian densities. We defined 46 Korean PLUs, and used the PTM model to reduce the data size of HMM acoustic models while considering the coarticulation effect. To reduce the computational burden in HMM decoding process, we employed both tree structured search method and state-based Gaussian selection (SBGS) algorithm [4]. 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 TMS320C6711 DSP Starter Kit (DSK). According to the profiling results for the implemented system on the TMS320C6711 DSK that runs with 150MHZ clock, the maximum number of cycles required for feature extraction and decoding process for a frame length of 32ms speech signal was 4,380,194 cycles, which corresponds to 29.2ms and guarantees the real-time operation.

The rest of this paper is organized as follows. In section 2, an implemented PC-based speech recognizer is explained with description of recognition techniques applied to it. In section 3, the implemented system on the TMS320C6711 DSK is described in the viewpoint of used memory size and processing time with results of recognition experiment. And finally conclusion is given in section 4.

2. HMM-BASED SPEECH RECOGNITION SYSTEM

2.1. Feature extraction

The input speech signal has sampling rate of 8kHz with 16bits resolution. Analysis conditions for the input speech are given in Table 1. The input speech is preemphasized with a factor of 0.97. The analysis frame size is set to be 32ms, and is Hamming windowed with frame rate of 16ms. After calculating 256-point FFT, 19 mel-scale filter bank outputs are obtained using triangular filters distributed with uniform interval on the mel-frequency scale. From the filter bank outputs, then 12 mel-frequency cepstral coefficients (MFCCs) are computed as basic feature parameters. In addition, we also include the first derivative and second derivative of MFCC and energy in the feature vector, which makes totally a 38-dimensional feature vector.

Table 1. Analysis conditions for the input speech signal

Pre-emphasis factor	0.97
Analysis window	Hamming
Analysis frame size	32ms (256 samples)
Frame rate	16ms (128 samples)
Feature parameters(38)	Delta Energy(1) Delta-delta Energy(1) MFCC(12) Delta MFCC(12) Delta-delta MFCC(12)

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 ($M \geq N$). At thesame time a dimensionality reduction of the vector space can optionally be performed [2]. We reduce the order of feature parameters from 38 to 16 by using the LDA, which results in the reduction of computational burden.

Since the spectral characteristics show much difference between training speech data obtained in quiet environment such as ETRI 445DB and test speech data obtained in real environment such as mobile phones, to get robustness it is necessary to compensate for the difference in some way. This channel difference is one of the main causes to degrade the recognition performance. Cepstral mean normalization (CMN) is generally used to compensate for the acoustic mismatch between them. In this paper, however, for real-time operation we use the real time cepstral normalization (RTCN) method given by the eq.(1).

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

where $\bar{x}_t$ is the recursively estimated mean of cepstrum for t th input vector and x_{t-1} is the cepstrum mean of $t-1$th input vector itself, and α is a forgetting factor set as 0.125, empirically.

2.2. Speech recognition system

The continuous HMM is used to represent an acoustic model for each PLU. The topology of an HMM is a 3-state left-to-right Bakis model as shown in Figure 1. Omission of state is allowed to take the short length of basic phonemes into consideration. By removing some indistinguishable Korean phonemes whose frequency of occurrence is low, the total number of PLUs was set to be 46 as given in Table 2. The HMM for each PLU in the speech recognizer was trained with 445DB made by Electronics and Telecommunications Research Institute (ETRI) in Korea. HMM parameters were estimated using segmental K-means algorithm and Baum-Welch reestimation [5,6].

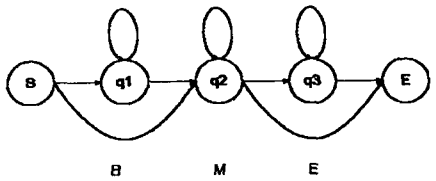

Figure 1. The topology of a state transition model for a phoneme-liked unit HMM

From the Gaussian distributions for each PLU, the PTM is synthesized to get context-dependent phoneme models with different mixture weights depending on class. The procedure for building a PTM model is as follows. First, monophone HMMs that have a large number of Gaussian mixtures for each state are obtained using the training database. Then, PTM HMM is synthesized from monophone HMM and class information. The class classification table for adapting PTM is shown in Table 3.

Table 2. Korean 46 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	Nge 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	-	-

The SBGS limits the number of Gaussian components associated with a state, i.e., the number of Gaussian functions that mapped to one codeword. It is adopted in the recognition system because it can reduce computational burden dramatically compared to standard GS while maintaining good recognition performance. In SBGS many Gaussian components are assigned to the states near the center of the cluster to model them accurately. We use a multi-ring method that assigns more Gaussian components to the states that are often appeared in the training data and assigns fewer to those less appeared [4].

