A Free Kazakh Speech Database and a Speech Recognition Baseline

Ying Shi*, Askar Hamdulla†, Zhiyuan Tang*, Dong Wang*‡, Thomas Fang Zheng*
* Center for Speech and Language Technologies, Research Institute of Information Technology,
Department of Computer Science and Technology, Tsinghua University, China
† Key Laboratory of Signal and Information Processing, Xinjiang University
‡ Corresponding Author E-mail: wangdong99@mails.tsinghua.edu.cn

***Abstract*—Automatic speech recognition (ASR) has gained significant improvement for major languages such as English and Chinese, partly due to the emergence of deep neural networks (DNN) and large amount of training data. For minority languages, however, the progress is largely behind the main stream. A particularly obstacle is that there are almost no large-scale speech databases for minority languages, and the only few databases are held by some institutes as private properties, far from open and standard, and very few are free. Besides the speech database, phonetic and linguistic resources are also scarce, including phone set, lexicon, and language model.**

In this paper, we publish a speech database in Kazakh, a major minority language in the western China. Accompanying this database, a full set of phonetic and linguistic resources are also published, by which a full-fledged Kazakh ASR system can be constructed. We will describe the recipe for constructing a baseline system, and report our present results. The resources are free for research institutes and can be obtained by request. The publication is supported by the M2ASR project supported by NSFC, which aims to build multilingual ASR systems for minority languages in China.

I. Introduction

Recently, automatic speech recognition (ASR) has gained significant improvement, partly due to the emergence of deep neural networks (DNN) and increasing amounts of training data [1], [2], [3]. For the major languages such as Chinese and English, large-scale ASR systems have been deployed by several big companies to provide ubiquitous service via numerous applications [4], [5].

These significant achievements, however, are less seen in minority languages. There are many reasons including complex economic and educational issues, but an obstacle is the lack of open, free, and standard resources. For many minor languages, there are almost no large-scale speech databases, and the only few databases are held by several institutes as private properties, far from open and standard, and very few are free. Besides the speech database, phonetic and linguistic resources are also scarce, including phone set, lexicon, and language model.

Recently, we started a multilingual minor-lingual ASR (M2ASR) project, supported by the national natural science foundation of China (NSFC). The project is a three-party collaboration, including Tsinghua University, Northwest National University, and Xinjiang University. The aim of this project is to construct speech recognition systems for five minor languages in China (Tibetan, Mongolia, Uyghur, Kazakh and Kirgiz). However, our ambition is beyond that scope: we hope to construct a full set of linguistic and speech resources for the 5 languages, and make them open and free for research purposes. We call this the M2ASR Free Data Program. All the data resources, including the ones published in this paper, are released through the website of the project[1].

In this paper, we report our progress on Kazakh resource construction. We will release a large-scale speech database and associated resources including the phone set, lexicon and language model (LM). The speech database consists of about 78 hours of speech signals recorded by 96 speakers. These resources include all required to establish a Kazakh ASR system. We will also publish a Kazakh ASR baseline system, so that researchers on Kazakh ASR can have a benchmark to evaluate their research.

The rest of the paper is organized as follows: Section II presents some work on free speech databases, and Section III describes the characteristics of Kazakh. Section IV presents database that we will release. The construction of the Kazakh ASR baseline is reported in Section V. The conclusions plus the future work are presented in Section VI.

II. Free databases for ASR

There are several famous speech databases, mostly are in English, such as WSJ[6], Switchboard[7], TIMIT[8]. For Chinese RAS 863 corpus [9] is mostly used. These databases can be publicly obtained, though most of them are distributed with commercial licences which are often expensive.

There have been some initial attempts towards free speech databases. For example, the EU-supported AMI/AMID project released all the data (both speech and video recordings on meetings), the VoxForge project[2] provides a platform to publish GPL-based annotated speech audio. OpenSLR is another platform to publish open resources for speech and language research, including the LibriSpeech[3]. The Sprakbanken database[4] for Swedish, Norwegian, Danish. In Chi-

[1] http://m2asr.cslt.org
[2] http://www.voxforge.org/
[3] http://www.openslr.org/12/
[4] http://www.nb.no/sbfil/talegjenkjenning/

nese, CSLT@Tsinghua University has released a free database THCHS30 [10], that contains about 30 hours of reading speech. A Kaldi recipe was also released to build a complete Chinese ASR system with THCHS30.

