information

MDPI

Article

M2ASR-KIRGHIZ: A Free Kirghiz Speech Database and Accompanied Baselines

Ikram Mamtimin [1,†], Wenqiang Du [2,†] and Askar Hamdulla [1,*]

[1] School of Information Science and Engineering, Xinjiang University, Ürümqi 830017, China
[2] Center for Speech and Language Technologies, BNRist, Tsinghua University, Beijing 100084, China
[*] Correspondence: askar@xju.edu.cn; Tel.: +86-139-9922-1222
† These authors contributed equally to this work.

Abstract: Deep learning has significantly boosted the performance improvement of automatic speech recognition (ASR) with the cooperation of large amounts of data resources. For minority languages, however, there are almost no large-scale data resources, limiting the development of ASR technologies in these languages. In this paper, we publish a free Kirghiz speech database accompanied by associated language resources. The entire database involves 128 h of speech data from 163 speakers and corresponding transcriptions. To our knowledge, this is the largest Kirghiz speech database that is dedicated to the ASR task and is publicly free so far. In addition, we also provide several baseline systems based on Kaldi and WeNet to demonstrate how these public data resources can be used to facilitate the Kirghiz ASR research. This publication is a part of the M2ASR project, and all the resources can be downloaded at the project webpage .

Keywords: automatic speech recognition; speech corpus; Kirghiz; deep learning

1. Introduction

The development of an automatic speech recognition (ASR) system for a new language requires a huge amount of annotated speech data and associated texts, especially with modern deep learning methods. However, for most low-resource languages, there are no existing large and standardized speech and text corpora. Unfortunately, most of the existing 7000 languages in the world are low-resource languages. This severely limits the application of ASR technology to people who are native in these languages.

The Multilingual Minorlingual Automatic Speech Recognition (M2ASR) project [1] aims to change the situation. The aims of this project are to construct a full set of speech and text resources for five minor languages (Tibetan, Mongolian, Uyghur, Kazakh, and Kirghiz) in the northwest of China, and make the resources open and free for research purposes. The project team involves researchers from three institutes, including Tsinghua University, Northwest Minzu University, and Xinjiang University. By the end of December 2021, the M2ASR project published about 450 h of speech data in four languages: Tibetan, Mongolia, Uyghur, and Kazakh, namely M2ASR-Tibetan (72 h) [2], M2ASR-Mongo (170 h) [3], M2ASR-Uyghur (136 h) [4], and M2ASR-Kazakh (78 h) [5], respectively. These databases have been used by dozens of research institutes so far.

In this paper, we report the data resources we have published for Kirghiz, including a speech database that contains 128 h of speech from 163 individual and a lexicon that involves 45,815 words. We name this new database *M2ASR-Kirghiz*. All the resources are available on the project webpage (http://m2asr.cslt.org, accessed on 8 October 2022) and can be obtained on request. Prior to our work, building resources for Kirghiz was rare. In particular, there were no high-quality, large-vocabulary Kirghiz speech data for free. The only Kirghiz speech data we found in literature were from the Common Voice project [6], which has collected 46 h of Kirghiz speech data via crowdsourcing (until 21 September 2022). As far as we know, M2ASR-Kirghiz is the largest free Kirghiz speech

check for updates

Citation: Mamtimin, I.; Du, W.; Hamdulla, A. M2ASR-KIRGHIZ: A Free Kirghiz Speech Database and Accompanied Baselines. *Information* **2023**, *14*, 55. https://doi.org/10.3390/info14010055

Academic Editor: David Martins de Matos

Received: 1 November 2022
Revised: 18 December 2022
Accepted: 6 January 2023
Published: 16 January 2023

database so far. With these data resources, we built two ASR baselines based on Kaldi and WeNet, respectively. The recipes have been published online (https://github.com/M2ASR, accessed on 8 October 2022).

The rest of the paper is organized as follows: Section 2 presents an introduction of the Kirghiz language; Sections 3 and 4 describe the collection details of text data and speech data, respectively; Section 5 reports the experiments with two baseline systems. The conclusion, analysis, and future work are presented in Section 6.

2. Kirghiz Language

Kirghiz is one of the ethnic minorities in China, and most of the Kirghiz population is distributed in the Xinjiang Uyghur Autonomous Region, especially in Kizilsu Kirghiz Autonomous Prefecture. There are two dialects of Kirghiz, northern dialect and southern dialect, with the Kizilsu River as the boundary. Both dialects have the same vocabulary while the pronunciations are different. A review of the phonology, orthography, and morphology of the Kirghiz language is presented in the following paragraphs.