3. EXPERIMENTS AND DISCUSSION

3.1. Recognition experimental results

Recognition experiments were carried out for two sets of test data. Both data sets were collected by Samsung Electronics corporation in real environment like at the office, roadside, subway, bus terminal using different model of mobile phones such as HHP SCH-600 and HHP SPH-7000. The test data sets contain 20 different words uttered by male and female speakers with dialects. The signal was sampled at 8kHz with 13bits resolution. The test data DB is described in Table 4.

In the recognition system the dimension of an input vector as feature parameters is 16, and the number of Gaussian mixture representing the observation distribution of a state in the HMM is set to 16 after considering recognition performance and required memory size for acoustic data models as well as computational burden. Table 5 shows the recognition results for test data sets depending on the number of tokens used in Viterbi decoding process. It is shown that as we decrease the number of tokens from 100 to 60 in decoding, the recognition performance degrades very slightly from 96.50% to 96.05%. But we can reduce the computational burden as shown in Table 6 at the cost of a little performance degradation.

3.2. Implemented recognizer on the TMS320C6711 DSK

TMS320C6711 is a powerful floating-point DSP from the Texas Instrument. It operates at 150MHz clock speed, and has 64kbytes of internal memory. TMS320C6711 DSK provides a low-cost, and convenient hardware development environment. It contains the 'C6711 processor and provides 16-bit A/D converter, 16Mbytes of SDRAM as an external memory, and parallel port cable, power supply, etc. For implementing an HMM-based speech recognizer on the TMS320C6711 DSK, PC-based speech recognition program is first optimized and compiled using appropriate compile options. Especially fft function in the source code is replaced with computationally efficient dsp support library function.

To investigate the execution time of the implemented system, TI's TMS320C6x profiler in code composer studio (CCS) was used [7]. Table 6 shows the profiling results for processing one frame, i.e., 32ms length of speech with the number of token 60 and 100 in the decoding process. We can see that real-time operation cannot be achieved with the number of token as 100 since maximum required time for 32ms of speech is about 44.5ms. But with the number of token as 60, maximum required time reduces to 29.2ms that corresponds to 4,380,194 clock cycles, which results in real-time operation. By adding microphone interface to the implemented system with the number of token as 60, proper real-time operation was confirmed with correct recognition result as expected.

Total memory size for data and 46 acoustic models is about 486 kbytes, and for program code 24.5kbytes. Thus total memory size of both program and data is about 510.5kbytes. Internal and external memory usage for the implemented system on the TMS320C6711 DSK is shown in Table 7.

Table 4. Test data sets for recognition experiments

Phone model	Number of speakers	Number of utterances
HHP SCH-600 (Cellular phone)	3 males 2 females	500
HHP SPH-7000 (PCS phone)	15 males 10 females	2,500

Table 5. Correct recognition rate for test data sets

Number of tokens	HHP SCH-600	HHP SPH-7000	Average
60	95.8%	96.3%	96.05%
100	96.2%	96.8%	96.50%

Table 6. Profiling results for executing one frame of speech

Number of tokens	Number of clocks			Max. required time
	Max.	Min.	Average	
60	4,380,194	2,825,099	3,640,497	29.20ms
100	6,730,899	2,893,097	5,471,819	44.87ms

Table 7. Memory size of an implemented system

Internal memory	55.783 bytes	Program code	24,480 bytes
External memory	454,694 bytes	Data and acoustic models	485,997 bytes
Total	510,477 bytes	Total	510,477 bytes

4. CONCLUSION

We have implemented an HMM-based speech recognizer that can handle several hundred words of vocabulary size as well as speaker independency, considering its application to hand-held terminals. First, we developed a PC-based system that operates on the frame basis with parallel processing of feature extraction and decoding to make the processing delay as small as possible. Many techniques such as LDA, SBGS, and PTM were used for reduction of computational burden and size of memory for program code and database of total acoustic models as small as possible.

With the number of token in the decoding process as 60, average correct recognition rate of 96.05% is achieved for test database collected in real environment. And its proper real-time operation was confirmed by adding microphone interface to the implemented system with memory size of about 510.5kbytes including both acoustic models, data and program code.

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] Akinobu Lee, Tatsuja Kawahara, Kiyoshiro Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," International Conference on Acoustic, Speech and Signal Processing, vol. 3, no. 2, pp. 1269-1271, June, 2000.

[2] R. Haeb-Umbach, H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," International Conference on Acoustic, Speech and Signal Processing, vol. 1, pp.13-16, 1992.

[3] Markus Lieb, R. Haeb-Umbach, "LDA-derived Cepstral Trajectory Filters in Adverse Environmental Conditions," International Conference on Acoustic, Speech and Signal Processing, 2000.

[4] 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*, vol. 7, no. 2, pp. 152-161, Mar., 1999.

[5] 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*, vol. 38, no. 9, pp. 1639-1641, Sep. 1990.

[6] 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*, vol. 1, No. 3, pp. 345-361, July 1993.

[7] Texas Insrtument, *TMS320C6000 Programmer's Guide*, 2000.