For minor languages, public and free databases are still rare. The AP16-OLR challenge[5] released a database consisting of seven oriental languages, including Mandarin, Cantonese, Indoesian, Japanese, Russian, Korean, Vietnamese. This database is free for the challenge participants, but still expensive for other researchers. Recently, the Babel project has released several very cheap speech databases for minority languages, including Assamese, Bengali, Cantonese, Georgian, Pashto, Tagalog, Turkish. Each of the database is just several dollars, nearly free.

Last year, CSLT published a free speech database for Uyghur. THUGY20 [11]. This database consists of 20 hours speech signals recorded by more than 340 speakers. Accompanying to this database, a complete set of resources that can be used to construct a full-fledged Uyghur ASR system was also published. The database and the associated resources can be downloaded from CSLT's webiste[6], and the recipe can be downloaded from github[7].

Supported by the M2ASR project, we will publish another set of free databases and associated resources. Currently, the data ready for publish include the phone set, lexicon and speech data of Tibetan, as well as the phone set and lexicon of Mongolia. The Kazakh speech database and the associated resources are part of the release. More information of the release can be found in the M2ASR project website.

III. Kazakh language

This section reviews some properties of Kazakh, and the difficulties in ASR.

A. Characters and phones

Kazakh is one of the Turkic languages and belongs to Altai language family. Most of the speakers reside in Kazakstan, Mongolia, and the Xinjiang Uygur Autonomous Region in China. There are three written forms for Kazakh alphabets, Arabic, Latin and Cyrillic. The Arabic characters are mostly used by the Kazakh people living in China. The Cyrillic characters are widely used by the Kazakh people who live in Kazakstan and Mongolia. The Latin characters were used for Kazakh people in China in 1964-1984, but it has been substituted for Arabic characters. With the wide spreading of online communication tools (e.g., WeChat), the young generation is used to using Latin characters for easier input. In most cases, pronunciation of Kazakh is strictly regularized, following the rule *reading as writing*. This means that the characters and phones are largely the same, and so whenever the writing form of a word is presented, its pronunciation can be obtained directly.

[5]http://olr.clst.org

[6]http://data.cslt.org/thuyg20/README.html

[7]https://github.com/wangdong99/kaldi/tree/master/egs/thuyg20

Fig. 1 shows all the Kazakh alphabets in different written forms. There are 33 phones in Kazakh, 9 of them are vowels and 24 consonants. The vowels can be further categorized into two groups: 4 front vowels and 5 back vowels, as shown in Table I. Note that the character 'v' (in Latin) shown in Fig. 1 is used only as a functional letter used to indicate the pronunciation change of the following vowel character (see below), and is not pronounced by itself.

Arabic	ڭ	ن	م	ل	ق	ك	ي	ز	ج
Latin	N	n	m	l	q	k	y	z	j
Cyrillic	Ң	Н	М	Л	Қ	К	Й	З	Ж
Arabic	ه	د	ع	گ	ۋ	ب	ءا	ا	ء
Latin	E	d	G	g	V	b	A	a	v
Cyrillic	Е	Д	Ғ	Г	В	Б	А	Ә	None
Arabic	ءى	ى	ش	چ	ھ	ح	ف	ءۇ	ۇ
Latin	i	e	x	c	H	h	f	U	u
Cyrillic	І	Ы	Ш	Ч	Һ	Х	Ф	Ү	Ұ
Arabic	ۇ	ت	س	ر	پ	ءو	و		
Latin	w	t	s	r	p	O	o		
Cyrillic	У	Т	С	Р	П	Ө	О		

Fig. 1. Kazakh characters in Arabic, Latin and Cyrillic.

TABLE I
Front vowels and back vowels in Kazakh.

Type	Written form				
Front vowel	A(va)	O(vo)	U(vu)	i(ve)	-
Back vowel	a	o	u	e	E

B. Vowel harmony in Kazakh

According to Kramer[12], "*Vowel harmony is the phenomenon where potentially all vowels in adjacent moras of syllables within a domain like the phonological of morphological word systematically agree with each other with regard to one or more articulatory features.*" This means that vowel harmony sets a constraint on which vowels may be found near each other in a word or word group.

