DEVELOPMENT OF LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM FOR MONGOLIAN LANGUAGE

Seiichi Nakagawa, Erdenebat Turmunkh, Hiroshi Kibishi, Kengo Ohta, Yasuhisa Fujii, Masatoshi Tsuchiya and Kazumasa Yamamoto

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

ABSTRACT

We developed a large vocabulary continuous speech recognition system(LVCSR) for Mongolian language. It is the first LVCSR system of Khalkha dialect in Mongolia. Firstly, we created Mongolian speech corpus for acoustic model and it contains over 6000 utterances in total recorded from 700 different sentences spoken by 40 male speakers, and then we created monophone and triphone based HMMs. Secondary, phoneme, morphone and word based n-gram language models were prepared by using 6 million words in a text corpus. Finally, we conducted continuous speech recognition experiments and obtained the phoneme correct rates of 56% and 67% by using monophone HMMs and triphone HMMs, respectively. We also obtained the word correct rates of 63% and 68% by using monophone HMMs & word based trigram and triphone HMMs & word based trigram, respectively.

Index Terms— Large vocabulary continuous speech recognition (LVCSR), Mongolian, Khalkha, Morpheme

1. INTRODUCTION

For Mongolian speech recognition, although a large vocabulary continuous speech recognition (LVCSR) system for Chakhar (Inner-Mongolia) dialect in China has been developed [1], LVCSR system for Khalkha Mongolia dialect in Mongolia has not yet been developed. This paper describes the first LVCSR system for the Khalkha Mongolian language, although there are researches for spoken word recognition [2] and domain-specific continuous speech recognition [3]. We compare the performance of Mongolian speech recognition with different acoustic models (monophone v.s. triphone) and different language models (phoneme v.s. morpheme v.s. word) on perplexity, continuous phoneme recognition and continuous word recognition, respectively. We used the standard automatic speech recognition techniques and tools. They are HTK [4], CMU-Cambridge toolkit [5] and SPOJUS++ [6] developed in our laboratory. By using a speech corpus of 15 hours by 40 male speakers and a text corpus consisting of 5 million words, we obtained the phoneme correct rates of 56% and 67% by using monophone HMMs and triphone HMMs, respectively. We also obtained the word correct rates of 63% and 68% by using monophone HMMs & word based trigram and triphone HMMs & word based trigram, respectively.

2. MONGOLIAN LANGUAGE AND PHOMEMIC SYSTEM

2.1 Grammar, Word, Morpheme [7]

Mongolian belongs to the Altaic language family of the Altaic language system like Uyghur[8]. There are several dialects such as Inner-Mongolian in China and Khalkha Mongolian in Mongolia. In this paper, we focus on the latter. The population talking Khalkha dialect is about 2.7 million. The syntactic structure of Mongolian is Subject-Object-Verb (SOV) like Japanese. The written style is divided by a space between words like European language, unlike Japanese. A word is composed of stem + suffix 1 + suffix 2 (.

Mongolian morphemes are classified into word stem and inflectional suffixes. Word stem keeps the original meaning of the word, and usually appears at the beginning. Suffix morphemes have lexical meaning and build a new word. Each suffix represents only a grammatical meaning. Suffixes related to voice, aspect, or mood can be added to verbs in the prescribed order. There are no irregular verbs. Noun stems can be marked for plurality, case, possessiveness, etc. in the prescribed order. Contrary to other agglutinative languages as Turkish, there are no person or number suffixes in Mongolian verbs like Japanese. Figs. 1 and 2 show examples of inflection for noun “child”, and verb “eat”, respectively.

(1) stem： хүүхэд； child

(2) stem + plural： хүүхдүүд； children

(3) stem + plural + postposition： хүүхдүүдэд； to children

(4) stem + plural + postposition+possesive：

хүүхдүүддээ； to one’s children

Fig. 1 An example of inflection for noun “child”

(1) stem： ид； eat

(2) stem + passive： идэгд； be eaten

(3) stem + intent： идүүл； will eat

(4) stem + past： идэв； ate

(5) stem + present perfect： идчихсэн； have eaten

Fig. 2 An example of inflection for verb “eat”

2.2 Phonemic System

In Khalkha dialect in Mongolian, there are 7 short vowels (+ reduced vowel [Ə] , 7 long vowels and 23 consonants as shown in Tables 1 and 2. Besides the above, there are 5 diphthongs and 5 “y–vowels”.