2.1. Phonology

The phonological system of the Kirghiz language has 36 phonemes, including 14 vowels (8 basic and 6 long) and 22 consonants. The vowels and consonants, 36 phonemes in total, are written by 30 Arabic letters. Specifically, the eight basic (short) vowels are: ɑ, e, ə, i, o, ø, u, y, and the six long vowels are: ɑɑ, ee, oo, øø, uu, yy. The 22 consonants in Kirghiz are: b, p, f, v, d, t, g, ʁ, k, q, m, j, s, z, ʤ, ʧ, ʃ, n, ŋ, l, r, x. Among these, v, f, and x are used exclusively to spell loanwords. Table 1 summarizes the letters and phonemes. More details are provided below.

Table 1. The phonological system of Kirghiz [7].

vowel	ا	و	ۇ	ۉ	ۆ	ئ	ە	ى
	[a]	[o]	[u]	[v]	[ø]	[i]	[e]	[ə]
	اا	وو	ۇۇ	ۉۉ	ۆۆ	ەە		
	[aa]]	[oo]	[uu]	[vv]	[øø]	[ee]		
Consonant	ب	پ	ن	ت	ج	چ	ح	ع
	[b]	[p]	[n]	[t]	[ʤ]	[ʧ]	[x]	[ʁ]
	ق	ف	ك	گ	ل	ڭ	م	ۋ
	[q]	[f]	[k]	[g]	[l]	[ŋ]	[m]	[w]
	س	ش	د	ر	ز	ي		
	[s]	[ʃ]	[d]	[r]	[z]	[j]		

2.1.1. Vowels

The eight basic vowels are spelled with a single letter, while the six long vowels are spelled with a repetition of the corresponding basic vowel letters. The 14 vowels in Kirghiz can be further classified according to the mouth's config (opening or closing), the tongue's position (front or back), and the lip's shape (rounded or unrounded). Table 2 shows properties of each vowel [8]. Note that long vowels exist in other Altaic languages, e.g., Khakasi, Tuvan, Yakut. In these languages, the number of long vowels is 8 rather than

6 as in Kirghiz. This is a characteristic that distinguishes Kirghiz from some other Altaic languages [9].

Table 2. The vowels of Kirghiz.

	Front				**Back**			
	Unrounded		**Rounded**		**Unrounded**		**Rounded**	
	Short	**Long**	**Short**	**Long**	**Short**	**Long**	**Short**	**Long**
Close	i		y	yy	ə		u	uu
Open	e	ee	ø	øø	ɑ	ɑɑ	o	oo

There are two types of long vowels: natural long vowels (also called primary long vowels) and derived long vowels (also called secondary long vowels). Natural long probably existed in the earliest stages of the language, while deriving long vowels emerged during the long-term language development, by either the assimilation of two vowels into a long vowel after the consonant between the two vowels has fallen off or the gradual change of the last vowel into a long vowel after the consonant at the end of the word has fallen off. Evidence from experimental phonetics shows that the average duration of long vowels is 686 milliseconds, and the average duration of short vowels is 285 milliseconds [8]. Besides the discrepancy in duration, long and short vowels are also different in muscle contraction and articulatory vibration [10].

The long vowels of Kirghiz function in two ways: (1) distinguish word meaning. For example, tɑm (wall) -> tɑɑm (taste), tɑr (narrow) -> tɑɑr (pocket), and mɑl (livestock) -> mɑɑl (time). (2) Add additional suffixes. For example, qon (live) –> qonuu (staying), kør (look) –> køryy (looking), jɑz (write) –> jɑzuu (text).

Kirghiz vowels can occur anywhere in the word. In general, vowels occur more frequently at the beginning or in the middle of a word while rarely occurring at the end. The occurrence of long vowels is more constrained than short vowels, especially the long vowel *yy*, which does not appear at the beginning of a word, and its occurrence in other places is also infrequent

2.1.2. Consonants

There are 22 consonants in Kirghiz, including b, p, f, v, d, t, g, ʁ, k, q, m, j, s, z, ʤ, ʧ, ʃ, n, ŋ,l ,r, x. Among these consonants, f, v, and x appear in foreign words only. The 22 consonants are roughly divided into 9 voiceless consonants and 13 voiced consonants, according to whether the vocal cord vibrates or not. They can also be divided into six classes according to the manner of articulation: Stop, Affricate, Fricative, Approximant, Trill, and Nasal. Finally, according to the place of articulaton, the consonants can be divided to five classes: Labial, Dental, Palatal, Velar and Uvular. Table 3 shows the categories of all the consonants.

