Free Linguistic and Speech Resources for Tibetan

Guanyu Li[1], Hongzhi Yu[1], Thomas Fang Zheng[2], Jinghao Yan[1]
[1]Key Laboratory of National language intelligent processing, Gansu Province, Northwest Minzu University, Lanzhou, China
E-mail: guanyu-li@163.com Tel: +86-13809316272
[2]Center for Speech and Language Technologies, Tsinghua University, Beijing, China
E-mail: fzheng@tsinghua.edu.cn Tel: +86-13801012234

***Abstract*— Tibetan is an important low-resource language in China. A key factor that hinders the speech and language research for Tibetan is the lack of resources, particularly free ones. This paper describes our recent progression on Tibetan resource construction supported by the NSFC M2ASR project, including the phone set, lexicon, as well as the transcription of a large scale speech corpus. Following the M2ASR free data program, all the resources are publicly available and free for researchers. We also release a small Tibetan speech database that can be used to build a proto type Tibetan speech recognition system.**

I. INTRODUCTION

Tibetan language is a key member in the family of minor languages in China. It belongs to the Sino-Tibetan language family, the Tibeto-Burman subgroup. The speakers are about 6 million people, mainly distributed in China (Tibet, Qinghai, Gansu, Sichuan and Yunnan provinces), India, Bhutan, and Nepal. Compared to the major languages such as English and Mandarin, the research for Tibetan is far from extensive, on both linguistics and speech processing. A key factor that hinders the research is that the resources are very limited and far from being standard. For example, the lexicon is still in a small scale, and large-scale speech databases are very rare. Most seriously, most of the resources are held by individual institutes, with very limited sharing and openness.

The Multilingual Minorlingual Automatic Speech Recognition (M2ASR) project aims to change the situation. An ambition of this project is to construct a full set of language and speech resources for 5 minor languages (Tibetan, Mongolia, Uyghur, Kazak and Kirgiz), and make the resources open and free for research purposes. In this paper, we report our progress on Tibetan resource construction, including the phone set, the lexicon, transcriptions and speech databases. All the resources are available on the project webpage (http://m2asr.cslt.org), and can be obtained by either free download or delivery on request.

Note that there are 3 Tibetan dialect areas in China: U-Tsang, Amdo, Kham. People in the three areas use the same written form, but pronounce very differently. In U-Tsang, the most popular dialect is the Lhasa Tibetan, and in Amdo, the Xiahe Tibetan (or Labrang Tibetan) is the mostly influential. The M2ASR project focuses on the two dialects as they are spoken by most of Tibetan people. In the following sections, we will first briefly summarize the written and pronunciation system of Tibetan, and then propose our work on resource construction for the two dialects respectively.

II. CHARACTERS, SYLLABLES AND WORDS IN TIBETAN

Tibetan scripts are written in alphabets. From view of written form, there are 30 consonant letters and 4 vowel signs in Tibetan (note all dialects are the same in writing). Each syllable is a combination of several consonant letters and a vowel sign. Words are comprised of one or several syllables. In the Tibetan script, syllables and words are written from left to right, and are separated by the same delimiter " · " (called ཚེག (/tsheg/) in Tibetan).

Each syllable involves a radical consonant letter, and other consonant letters could be appended to the radical consonant as superscript, subscript, prescript, postscript and post-postscript to form a syllable (Fig 1). A syllable must contain a vowel sign, but a vowel sign corresponds to a sound /a/ can be omitted. In general, the vowel signs ི, ུ, ེ, ོ sound /i/, /u/, /e/, /o/ respectively, but exceptions also exist, as their pronunciations can be changed following some regular rules. Note that in all the dialects of Tibetan, two syllables may be pronounced the same but each syllable has only a single pronunciation. In

Supported by the National Natural Science Foundation of China under Project 61633013

other words, there are many homophones but no polyphones in Tibetan.

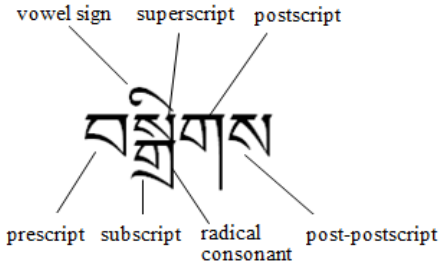

Fig 1 Constitution Of Syllable

The radical consonant, the prescript and superscript consonants together form the initial part of a syllable, and the vowel sign, the postscript and post-postscript consonants altogether form the final part. In ancient Tibetan, there are many consonant compositions. These consonant compositions are largely preserved in the modern Xiahe dialect, however in the Lhasa dialect, most consonant compositions sound just like a single constant. Another distinction between the Lhasa dialect and the Xiahe dialect is that the former is tonal (four tones in total: 43, 44, 12 and 113[1]) while the latter is toneless. For these reasons, the two dialects sound very different and should be treated different in resource construction.

