ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

Large Vocabulary Continuous Speech Recognition in Uyghur:

Data Preparation and Experimental Results

Nasirjan Tursun[1,2], Wushour Silamu[3]
1 School of Electronics and Information Engineering, Xian Jiaotong University, Xi'an, China
2 College of Mathematics and Systems Science, Xinjiang University, Urumqi, China
3 School of Information Science and Engineering, Xinjiang University, Urumqi, China
{nasir, wushour}@xju.edu.cn

Abstract—**Uyghur language is an agglutinative language. It is one of the least studied languages on speech recognition area. In this work, we present the research process of Uyghur Large Vocabulary Continuous Speech Recognition based on HMM (Hidden Markov Model). This paper introduce the process of data collection (text corpus and speech corpus), the unit selection for speech recognition, the creation of acoustic and language model for Uyghur language. Also presents the experimental results of Uyghur continuous speech recognition in different recognition units.**

Keywords- Uyghur; speech recognition; acoustic model; language model; recognition unit

I. INTRODUCTION

The speech recognition technology has appeared the isolation word recognition, the connection numeral recognition, the connected word recognition, the continuous speech recognition, the dialog system, etc., since 1962. The vocabulary also expands from the few vocabularies to the very large vocabulary [1].

Large vocabulary continuous speech recognition is only at the beginning of development for Uyghur language. In languages like Chinese, English, and Japanese etc., many large vocabulary continuous speech recognition engines have been evaluated on several different tasks, such as IBM ViaVoice, Dragon Naturally Speaking, Julius.

The Uyghur language belongs to the Turkic language family, which is grouped to the Altaic languages system [2]. Uyghur language is an agglutinative language. It is possible to produce a very high number of words from the same root with suffixes, and increase the number of OOV (Out of Vocabulary).

From 1990 to 1994, we had developed the Uyghur isolated-word speech recognition and synthesis system, which include 40,000 words and 1,200 common used syllable of Uyghur language. In 2005, the Key Laboratory of Multilingual Information Technology of Xinjiang University had organized a speech information processing research group, and begins to research Uyghur continuous speech recognition and speech synthesis.

On Jan. 2007 and Aug. 2007, we had established a speech database for Large Vocabulary Continuous Uyghur Speech Recognition. The first part of this speech database were includes 3,000 common used Uyghur sentences selected from Uyghur text corpora and 10 speakers (6 male, 4 female, almost all speaker come from university), each speaker read 500 sentences. The second part of this speech database were includes 1,018 common used Uyghur sentences (these sentences not included in above 3000 sentences) selected from 25,000 sentences and 64 speakers (32 male, 32 female, almost all speaker come from university), each speaker read 100 sentence randomly. In this paper we will mainly introduce the second part of this speech database. The total utterances of the second part of this speech database are 6500. The size of this speech database about 1.5GB, and 10 hours long.

In this paper, we present our work on the research of Uyghur Large Vocabulary Continuous Speech Recognition, and the experimental results of Uyghur speech recognition. In section 2, we briefly introduce the Uyghur language characteristics. In section 3, we describe the data collection (design of text corpus and speech database). In section 4, we describe the unit selection for Uyghur speech recognition. In section 5, we describe the creation of acoustic and language model for Uyghur language. At last, we present the Uyghur speech recognition experiments and the results.

II. UYGHUR LANGUAGE CHARACTERISTICS

Uyghur language written with the Arabic alphabet includes 8 vowels and 24 consonants [2].

The syllable is the basic pronunciation constitution unit of Uyghur language. Each syllable must have a vowel, and a single vowel could make a syllable. The Uyghur inherent syllable types are V, VC, CV, VCC, CVC, CVCC (V represents vowel, C represents consonant). The modern Uyghur language also have the additional syllable types, such as CCV, CCVC, CCVCC, CVV, CVVC, because the Uyghur language includes considerable amount loanwords from Chinese, Arabic, Persian and Russian.

The Uyghur word is the minimum meaningful language unit, two words splits by a blank space in a sentence. Uyghur language is an agglutinative language. It is possible to produce

a new word by using the word-building suffixes. A same root word may connect many different suffixes and form a different word. For example,

sawaK	knowledge
sawaK+dax=sawaKdax	schoolmate
sawaK+dax+lar=sawaKdaxlar	schoolmates

The prosody of Uyghur language mainly includes the stress, intonation, and stop. The stress mainly falls on the second syllable in the two syllable words.

III. DATA COLLECTION

In speech technology, the maturity and the scientificity of speech database is very important [3]. The speech database almost includes the whole linguistics feature and prosody of a certain language, and is suitable for the design request of speech recognition systems. High-quality, large-scale and diversified speech database is of important significance to promote the development and application of speech recognition technology.

Speech Database includes the training set and testing set, and it is used for training and testing of the acoustic model respectively. The processes of building speech database include preparations (e.g. recording devices, recording software, etc.), text collection and selection, speaker selection, recording, and labeling.