Table 3. The consonants of Kirghiz.

		Labial	**Dental**	**Palatal**	**Velar**	**Uvular**
Stop	voiceless	p	t		k	q
	voiced	b	d		g	
Affricates	voiceless			ʧ		
	voiced			ʤ		
Fricative	voiceless	f	s	ʃ		x
	voiced	v	z			ʁ
Approximant			l	j		
Trill			r			
Nasel		m	n		ŋ	

The distribution of consonants in Kirghiz has the following characteristics: (1) The consonants r, l, ʁ, j, ŋ do not appear at the beginning of the word. (2) The consonants b, d, g, ʤ, ʁ do not appear at the end of a word. (3) j does not usually appear at the beginning of words and becomes ʤ at the beginning. ll is not allowed. (4) k and g only cooccur with front vowels (i, e, ø, y), and q and ʁ only cooccur with back vowels (ɑ, ə, o, u). However, these constrains do not apply to words borrowed from another language.

2.1.3. Vowel Harmony

An important phonetic feature of Altaic languages is phonetic harmony, which refers to the coordination and constraint of vowels within a component or between two components in a word. For instance, vowels within the same syllable or between two consecutive syllables should be 'consistent', meaning that they should share the same tongue position (front or back) and lip shape (round or unround) [10]. Compared to other Altaic languages, vowel harmony in Kirghiz is maintained rather strictly. Table 4 shows some vowel combinations allowed by the harmony rule.

Note that the harmony rules could be broken in foreign words borrowed from other languages (e.g., Russian, Chinese, and Arabic), as the pronunciation of these words in their original languages is often retained. To mention some examples: partʧa (party), santmeter (centimeter), tonna (ton), masele (problem) [11].

Table 4. The combinations of vowels in harmonic disyllabic words in Kirghiz.

Initial-Syllable Vowel	The Vowel on the Second Syllables	Examples (Disyllabic Words)
e	e, i, ee	ene, bejit, keʧee
i	i, e	ijin, tiken
ɑ	ɑ, ə, ɑɑ	ɑtɑ, ɑqən, ʃɑlbɑɑ
ə	ə, ɑ, ɑɑ	qəməz, əlɑj, ʃənɑɑ
o	o, u, oo, uu	qojun, otuz, qoroo, qoŋʃuu
u	u, o, uu, oo	ourut, suro, uruu, suroo
y	y, ø, yy, øø	bygyn, bygø, kyʧyy, yrøøn
ø	ø, y, øø	øpkø, øqyz, bøgøøn
ee	e, i	meelej, teerlik
ɑɑ	a, ə	aara, aalam
oo	o, u	ʤooluq, boordoʃ
uu	u	uuru
øø	ø	ʧøølmøk

2.1.4. Consonant Harmony

Consonant harmony means that a word ending with a voiceless/voiced consonant can be augmented by a suffix only if the beginning consonant of the suffix is voiceless/voiced [10]. Examples of consonant harmony between Kirghiz stem and suffixes are given in Table 5.

Table 5. Consonant harmony in Kirghiz stem and suffixes.

	Stem Foneme	Stem	Affix
voiceless	k-	terek	te
	q-	qonoq	tor
	s-	beʃes	ke
	p-	kitøp	tør

Table 5. *Cont.*

	Stem Foneme	Stem	Affix
voiced	l-	koŋyl	dyn
	n-	qulun	ka
	r-	ømyr	dø
	m-	tam	dar

2.1.5. Other Phonetic Phenomena

Kirghiz also involves some subtle phonological variation when suffixes are appended to a word. For instance, some consonants are weakened or eliminated, vowel's pronunciation is changed (vowel reduction), new consonants are inserted. All these phenomena may cause additional complexity when building speech recognition systems.