III. RESOURCES CONSTRUCTION FOR LHASA TIBETAN

In this section, we describe our work on resources construction for the Lhasa dialect. The resources include the phones set, the syllable lexicon, the word lexicon, text database and speech database.

A. Phone set

The phone set involves small and distinct pronunciation units. These units are related to the consonants and vowels in the written form, and more reflect the true pronunciation. We follow the seminal work by Ge Sang Ju Mian [1] and define 29 consonants and 8 cardinal vowels in Lhasa Tibetan. The phone clusters of the consonants are presented in Table I, where the consonant /f/ only appears in foreign words. The vowels in Lhasa Tibetan are listed in table II [1][2][3]. There is a long vowel form for each of the 8 cardinal vowels; 5 vowels (/ɛ/, /e/, /ø/, /i/ and /y/) have a glottalized form and 3 vowels (/e/, /i/ and /y/) have a nasalized form. Additionally, there are 4 consonants (/k/, /m/ and /p/) that can be augmented to the end of a vowel to form a coda, and there are two compound vowels: /au/ and /iu/.

Table Ⅰ Clusters Of Lhasa Tibetan Consonants

			bilabial	labiodental	Apical alveolar	retroflex	Palatal	Velar	Labial velar	glottal
Plosive	voiceless	unaspirated	p		t		c	K		ʔ
		aspirated	p^h		t^h		c^h	k^h		
Affricate	voiceless	unaspirated		f	ts	ʈʂ	tɕ			
		aspirated			ts^h	$ʈʂ^h$	$tɕ^h$			
Fricative		unaspirated			s	ʂ	ɕ			h
nasal	voiced	unaspirated	m		n		ɲ	ɴ		
approximant	voiced	unaspirated					j		w	
Lateral fricative					ɬ					
Lateral approximant	voiced	unaspirated			l					
trill	voiced	unaspirated			r					

Table Ⅱ Vowels Of Lhasa Tibetan

characters			vowels	long	glottalized	nasalized	Post consonant				
tongue position		rounded					k	m	p	u	ŋ
front	low		a	a:			ak	am	ap	au	aŋ
front	medium low		ɛ	ɛ:	$ɛ^ʔ$						
front	medium high		e	e:	$e^ʔ$	ẽ		em	ep		eŋ
front	medium high	rounded	ø	ø:	$ø^ʔ$						
front	high		i	i:	$i^ʔ$	ĩ	ik	im	ip	iu	iŋ
front	high	rounded	y	y:	$y^ʔ$	ỹ					
back	medium	rounded	o	o:			ok	om	op		oŋ
back	high	rounded	u	u:			uk	um	up		uŋ

For speech recognition, we made a slight modification to construct the ASR phone set. The first modification is that we merge a cardinal vowel with its corresponding long vowel form. The second modification is to treat /au/ and /iu/ as two single phones rather than splitting them into ingredient phones. To ease the text processing with computers, all these phones (IPAs) are transformed into Latin letter, as shown in Table III.

Table Ⅲ Phones And Latin Transformation OfLhasa Tibetan

IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin
c	c	l	l	s	s	ŋ	ng	tɕ	tx	ɛ	Ec	$ø^ʔ$	edb	ỹ	yy
c^h	ch	m	m	t	t	ɲ	nn	$tɕ^h$	txh	$ɛ^ʔ$	Ecb	i	i	o	o
h	h	n	n	t^h	th	ɕ	x	ʔ	ab	e	E	$i^ʔ$	ib	u	u
j	j	p	p	ʈʂ	q	ʂ	ss	f	f	$e^ʔ$	Eb	ĩ	ii	au	au
k	k	p^h	ph	ts^h	tsh	ʈʂ	ds	ɬ	lh	ẽ	Ee	y	y	iu	iu
k^h	kh	r	r	w	w	$ʈʂ^h$	dsh	a	a	ø	Ed	$y^ʔ$	yb		

B. Syllable lexicon

A syllable lexicon involves a set of syllables whose pronunciations are defined. There are more than 8000 possible syllables in Tibetan, including syllables for foreign words. We construct the syllable lexicon by constructing a text corpus involving 420,000 sentences (including both written and spoken), and then selected the most frequent

syllables from this corpus. After removing some syllables that are for transliterating Sanskrit words only, we obtained a syllable lexicon consisting of 6013 syllables. By applying the pronunciation rules, these syllables were segmented into initials and finals, and the initials and finals were further split into phones [4]. All these syllables and their phone sequence forms were manually checked to ensure the quality.