The Uyghur text is written from right to left like Arabic, and the transcription inconvenient. For easily processing Uyghur texts, we described the Uyghur characters by using the Latin characters [4], which shown in figure 1.

ن	م	ل	ك	ج	ئى	ھ	غ	گ	ف	ې	ئە	د	چ	ب	ئا
n	m	l	k	j	i	H	G	g	f	e	E	d	q	b	a
ژ	ز	ي	خ	ۋ	ئۇ	ئۆ	ت	ش	س	ر	ق	پ	ئۆ	ئو	ك
Z	z	y	h	w	v	u	t	x	s	r	K	p	O	o	N

Figure 1. Uyghur-Latin characters

A. Text Corpus

Our text corpus for recording was collected from Xinjiang Daily, Uyghur WebPages, Uyghur novels, Uyghur dictionary, Research papers and Telescripts. The proportion of each text type is shown in table 1. These texts were pre-processed and divided into sentences. They were checked manually, and only the sentences that were complete and well-formed were included while the rest were discarded. At last our original text corpus includes more than 30,000 sentences. These sentences have been used to select the recording text and building the language model.

The principle of text corpus selection of speech recognition is to include the more natural language phenomena by using small text corpus. According to the phonetic context, we chose the sentences from original text corpus by using Greedy algorithm [5]. For the continuous Uyghur speech recognition, we selected 1018 sentences which include more than 5,500 commonly used Uyghur words, more than 5,000 triphones; each sentence mostly includes 5-15 words.

TABLE I. PROPORTION OF TEXT TYPE

News	Novels	Dictionary	Research	Telescript
75.6%	12.5%	7.4%	1.5%	3.0%

B. Speech Database

Based on the process of Uyghur Speech Database [4], we established a Speech Recognition Speech Database by using the 1018 sentences.

Our speech database consists of 64 speakers (32 male, 32 female), most of them speaking standard central dialect of Uyghur, and came from Urumqi. The average age is 22; each speaker read 100 sentences randomly in their normally reading speed. The sampling rate of collecting speech data was set to 16,000 Hz with 16 bit per sample. Each sentence was saved to one wav file. The size of this speech database is about 1.5GB, and 10 hours long. The 54 speakers' speech was used to training acoustic model, and the 10 speakers' speech was used to test the Uyghur continuous speech recognition.

IV. RECOGNITION UNIT SELECTION

The selection of recognition unit is one basic and important problem in acoustics modeling. In continuous speech recognition we could use the phrases, words, syllables, and phonemes as the recognition unit [6]. Usually, we don't use the phrases and words, because the context relation between words or phrases is too long than syllables or phonemes, causes the model training insufficient.

In the Uyghur continuous speech recognition, we may choose the syllable or phoneme as the recognition unit. But there are more than 5000 syllables in Uyghur, if the context relation between syllables, need to train more than 5000×5000 ×5000 models. We couldn't use the existing speech database to obtain the full training syllable models. Based on the above reasons, we select the phoneme as the recognition unit for Uyghur continuous speech recognition. There are 34 monophones including silence and short pause.

The coarticulation phenomenon is extremely common in continuous speech, therefore should be consider the context relations of monophone in continuous speech recognition. So that we would expands the monophone to the context-dependent triphone.

In this work, we will select the monophone and triphone as the recognition unit for Uyghur continuous speech recognition, and training the acoustic model by using our speech database.

V. ACOUSTIC AND LANGUAGE MODEL

This is the important part of the continuous speech recognition. In this section, we describe training of the acoustic model and language model by using speech database and text corpora respectively.

A. Acoustic Model

Acoustic model is used for simulating the speech signal, which describes the speech signal in order to simulate human speech production and perception characteristics. Acoustic model will extract voice feature vector, calculate the observation probability of feature vector, and build HMM modules for each recognition unit (the monophone and triphone), through the training of these models to enable the match between the voice and the acoustic signals [1].

According to the characteristic of Uyghur language, we use the 5 state HMM topology for monophone or triphone (the state 1 is beginning and the state 5 is ending), the 3 state HMM topology for silence SSS and short pause SP, and training the acoustic model for each unit by using 54 speakers' speech data.

We can get $34 \times 34 \times 34$ triphone by combining the 34 phonemes, but very many triphones cannot appear in the Uyghur language. For decrease the number of training triphone model, we establish the Uyghur question set, and tied the similar pronunciation triphone, let them sharing states each other by using decision tree method. Thus, our speech database is enough to train the regarded triphone models.

B. Language Model

A large vocabulary continuous speech recognition system must inspect whether it is on the encounter voice pronunciation border or not. So many different possible characters or words will be recognized from different sub-stream voice. These ambiguities between characters or words have been eliminated by using statistical language model. Statistical language models can provide the context information between characters or words, and solve the inherent deficiencies caused by the acoustic model [7].