2.2. Alphabets

The Kirghiz language in Xinjiang China uses Kirghiz letters based on the Arabic alphabet, established in 1954 [12]. With the wide spread of online communication tools (e.g., WeChat), the young generation uses Latin characters for easier input. Table 6 presents the Arabic letters and the corresponding Latin letters.

Table 6. Arabic and Latin alphabet for Kirghiz.

Latin	Arabic	Latin	Arabic	Latin	Arabic	Latin	Arabic	Latin	Arabic
a	ا	O	ۅ	b	ب	n	ن	N	ڭ
A	ە	U	ۉ	G	ع	t	ت	d	د
e	ى	x	ش	p	پ	q	ق	y	ي
i	ئ	s	س	c	چ	f	ف	w	ۋ
o	و	m	م	H	ح	k	ك	z	ز
u	ۇ	l	ل	j	ج	g	گ	r	ر

2.3. Morphology and Syntax

Kirghiz is an agglutinative language, where a word is formed by augmenting multiple suffixes to a root. Combined with compounding and inflections, this leads to millions of different but infrequent word forms [12].

Each component (root or suffix) involves one or several syllables.

Table 7 shows several syllable structures in Kirghiz, where V represents vowel, and C represents consonant.

Table 7. Syllable structure of Kirghiz phonetic sections.

Syllable Structure	Examples	Syllable Structure	Examples
v	uu-master	cvc	sol-hand
vc	ot-fire	vcc	alp-giants
cv	bee-mare	cvcc	qart-ancient

From the perspective of linguistic roles, the words can be divided to 11 categories: nouns, pronouns, adjectives, numerals, verbs, adverbs, conjunctions, postpositions, interjections, onomatopoeia, and auxiliaries. Nouns involve variations on individual, number, and case. Adjectives and some adverbs have gradation. Pronouns can be divided into personal pronouns, demonstrative pronouns, and so on. Numerals change according to case. Verbs are divided into transitive and intransitive verbs and can vary in tense, form (imperative, suppose, wish), person, and aspect.

Like other Altaic languages, Kirghiz has a basic Subject–Object–Verb form when constructing sentences, with attributes and adverbs modifiers. Plural agreement between subject and predicate is observed more often than in other Altaic languages. Conjunctions are few, and their use is limited.

The agglutinative nature of Kirghiz results in a major difficulty when building speech recognition systems. The large vocabulary requires a large amount of text data to train the language model, and Out-Of-Vocabulary (OOV) words are often encountered.

3. Text Data Collection

With knowledge of the language, we start to construct the database. The first step is to select an appropriate set of sentences. Because Kirghiz has long vowels and is constructed by strict vowel harmony rules, its vocabulary has rich affix variants. Choosing sentences covering the phonological variation as much as possible is important to construct a high-value database. We collected the raw data from several Kirghiz websites using an Internet crawler, and the topics cover news, stories, and articles.

3.1. Text Normalization

The raw data obtained from the Internet contains lots of noise, such as images, links, and HTML tokens. These noise content should be detected and removed. The cleaned text materials are further segmented into sentences, based on period marks.

Next, all punctuation marks and special characters were removed, and sentences with foreign graphemes were discarded. Moreover, the sentences too long and too short were removed, to ensure the selected sentences are 6–18 words in length. Finally, we manually checked the collected text. Ungrammatical sentences were removed, and spelling errors were corrected. The final text collection after the above normalization process involves about 10,000 sentences.

3.2. Prompt Design and Generation

With the selected text collection, we chose a subset of the sentences to design prompts. The goal was to let the selected sentences cover phonological variations as many as possible.

In this paper, we chose triphone coverage as the selection criterion. There are 36 phonemes in Kirghiz, and theoretically, there are 36*36*36 = 46,656 (not including the silent phonemes in pronunciation) triphones. After considering the syllable structure, there are 20,328 valid triphones. We designed a greedy algorithm to perform the selection [13], including the following steps:

1. Initialize the global variables: the set of all sentences S, the set of units to be covered T, and the selected corpus set U. Set U to be empty;
2. Score sentences in S as the improvement on triphone coverage if it is selected into the set U. Reorder the sentences in S according to this score;
3. Put the sentence with the highest triphone coverage improvement to the set U, and the units covered by this sentence are removed from T;
4. Repeat steps 2 and 3 until S or T becomes the empty set.

Once the selection process is completed, the set U is used as the set of prompt sentences. In our study, 3880 sentences were finally selected.