C. Word lexicon

The word lexicon translates words into syllable sequences. To construct the lexicon, the same text corpus used in the syllable lexicon construction is used to form the word list. This is performed by a word segmentation followed by a frequency-based filtering. By this approach, we obtained 27000 frequent words. This primary set was further extended by adding two extra sets: a set of 10000 nouns and a set of 14000 verbs. This results a word lexicon consisting of 51000 words in total. Again, the quality was ensured by manual check.

The treatment for abbreviations and foreign words deserve some discussion, as this may have impacted the quality of the resultant lexicon.

1) Abbreviations

In Tibetan, a case particle might be added to the end of a word to form a new syllable. These case particles include ར, འི, ས, འོ, འང and འམ, and should be treated carefully. In addition, some shorthand notations are commonly seen in normal printed text, they should be treated as new Tibetan letters, e.g., ཏ is used as the abbreviation of གིས [5]. The pronunciations for these new forms should be carefully settled.

2) Foreign words

Some syllables to represent Sanskrit words are also used to represent foreign words in Modern Tibetan. These syllables are added to the word lexicon, e.g. "ཀྵ". As another example, to represent the foreign pronunciation /f/, ཕྷ (/fa/) with vowel signs are added the word lexicon as well. ཧྭ is also added to the word lexicon for it is often used to represent the foreign pronunciation /hua/.

D. Text and transcription

We collected a 100M text database of Tibetan for language modeling. From this huge database, we also extract a transcription that will be used to collect the initial speech database that involves 50 hours of speech signals. The sources of the text include Tibetan newspapers, official documents and sentences of spoken Tibetan. The sentences in the text database are transformed into phones and triphones sequences, using the syllables lexicon. Then a maximum entropy method was used to select the most representative sentences. By this selection, about 30,000 sentences (including 13,654 triphones) were chosen from the original database that contains more than 400,000 sentences (20,015 triphones). The triphone coverage rate is 68.22%. The number of syllables in these sentences is between 5 and 25.

E. Speech database

Although the large-scale M2ASR speech database is still under construction, we release a small-scale database following the M2ASR free data program. The data was recorded by the first author of this paper. It consists of 15 hours of speech signals recorded from 34 speakers, where the sampling rate is 16k Hz and the sample size is 16 bits. All the speakers are in Lhasa dialect, and the recording style is reading. This database is free for research purpose, available on request.

IV. RESOURCES CONSTRUCTION FOR XIAHE TIBETAN

In this section, we describe our work on the XiaHe dialect. Only the work special to this dialect will be presented.

A. Phone set

There are 35 consonants in Xiahe Tibetan. The clusters of the consonants are presented in Table IV[1][2][3].

Table Ⅳ Clusters Of Xiahe Consonants

			bilabial	apical alveolar	retroflex	palatal	velar	labial velar	uvular	glottal
plosive	voiceless	unaspirated	p	t			k			
		aspirated	p^h	t^h			k^h			
	voiced	unaspirated	b	d			g			
affricate	voiceless	unaspirated		ts	ʈʂ	tɕ				
		aspirated		ts^h	$ʈʂ^h$	$tɕ^h$				
	voiced	unaspirated		dz	dz˛	dʑ				
fricative	voiceless	unaspirated		s	ʂ	ɕ	χ		ʁ	h
		aspirated		sh						
	voiced	unaspirated		z		ʑ				
nasal	voiced	unaspirated	m	n		ɲ	ŋ			
approximant	voiced	unaspirated				j				
lateral fricative	voiceless	unaspirated		ɬ						
lateral approximant	voiced	unaspirated		l						
trill	voiced	unaspirated		r						

There are 3 types of consonant compositions: pre-add /n/, pre-add /h/ and others (e.g., /xw/ and /kw/). These are shown in Table Ⅴ and TableVI [1][2][3].