Vowel harmony may result in departure of the true pronunciation from the spelling form which will hence the errors in the lexicon. These errors can be corrected by the vowel harmony rules. Unfortunately, the vowel harmony rules about Kazakh have not been fully investigated by researchers so far, and what we have known are just two rules. Firstly, if the functional character 'v' appears at the beginning of a word, all the back vowels in this word should be treated as the corresponding front vowels. For instance, the word 'vbare' should be actually pronounced as 'bAri'. Secondly, whenever the consonant characters 'k' and 'g' and the vowel character 'E' appear in a word, all the back vowels should be converted

to the corresponding front vowels. For example, 'kareN', 'elgEn' and 'esE' should be actually pronounced as 'kAriN' ,'ilgEn' and 'isE' respectively. Other vowel harmony rules about Kazakh may exist and we will leave the investigation as future work.

C. Agglutination

Kazakh is an agglutinative language. In this kind of language, a word consists of a stem and unlimited numbers of suffixes. This will result in a very huge lexicon, and new words can be easily produced by a flexible agglutinative rule. This agglutinative nature causes serious problems in language modeling, as the most of the words occur only a few times in the training data, and many words in the test may be out-of-vocabulary. This will result in inaccurate estimation of the LM, and a low coverage of the words in the test. A possible solution is to use a morpheme-based LM instead of the regular word-based LM, as will be seen in the experiment section.

IV. Database, lexicon and LM

This section presents Kazakh speech database M2ASR-Kazak-78. This database was recorded by the Xinjiang University and will be published as part of the M2ASR Free Data Program. Additionally, a full set of resources including the lexicon and LM will also be published, by which researchers can build a full-fledge Kazakh ASR system.

This database involves more than 78 hours of speech signals recorded by 96 Kazakh native speakers. All the speakers were the students of Xinjiang University and had no accent. The sentences (text prompts in recording) were selected from multiple sources, including news, novels and web pages. The number of selected sentences is $4,000$. The bi-phone coverage of the selected sentences is 89.4%, and the tri-phone coverage is 33.7%.

The recording equipments include notebook PCs, desktop PCs and mobile phones with both Android and iOS. The sample rate of the recording was set to 16 kHz, and the sample precision is 16 bits. The recording was done in silent environments, and the speakers uttered the sentences in reading style.

The entire database is split into a training set and a test set. The training set involves $4,000$ sentences, $34,392$ utterances, about 400 utterances per speaker. The test set is consisted of 350 sentences, $3,500$ utterances, about 350 utterances per speaker. There is no overlap between the two data sets, in both speakers and sentences. More details of the speakers and utterances for each data set are shown in Table II.

TABLE II
The data profile of M2ASR-Kazak-78.

Data set	No. of Speaker (Male/Female)	No. of Utter.	Hours
Training	86 (40/46)	34,392	78
Test	10 (5/5)	3,500	8

1) Lexicon: Kazakh is 'reading as writing', so it is not difficult to collect a list of words and create the pronunciation for each word. The difficulties are two folds: one is that there are a huge amount of Kazakh words and it is essential to select the most valuable words, the other is that the vowel harmony may cause pronunciation deviation. For the first issue, we chose a large text volume (see shortly after), and selected the most frequent words. Moreover, the words occurred in the transcription of the training set of M2ASR-Kazak-78 were selected and added to the lexicon. The final lexicon contains 100k words.

For the morpheme system, we constructed a morpheme lexicon based on the word lexicon. The words were firstly segmented into morphemes by a simple segmentation tool and the resultant morphemes were collected to form the lexicon. The pronunciations of the morphemes were just their spellings. The final morpheme lexicon contained 10k morphemes.

2) Text data and language model: We collected a Kazakh text corpus that contained 300k word tokens. As the writing style of Kazakh document is highly flexible with multiple character systems and codes, we selected the sentences from several news web sites run by the government. Based on this text corpus, we built two language models, a word-based 3-gram model and a morpheme-based 6-gram model. The two language models were both trained using the SRILM toolkit [13]. We assume that the morpheme-based LM is weaker, but it is capable of recognizing OOV words. Note that for the morpheme LM, the vocabulary involves a special token to represent word boundaries, so a morpheme sequence decoded using this LM can be easily converted to a word sequence.

V. Baseline system

We constructed a Kazakh ASR baseline system using the Kaldi toolkit [14], using the M2ASR-Kazak-78 database and the associated resources as described in the last section.