Table 8 shows the triphone coverage at different stages of the selection process. It can be seen that the selected 3880 sentences covered 58% triphone classes, just a little drop compared to the original 10,000 sentences. Figure 1 shows the phone balance before and after the selection process.

Table 8. Comparison of triphone coverage at different stages in the selection process.

	Original (10,000)	**Selected (3880)**
Triphone tokens	856275	218,710
Triphone classes	12,336	11,776
Triphone coverage %	61%	58%

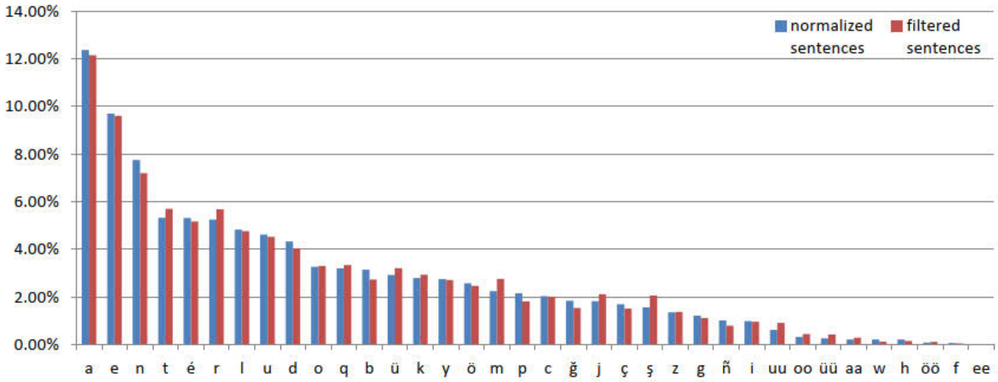

Figure 1. The phone distribution before (normalized sentences) and after (filtered sentences) the selection process.

3.3. Lexicon Generation

Lexicon maps the orthographic representation of a word to a phonemic sequence [14].

Kirghiz is a phonemic language, meaning that the written form and the pronunciation precisely match. Hence, if a list of words have been designed, creating the lexicon is trivial. On the other hand, Kirghiz is an agglutinative language, and a stem can be augmented by a large number of suffixes. This leads to an almost endless word list, and one must select the most useful ones. We performed the word selection based on frequency and selected 45,815 unique words to construct the lexicon.

4. Speech Data Collection

4.1. Speaker Selection

The speakers who participated in the M2ASR-Kirghiz recording were all Kirghiz students from Xinjiang University who came from various regions of Xinjiang without accents. The speech corpus consists of 58,097 audio files collected from 163 Kirghiz native speakers, and there are 63% males and 37% females. The ages of the speakers range from 19 to 25, and the average age is 23. Table 9 shows more details about the gender and age distribution.

Table 9. Speaker age and gender information.

Age Range	**# Speakers**	**Male**	**Female**
19–25	163	100	63

4.2. Speech Recording

All the speech data are recorded using an audio recording application that is capable of running on personal computers and smartphones. In order to use the application, the speaker needs to create an account and fill in some personal information such as name, gender, and age.

Before recording, the speaker was given a short description of the recording system. There was no constrained on the manner and speed, only if the reading was correct.

The recording software randomly sampled one sentence at a time, and when a sentence was finished, the reader could listen to the recorded speech and choose whether to accept it or record again. Quality check was performed in a random way. The checker randomly listened to some recordings and judge there were obvious mispronunciations or harsh noises. If the quality was low, the speaker was notified to record again.

4.3. Details of Speech Corpus

The details of M2ASR-Kirghiz are reported in Table 10. In total, it consists of 46,675 utterances recorded from 163 speakers, and the total duration is around 128 h. The corpus is split into training and test sets, as follows. The training set contains 41,071 utterances from 151 speakers. The dev set contains 1334 utterances from 3 speakers, and the test set contains 4270 utterances from 12 speakers.

Table 10. Data specification of M2ASR-Kirghiz database.

	Train	Test	dev	Total
# Hours	110	14	4	128
# Utterances	41,071	4,270	1334	46675
# Words	617,709	74,511	20,589	712,809
# Speakers	148	12	3	163

5. Baseline Systems

We constructed two Kirghiz ASR baseline systems, one using the Kaldi toolkit [15], and another using the WeNet toolkit [16]. The two baselines represent the conventional DNN-HMM architecture and the modern end-to-end architecture, respectively.