In our recognition system, we used 8 vowels including “Ə”, 22 consonants, and two pauses (short pause and silence). The long vowel and diphthong are expressed by the concatenation of short vowels.

Table 1 Mongolian vowels

short vowels	a, o, u, ü (ue), ö (oe), e,I, Ə
long vowels	a:, o:, u:,ü:(ue:), ö:(oe:), e:, i:
dipthongs	ai, oi, ui,üi(uei), ei
“y” vowels	ya, yo, ye, yu, yü

3. SPEECH CORPUS AND ACOUSTIC MODEL

3.1 Speech Corpus

We collected and recorded the speech corpus of Khalkha dialect in Mongolia. The corpus consists of 4701 spoken sentences and 1427 spoken words uttered by 40 adult males (range from 18 to 35 years old). The amount of speech is about 15 hours. The set of sentences is composed of 700 different sentences, which is a subset of 1500 phoneme balanced sentences [9]. The speech was sampled by the rate of 44.1 kHz and then down-sampled by 16 kHz. From these sampled speech signals, we extracted the feature parameters of MFCC, their delta, delta-delta coefficients and delta, delta-delta power in total 38 dimensions at every 10 ms. The test corpus for speech recognition consists of 200 sentences uttered by two other males, which were selected from the rest of 1500 sentences (i.e., 800 sentences).

3.2 Acoustic Model

We made speaker-independent monophone-based HMMs and triphone-based HMMs by using HTK tools. Each HMM consists of 3 states with diagonal-type Gaussian Mixture Models (GMM). In the case of monophone models, we used 8, 16 and 32 mixtures for GMMs. In the case of triphone models, we tied the acoustic similar states with the same GMMs and generated/learned different 600 states. The mixtures of GMMs were set to 4, 6 and 16. For the state tying, we used a technique based on the decision tree, where the questions for making the tree were referred to English phonemic system [8].

4. TEXT CORPUS AND LANGUAGE MODEL

4.1 Text Corpus

We borrowed the Mongolian text raw corpus of 6 million words collected by the National University of Mongolia in Mongolia, which consisted of daily online or printed newspapers, literature, and laws [9]. In this corpus, there are about distinct 200,000 words. The word in this text corpus is divided into morpheme sequences by using a morpheme analyzer [10] and then transformed to phoneme sequences.

4.2 Language Model

We constructed the statistical n-gram language models of Khalkha dialect Mongolian by using CMU-Cambridge language toolkit. There are phoneme-based, morpheme-based and word-based n-grams. We excluded sentences which were not analyzed correctly by the morpheme analyzer because of including numeral numbers or English words. After cleaning, we obtained 5, 455k morphemes in total and 123k different morphemes, respectively. We calculated morpheme-based n-grams by using the vocabulary size of 20k morphemes, including 88 different

Table 2 Mongolian consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvuler	Glottal
Plosive	p b			t d				k g	G	
Nasal	m			n				ŋ	N	
Trill				r						
Fricative		f v		s z	ʃ dz	ts			χ	h
Approximant							j			
Lateral approximant				l						

suffixes. The coverage rates were 96.6% and 93.8% for the training data and test data, respectively. For the word-based n-grams, we used 3,764k words in total and 144k different words, respectively. By using the vocabulary size of 40k words, the coverage rates were 96.1% and 93.2% for the training data and test data, respectively. From the above statistics, we can estimate that a word is composed of about 1.5 morphemes on the average. Therefore, from a view point of language constraint, a word based bigram LM corresponds to a morpheme based bigram or trigram LM, and a word based trigram LM corresponds to a morpheme based 4 gram LM,
respectively. The perplexity for the test corpus is shown in Table 3. The perplexity by word-base trigram is sufficiently small in spite of Mongolian for the agglutinative language, if a test set is included in the training set. However, the perplexity becomes very large, if a test set of 200 sentences is not included in the training set. This shows that the
training size (about 4 million words) is insufficient. For a such a case, a morpheme-based language model will be suitable .

Table 4 summarizes the hit rate for various language models. For the case of excluding the test set in training data for building a word-based tri-gram language mode, the hit rate of trigram decreased to 24.4%.