Statistical language model calculate the probability of various words and conditional probability by using the text corpus. In this work, we built the 2-gram language model by using the text corpus of Uyghur (including 30,000 sentences). The OOV rate of this language model was 8.7% for the 1018 sentences.

VI. Experiments and Results

A. Experiments

The experiment of Uyghur continuous speech recognition is based on the HTK3.3 [8]. The training set is the 54 speakers' speech data of second part speech database. In experiment I, the test setI is the 10 speakers' speech data of second part speech database. In experiment II, the test setII is the 100 utterances of first part speech database.

In these experiments, we compared the recognition performance of three systems based on different recognition unit. The first is monophone unit, the second is triphone unit with untied and tied triphone, and the third is mixture triphone unit with different mixture number.

The recognizers in these experiments used MFCC_0_D_A with 12 MFCC coefficients, the feature vector have 39 dimensions with cepstrum coefficients, delta coefficients and coefficients acceleration.

B. Results

1) Experiment I

Table 2 shows the results of unit type experiment with word accuracy (Word Acc) and sentence accuracy (Sent Acc) for the test setI. The triphone unit recognition has better performance than the monophone unit recognition, demonstrating the effectiveness of triphone modeling. The mixture triphone unit recognition has better recognition accuracy than the other recognitions, showing that this unit suitable for Uyghur large vocabulary continuous speech recognition.

Untied and tied triphone recognition results indicate that the state-sharing in acoustic modeling increase the accuracy of speech recognition.

The recognition performance was increased along with the increasing number of mixture number. When the mixture number was increased to 10, the recognition performance start to drop, this explains the training data was not enough to train these models.

TABLE II. Recognition performances

recognition unit		Word Acc	Sent Acc
monophone		58.90%	4.81%
triphone	*untied*	86.05%	30.26%
	tied	89.27%	38.11%
mixture	*4*	93.65%	57.56%
	6	94.12%	60.30%
	8	94.44%	63.36%

2) Experiment II

Table 3 shows the results of unit type experiment with word accuracy and sentence accuracy for the test setII. These results show that the experiment II has the poorer recognition performance than experiment I. This indicates that the speech data for training should include more unit models.

TABLE III. Recognition performances

recognition unit		Word Acc	Sent Acc
monophone		35.60%	1.00%
triphone	*tied*	69.93%	22.00%
mixture	*6*	76.73%	41.00%

This is the important part of the continuous speech recognition. In this section, we describe training of the acoustic model and language model by using speech database and text corpora respectively.

VII. CONCLUSION

In this paper, we present our study on large vocabulary continuous speech recognition for Uyghur with a reading speech database. The results show that tied mixture triphone unit selection based on our knowledge of Uyghur acoustic-phonetics is the better choice in comparison with the monophone unit and untied triphone unit.

In comparison with large vocabulary continuous speech recognition for Chinese and English, our system has less performance with 94.44% word accuracy and 63.36% sentence accuracy. This could be caused by the limited number of speakers and the text corpus for building language model is not enough to cover the Uyghur language phenomena. Our future works focus on collecting more Uyghur text, increasing the number of the speaker, collecting the spontaneous speech data and the convertional speech data.

ACKNOWLEDGMENT

This work was supported by the China National Natural Science Foundation NSFC (60762006), the China Education Ministry Fund (MZ115-75), and the Scientific Research Program of the Higher Education Institution of Xinjiang (XJEDU2006S10).

REFERENCES

[1] L. R. Rabiner and B. H. Juang, "*Fundamentals of Speech Recognition*", Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] Mirsultan, "*Modern Uyghur Language*", Peoples Publishing Home of Xinjiang, Urumqi, 1987.

[3] Li Aijun, Wang Tianqing, and Yin Zhigang, "RASC863 Speech Recognition Speech Database of 863 project−Four Regional Accent Speech Database Chinese", *In proceedings of NCMMSC7*, 2003, pp. 274-277.

[4] N. Tursun, W. Silamu, M. Tursun, "Research of Large Vocabulary Continuous Uyghur Speech Recognition—The Design of Speech Database", *11th Symposium of the National Language and Information Technology*, Xishuang Banna, China , 2007, pp. 379-385.

[5] M. Boldea, C. Munteanu, A. Doroga, "Design, Collection and Annotation of a Romanian Speech Database", *In Proceedings of the First LREC - Workshop on Speech Database Development for Central and Eastern European Languages*, Granada, Spain, 1998.

[6] S. J. Young and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling", *Proc. Human Language Technology Workshop,* 1994, pp. 307–312.

[7] P. Clarkson and R. Rosenfeld, "Statistical language modeling using CMU-Cambridge toolkit", *In EURO SPEECH-97*, 1997, pp. 2707–2710.

[8] S. J. Young, G. Evermann, M. Gales, etc, "*The HTK Book (for HTK Version 3.3)*". Cambridge University Engineering Department, 2005.