A. Settings

The acoustic model of the Kazakh ASR baseline system was built following the Kaldi WSJ s5 nnet3 recipe. A monophone GMM system was firstly trained with the standard 13-dimensional MFCCs plus the first and the second order derivatives, followed by which a tri-phone GMM system was trained, with the LDA and MLLT feature. The final GMM system was used to generate state alignment of the training data.

A time-delay neural network (TDNN) was trained based on the alignments generated from the GMM system. The features were 40 dimensional FBanks, with a symmetric 4-frame window to splice neighboring frames. The deep structure contained 6 hidden layers, and the activation function was p-norm, by which the input dimension was $2,000$ and output dimension was 250. The final output layer was composed of $3,725$ units, equal to the total number of Gaussian mixtures in tri-phone based GMM system. We employed the natural stochastic gradient descent algorithm (NSGD) [15] as the

optimization algorithm, and cross entropy as the cost function to train the TDNN model. More details about the training procedure can be found in the recipe published together with the database.

B. Results

The performance of the TDNN-HMM system and tri-phone GMM-HMM system are presented in Table III. The word system is evaluated in terms of word error rate (WER). For the morpheme system, we evaluate the performance in terms of both WER and morpheme error rate (MER). The WERs of the two systems are comparable. Considering the data volume, this performance seems a reasonable reference.

TABLE III
PERFORMANCE OF THE KAZAKH ASR BASELINE

	Word (WER%)	Morpheme (WER%/MER%)
GMM	29.46	28.74/18.18
TDNN	25.06	24.12/15.16

VI. DISCUSSION AND CONCLUSION

In this paper we reported our recent progress on Kazakh speech and language resource construction, including phone set, lexicon, text data, language model and speech database, M2ASR-Kazak-78. We will release all the resources to research institutes for free. Moreover, we constructed a Kazakh ASR baseline system. This system, although not very powerful yet, is a good reference for Kazakh ASR researchers. The Kaldi recipe of this baseline system will be published together with the M2ASR-Kazak-78 database. We hope that these publications can establish a standard resource set, a standard training procedure, and a standard evaluation metric for Kazakh ASR research.

There are many things we plan to do in the future. Firstly, the M2ASR-Kazak-78 was not designed very well. Although the database consists of 78 hours speech signals, it only contains $4,000$ sentences, and most of the participants recorded the same transcriptions. These heavily impacts the phone coverage and reduces the diverse convolution between linguistic contents and speaker traits. Moreover, the silent recording environment is far from practical scenarios. We are collecting more data, hopefully in real-life scenarios.

Another shortage of the present research is that our language model is still weak. The writing style of Kazakh is highly flexible, especially in online chatting records. These noisy text is difficult to be used unless we can find an efficient normalization technique. Ironically, these noisy sources are most close to spoken language, and most useful for the ASR system. In the present study, to alleviate the impact of noise, we use only official web sites as the text source, which are obviously suboptimal. Text normalization will be the focus of our future work.

Finally, although we can make more efforts in speech database construction, this process is costly and time-consuming. Collecting minority language speech is much more challenging than collecting major languages speech, and it is not possible to expect a large volume speech data to be accumulated in a short time. More technical solutions should be investigated, particularly various transfer learning approaches, to make use of the similarity between languages in the same region (e.g., Uyghur and Kazakh), and the similarity of all human languages.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61371136 / 61633013 / 61462084 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

REFERENCES

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[2] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. Springer, 2014.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

[4] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, 2014.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[7] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[8] C. Lopes and F. Perdigao, "Phoneme recognition on the timit database," *Speech Technologies*, 2011.

[9] A. Li, Z. Yin, T. Wang, Q. Fang, and F. Hu, "Rasc863-a chinese speech corpus with four regional accents," *ICSLT-o-COCOSDA, New Delhi, India*, 2004.

[10] Z. Z. Dong Wang, Xuewei Zhang, "Thchs-30 : A free chinese speech corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882

[11] Z. Z. D. W. A. H. Askar Roze, Shi Yin, "Thugy20: A free uyghur speech database," in *NCMMSC'15*, 2015.

[12] M. Krämer, *Vowel harmony and correspondence theory*. Walter de Gruyter, 2003, vol. 66.

[13] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *Interspeech*, vol. 2002, 2002, p. 2002.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[15] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.