Table 3 Perplexity and OOV rate

(a-1) phoneme-based

1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
22.4	14.1	10.5	7.9	6.2	5.0

(a-2) phoneme-based (excluding test set)

1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
22.4	14.1	10.6	7.9	6.1	4.9

(b-1) morpheme-based

n-gram	vocabulary size	perplexity	OOV rate
1-gram	20000	1682.2	6.9
2-gram	20000	174.6	6.9
3-gram	20000	25.6	6.9
4-gram	20000	7.0	6.9

(b-2) morpheme-based (excluding test set)

n-gram	vocabulary size	perplexity	OOV rate
1-gram	20000	1657.3	7.2
2-gram	20000	365.6	7.2
3-gram	20000	288.9	7.2
4-gram	20000	295.6	7.2

(c-1) word-based

n-gram	vocabulary size	perplexity	OOV rate
1-gram	40000	7805.4	6.8
2-gram	40000	296.7	6.8
3-gram	40000	15.0	6.8

(c-2) word-based (excluding test set)

n-gram	vocabulary size	perplexity	OOV rate
1-gram	40000	7522.1	7.9
2-gram	40000	2112.3	7.9
3-gram	40000	2016.3	7.9

Table 4 N-gram hit rate (%)

(a-1) phoneme-based

1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
0.0	0.0	0.0	0.1	0.3	99.6

(a-2) phoneme-based (excluding test set)

1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
0.0	0.0	0.0	0.1	0.4	99.5

(b-1) morpheme-based

1-gram	2-gram	3-gram	4-gram
0.5	0.8	1.6	97.0

(b-2) morpheme-based (excluding test set)

1-gram	2-gram	3-gram	4-gram
15.9	34.6	25.7	23.8

(c-1) word-based

1-gram	2-gram	3-gram
0.8	0.9	98.3

(c-2) word-based (excluding test set)

1-gram	2-gram	3-gram
36.9	38.7	24.4

5. SPEECH RECOGNITION RESULTS

5.1 Continuous phoneme recognition

Firstly, we conducted continuous phoneme recognition by using monophone and triphone based HMMs. For the monophone-based recognition, we used the decoder of SPOJUS++ developed in our laboratory[6], which can decode speech by using left-context dependent HMMs. However, for the triphone-based recognition, we used the decoder of HTK tool (HDcode). The test set consists of 200 utterances by two male adults. Table 3 (a) summarizes the perplexity of phoneme-based n-gram and Table 5 shows the recognition results, where “Cor” and “Acc” are defined as follows:

Cor = 100 – substitution error rate – deletion error rate
Acc = 100 – substitution error rate – insertion rate – deletion error rate

Finally, we obtained the phoneme recognition rates of 66.8%(Cor) and 51.6%(Acc), respectively. We think it is a reasonable performance.

Table.5 Phoneme recognition result (%)
(a) monophone-based

mixtures / N-gram	8		16		32	
	Cor	Acc	Cor	Acc	Cor	Acc
1	49.2	19.1	52.5	21.6	54.9	23.4
2	50.1	28.1	52.8	30.3	55.1	32.6
3	50.8	29.5	53.7	32.1	56.2	34.7
4	50.8	30.3	53.8	32.9	56.4	35.5
5	50.7	30.5	53.8	33.1	56.3	35.9

(b) triphone-based

mixtures / N-gram	4		6		8	
	Cor	Acc	Cor	Acc	Cor	Acc
2	62.1	45.4	63.3	48.0	64.0	49.5
3	64.5	47.1	66.0	49.9	66.8	51.6

5.2 Continuous word recognition

Next, we conducted continuous word recognition, that is, large vocabulary continuous speech recognition (LVCSR). The test set is the same as Section 5.1. We used the decoder of SPOJUS++ for both monophone and triphone models. Table 3 (c) summarized the perplexity of word-based n-gram and Table 6 shows the word recognition results. By using triphone-based HMM and trigram language model, we obtained the word recognition rates of 67.5%(Cor) and 61.9%(Acc), respectively. When we used a word-based tri-gram language model excluding the test set in training (refer to Table 3 (c-2)), however, the rates decreased to 24.4% (Cor) and 21.1%(Acc), respectively.