Table Ⅴ Consonant Compositions (Pre-add /n/)

	b	dz	d	dz˛	dʑ	G
n	nb	ndz	nd	ndz˛	ndʑ	Ng

Table Ⅵ Consonant Compositions (Pre-add /h/)

	m	t	ts	N	l	ʈʂ	tɕ	ɕ	k	ŋ
h	hm	ht	hts	hn	hl	hʈʂ	htɕ	hɕ	hk	hŋ

There are 6 cardinal vowels in Xiahe Tibetan: /a/, /e/, /o/, /ə/, /i/, /u/. 7 consonants may be added to the end of /a/, /e/, /o/ and /ə/ to form coda consonants. [1][2][3].

Table Ⅵ Coda Consonants In Xiahe Dialect

	p	m	t	n	k	ŋ	R
a	ap	am	at	an	ak	aŋ	ar
e	ep	em	et	en			er
o	op	om	ot	on	ok	oŋ	or
ə	əp	əm	ət	ən	ək		ər

We constructed the phone set for Xiahe Tibetan based on the phoneme letters. The consonant compositions were treated as one phone, where the pronunciation of the second consonant is weary. The phone set is presented in Table VIII.

TABLE VIII. PHONES AND LATIN TRANSFORMATION OF LHASA TIBETAN

IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin
p	p	ts	ts	ɬ	lh	tɕ	tx	k	k	nb	nb	ht	ht	hk	hk	u	u
p^h	ph	ts^h	tsh	l	l	$tɕ^h$	txh	k^h	kh	ndz	ndz	hts	hts	hŋ	hng		
b	b	dz	dz	r	r	dʑ	dp	g	g	nd	nd	hn	hn	a	a		
m	m	s	s	ʈʂ	ds	ɕ	x	ɣ	xx	ndz	ndr	hl	hl	e	e		
t	t	sh	sh	$ʈʂ^h$	dsh	ʑ	zz	ŋ	ng	ndʑ	ndp	hʈʂ	hds	ə	ee		
t^h	th	z	z	dz̤	dr	ɲ	nn	ʁ	v	ng	ng	htɕ	htx	i	i		
d	d	n	n	ʂ	ss	j	j	h	H	hm	hm	hɕ	hx	o	o		

B. Lexica

We constructed a syllable lexicon consisting of 5900 syllables. The construction of the word lexicon is based on the Lhasa word lexicon, with some oral words different from the Lhasa dialect added.

V. SPEECH RECOGNITION PROTO TYPE

In the final part, we present a speech recognition proto type system based on the resources provided for the Lhasa dialect. This can be regarded as a preliminary check for the resources we released.

The experiment was conducted using the Kaldi toolkit, following the WSJ s5 recipe. After training the GMM-HMM system, a DNN system was constructed with layer-wised pre-training. The DNN model is optimized based on the loss function of cross-entropy, using a single GPU[6][7].

The training data consists of 20.5 hours of speech signals (23053 utterances of 41 speakers). The size of syllable lexicon is 6013. A trigram language model trained on 310,000 Tibetan text sentences was used in decoding. The test set consists of 1,910 utterances of 3 speakers. The syllable error rate of this proto type system is 24.6%. This result demonstrated that the resources we released are in reasonable quality.

VI. CONCLUSIONS

We described a set of free resources that we published under the M2ASR free data program. These resources include the phone set, syllable and word lexica, text database and speech signals. We also presented a proto type systemthat demonstrated the quality of the release.

Much work remains. Particularly, we will finish the speech recording of more than 200 hours and release the data for the speech community.

REFERENCE

[1] Ge Sang Ju Mian, Ge Sang Yang Jing, “A Introduction to Tibetan Dialects”, The Ethnic Publishing House, July 2002

[2] Hu Tan, Suo Nan Zhuo Ga, Luo Bingfen, Oral Lhasa Tibetan Reading, The Ethnic Publishing House, Jan 2013

[3] Gon Que Jiang Cuo, On Tibetan Language, Research of Tibetan, vol.3, pp 94-108,1997

[4] LI Yonghong, KONG Jiangping, YU Hongzhi, Rules for the Auto-transformation of Tibetan Text to IPA, Tsinghua Univ (Sci &Tech), Vol.48, No.S1, 2008

[5] CAI Zhi-jie, Identification of Abbreviated Word in Tibetan Word Segmentation, Journal Of Chinese Information Processing, vol. 23, pp. 35-43, Jan 2009

[6] Geoffrey Hinton et al, Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research, IEEE Signal Processing Magazine , 29 (6) :82-97, 2012

[7] Website, http://kaldi-asr.org/doc/index.html