5.1. Kaldi Settings

The first baseline system was developed using the Kaldi egs/aishell/s5 nnet3 recipe (https://github.com/kaldi-asr/kaldi, accessed on 8 October 2022). First, a GMM-HMM model was built using 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) as features, including standard 13-dimensional MFCCs and their first and second derivatives. After that, more powerful GMM models were constructed using advanced techniques applied to features, including cepstral mean and variance normalization (CMVN), linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and speaker adaptive training (SAT) based on feature-space Maximum Likelihood Linear Regression (fMLLR).

The GMM-HMM model with fMLLR was used to produce phone alignment for all the training speech, based on which a DNN-HMM hybrid system was constructed. The structure of the DNN model was time-delay deep neural network with factorization (TDNN-F) [17]. The input features were 40-dimensional Fbanks with a symmetric three-frame window to splice neighboring frames. The neural net contained 13 hidden layers and the activation function was ReLU. The size of the hidden layers was 2048. The output layer was composed of 3725 units, corresponding to the triphone units in the GMM-HMM system.

We employed the natural stochastic gradient descent algorithm (NSGD) [18] as the optimization algorithm, and LM-MMI as the cost function with a cross-entropy regularizer to train the TDNN-F model. More details about the training procedure can be found in the recipe published accompanied with the database.

5.2. WeNet Settings

We built two end-to-end ASR baseline systems, based on Transformer [19] and Conformer [20], respectively. We used the WeNet toolkit (https://github.com/wenet-e2e/wenet, accessed on 8 October 2022) in our experiment, and followed the aishell/s0 recipe. Default parameters were adopted, except that the batch size was reduced to 12 to meet the capacity of the hardware.

Specifically, the Transformer model involved 12 transformer layers in the encoder and six transformer layers in the decoder. We used Adam optimizer to train the model, with learning rate set to 0.002. There were 25,000 warm-up steps.

For the Conformer baseline, we only used conformer layers in the encoder, and the decoder was the same as the Transformer baseline. The advantage of Conformer compared to Transformer is that an additional CNN layer may help collect information from local context.

Finally, considering the recent success of self-supervised training methods on low-resource ASR [21–23], we built another Conformer-based system via a pre-training and fine-tuning strategy. We firstly adopted a pre-trained Conformer model *ch-w2v-conformer* (https://huggingface.co/emiyasstar/ch-w2v-conformer, accessed on 8 October 2022) as the starting point, and then fine-tuned the model with M2ASR-Kirghiz.

5.3. Experimental Results

The performance of the GMM-HMM, TDNN-HMM, Transformer, Conformer, and pre-training + fine-tuning Conformer (Conformer + pre-train) systems is presented in Table 11. Considering the agglutinative character of Kirghiz, we use letter error rate (LER) as the metric rather than word error rate (WER).

It can be observed that the LERs of the Conformer, Transformer, TDNN-HMM, and GMM-HMM are gradually increased, and the best result was achieved by the Conformer + pre-train system. It is worth mentioning that the data used to train *ch-w2v-conformer* contain some Altaic languages, such as Kazakh and Turkish, which are quite similar to Kirghiz. This is probably a reason explaining that the Conformer + pre-train system obtained such good performance.

Table 11. LER(%) results obtained with various ASR baseline systems.

System	LER%
GMM-HMM	29.28%
TDNN-HMM	14.59%
Transformer	8.06%
Conformer	7.82%
Conformer + pre-train	4.12%

6. Conclusions and Future Work

In this paper, we presented our recent effort on Kirghiz data resource construction. We collected a text corpus of 2.6 M tokens and built a speech corpus called M2ASR-Kirghiz that contains 128 h of speech data from 163 speakers. Several Kirghiz ASR baselines were constructed using the new resource, and the best performance obtained so far is 4.12% in LER. We hope that this new database encourages and makes it easier for researchers to start their research in Kirghiz.

Counting M2ASR-Kirghiz in, the M2ASR project has published datasets for five languages: Tibetan, Mongolia, Uyghur, Kazakh, and Kirghiz. These languages share similarities but also possess their own particularities. It would be highly interesting to utilize the similarities to boost the performance for each other and use the particularities to obtain more room for improvement.