Table 6 Continuous word recognition result by word-based trigram language model corresponding to Table 3(c-1)
(a) monophone-based HMM (#mixtures=32)

N-gram	Cor	Acc
2	46.6%	37.6%
3	63.2%	59.0%

(b) triphone-based HMM (#mixtures=8)

N-gram	Cor	Acc
2	56.7%	46.8%
3	67.5%	61.9%

6. CONCLUSION

In this paper, we described a Mongolian large vocabulary continuous speech recognizer (LVCSR). It was the first LVCSR system for Khalkha dialect in Mongolia. By training word-based n-gram language model using about 3.8 million words and triphone-based HMMs using speech of 15 hours for 40 adult male speakers, we obtained the correct rate of about 68% based on trigram LM. In future works, we should compare the word-based LM and morpheme-based LM！[11,12,13,14,15,16,17] by using larger text corpus and speech corpus. It may be also necessary to adapt the acoustic models of Mongolian under limited speech corpus from the acoustic models of rich resource language with large speech corpus.

7. ACKNOWLEDGEMENT

The authors would like to thank Professor A. Chagnaa of National University of Mongolia for supporting the text corpus and the collection of speech data.

8. REFERENCES

[1] G. Gao, Biligetu, Nabuqing and S. Zhang, ”A Mongolian speech recognition system based on HMM,” ICIC 2006, Lecture Notes AI 4114, Springer, pp. 667-676, 2006.

[2] A. Altangerel and B. Damdinsuren, “A large vocabulary speech recognition system for Mongolian language,” Proc. Oriental COCOSDA, 2008.

[3] I. Dawa, S. Nakamura, Y. Sagisaka and K. Shirai, “Spoken language corpora for Mongolian family language,” Computer Processing of Asian Spoken Languages, (Ed) S. Itahashi, L. Tseng, Hirakawa Kogyo sha Co., 2010.

[4] S.Y.Young, G. Everman, M.J.F.Gales, T. Hain, D. Kenshaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev and P.C. Woodland, The HTK book version 3.4 Manual, Cambridge, 2006.

[5] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge toolkit,” Proc Eurospeech, pp.2707-2710, 1997.

[6] Y. Fujii, K. Yamamoto and S. Nakagawa, “Large vocabulary speech recognition system: SPOJUS++,” Proc. 11th Wseas Int. Conf. MUSP-11, pp.110-118, 2011.

[7] P.Jaimai, T. Zundui, A. Chagnaa and C-Y. Ock, “PC-KIMMD-based description of Mongolian morphology,” Int. Jour. Information Processing systems, Vol.1, No.1, pp.41-48, 2005.

[8] S. J. Young, J. Odell and C. Woodland, “tree-based state tying for high accuracy acoustic modeling,” Proc. Human Language Technology Workshop, pp.307-312, 1994.

[9] J. Purey and C. Altangered, “Language resources for Mongolian,” Proc. Human Language Technology for Development, pp.56-61, 2011.

[10] S. Enkhbayar, T. Utsuro and M. Sato, “Mongolian morphological analysis based on phonological and morphological constraints,” IPSJ, Natural Language Processing Report, 2004 NL-164(7), 2004.11, (in Japanese).

[11] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara and A. Hamdulla, “Uyghur morpheme-based language models and ASR,” Proc. ICSP, 2010.

[12] O-W. Kwon and J. Park, “Korean large vocabulary continuous speech recognition with morpheme-based recognition units,” Speech Communication, Vol. 39, pp. 287-300, 2003.

[13] T. Hirsimaki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja and J. Pylkkonen, “Unlimited vocabulary speech recognition with morph language models applied to Finnish,” Computer Speech and Language, Vol. 20, pp. 515-541, 2006.

[14] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh and A. Stolcke,” Morphology-based language modeling for conversational Arabic speech recognition,” Computer Speech and Language, Vol. 20, pp. 589-608, 2006.

[15] E. Arisoy, H. Dutagaci and L. M. Arslan, “A unified language model for large vocabulary continuous speech recognition of Turkish,” Signal Processing, Vol. 86, pp. 2844-2862, 2006.

[16] T. Rotovnik, M.S. Maucee and Z. Kacic, “Large vocabulary continuous speech recognition of an inflected language using stems and endings”, Speech Communication, Vol. 49, pp. 437-452, 2007.

[17] P. Mihajlik, T. fegyo, O. Taske and P. Ircing, “A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages – Like Hungarian, Proc. Interspeech, pp.1497-1500, 2007.