Author Contributions: Conceptualization, I.M.,W.D. and A.H.; methodology, I.M.; software, W.D.; validation, I.M.,W.D. and A.H.; formal analysis, I.M.; investigation, W.D. ; resources, I.M. and W.D.; data curation, I.M. and W.D.; writing—original draft preparation, I.M.; writing—review and editing, I.M. and W.D.; visualization, W.D.; supervision, W.D. and A.H.; project administration, A.H.; funding acquisition, I.M.,W.D. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strengthening Plan of National Defense Science and Technology Foundation of China (2021-JCJQ-JJ-0059) and the Natural Science Foundation of China (U2003207,62171250,61633013).

Data Availability Statement: The data that support the findings of this study are openly available for the research purpose at http://m2asr.cslt.org/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, D.; Zheng, T.F.; Tang, Z.; Shi, Y.; Li, L.; Zhang, S.; Yu, H.; Li, G.; Xu, S.; Hamdulla, A.; et al. M2ASR: Ambitions and first year progress. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–6.
2. Li, G.; Yu, H.; Zheng, T.F.; Yan, J.; Xu, S. Free linguistic and speech resources for Tibetan. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 733–736.
3. Zhi, T.; Shi, Y.; Du, W.; Li, G.; Wang, D. M2ASR-MONGO: A Free Mongolian Speech Database and Accompanied Baselines. In Proceedings of the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore, 18–20 November 2021; pp. 140–145.
4. Rouzi, A.; Dong, W.; Zhiyong, Z.; Fang, Z. An open/free database and benchmark for Uyghur speaker recognition. In Proceedings of the 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). IEEE, Shanghai, China, 28–30 October 2015; pp. 81–85.
5. Shi, Y.; Hamdullah, A.; Tang, Z.; Wang, D.; Zheng, T.F. A free Kazakh speech database and a speech recognition baseline. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 745–748.
6. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
7. Yishake, C. The phonetic differences between Kirghiz and Uyghur and their causes. *J. Hotan Norm.* **2015**, *2*, 69–74.
8. Washington, J.N. Root vowels and affix vowels: Height effects in Kyrgyz vowel harmony. 2006, *unpublished*.
9. Tenishev, E.R. *Introduction to the Study of Turkic Languages*; China Social Science Press: Beijing, China, 1981.
10. Yishake, C.; Maituohuo, M.; Idris Abdu, H.; Tsakuwa, M.B. Long vowels in Kyrgyz language: Characteristics and evolutionary steps. *J. Lang. Linguist. Stud.* **2022**, *18*, 294–308.
11. Hu, Z. *A Brief History of the Kirghiz Language*; The Ethnic Publishing House: Beijing, China, 1986; Volume 1.
12. Hu, Z. *Studies on Kirghiz Language and Culture*; China Minzu University Press: Beijing, China, 2006.
13. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, USA, 2022.
14. Biadsy, F.; Habash, N.; Hirschberg, J. Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 31 May–5 June 2009; pp. 397–405.
15. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
16. Zhang, B.; Wu, D.; Yang, C.; Chen, X.; Peng, Z.; Wang, X.; Yao, Z.; Wang, X.; Yu, F.; Xie, L.; et al. Wenet: Production first and production ready end-to-end speech recognition toolkit. *arXiv* **2021**, arXiv:2102.01547.
17. Povey, D.; Cheng, G.; Wang, Y.; Li, K.; Xu, H.; Yarmohammadi, M.; Khudanpur, S. Semi-orthogonal low-rank matrix factorization for deep neural networks. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3743–3747.
18. Povey, D.; Zhang, X.; Khudanpur, S. Parallel training of DNNs with natural gradient and parameter averaging. *arXiv* **2014**, arXiv:1410.7455.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Kaiser, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. pp. 6000–6010.
20. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
21. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv* **2020**, arXiv:2006.13979.
22. Javed, T.; Doddapaneni, S.; Raman, A.; Bhogale, K.S.; Ramesh, G.; Kunchukuttan, A.; Kumar, P.; Khapra, M.M. Towards building ASR systems for the next billion users. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 2022; Volume 36, pp. 10813–10821.
23. Zhang, Z.Q.; Song, Y.; Wu, M.H.; Fang, X.; McLoughlin, I.; Dai, L.R. Cross-Lingual Self-training to Learn Multilingual Representation for Low-Resource Speech Recognition. *Circuits Syst. Signal Process.* **2022**, *41*, 6827–6843. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.