information

Article

Multilingual Speech Recognition for Turkic Languages

Saida Mussakhojayeva, Kaisar Dauletbek, Rustem Yeshpanov and Huseyin Atakan Varol *

Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Astana 010000, Kazakhstan
* Correspondence: ahvarol@nu.edu.kz

Abstract: The primary aim of this study was to contribute to the development of multilingual automatic speech recognition for lower-resourced Turkic languages. Ten languages—Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek—were considered. A total of 22 models were developed (13 monolingual and 9 multilingual). The multilingual models that were trained using joint speech data performed more robustly than the baseline monolingual models, with the best model achieving an average character and word error rate reduction of 56.7%/54.3%, respectively. The results of the experiment showed that character and word error rate reduction was more likely when multilingual models were trained with data from related Turkic languages than when they were developed using data from unrelated, non-Turkic languages, such as English and Russian. The study also presented an open-source Turkish speech corpus. The corpus contains 218.2 h of transcribed speech with 186,171 utterances and is the largest publicly available Turkish dataset of its kind. The datasets and codes used to train the models are available for download from our GitHub page.

Keywords: automatic speech recognition; multilingual speech recognition; Turkic languages; transfer learning; Common Voice; big data; lower-resourced languages; Kazakh Speech Corpus; Uzbek Speech Corpus; Turkish Speech Corpus

Citation: Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. *Information* **2023**, *14*, 74. https://doi.org/10.3390/info14020074

Academic Editor: Diego Reforgiato Recupero

Received: 30 November 2022
Revised: 13 January 2023
Accepted: 23 January 2023
Published: 28 January 2023

1. Introduction

The task of automatic speech recognition (ASR) refers to converting any acoustic signal containing human speech into the corresponding word sequence [1]. The development of the graphics processing units (GPUs) and deep neural networks (DNNs) [2], the availability of transcribed speech corpora in the public domain [3–5], and the wide use of voice interaction services that support hundreds of languages (e.g., Alexa, Google Voice Assistant, and Siri) have led to ASR solutions achieving—and even exceeding—human performance [6]. That said, most ASR efforts have been directed towards developing models for languages for which large corpora exist (i.e., higher-resourced languages (The terms *lower-* and *higher-resourced languages* are used throughout the paper to emphasize the continuum existing across languages in terms of resources available for speech technology development.)), such as English, Mandarin, and Japanese (see, e.g., [4,7,8]). In turn, ASR models built for lower-resourced languages can rarely boast robustness and reliability due to the insufficient amount of training data.

To address the problem of the inadequacy of training data for lower-resourced languages, such techniques as transfer learning [9], data augmentation [10], and high resource transliteration [11], to name the most notable few, have been proposed. Special attention has also been devoted to the development of multilingual models, which enables the use of common linguistic features across languages, thus alleviating challenging data requirements [12]. Most work on multilingual ASR for lower-resourced languages focuses on combining the data of similar languages and performing cross-language optimization, by utilizing positive transfer from higher-resourced languages during training [13–15]. Research in transfer learning, too, has shown that linguistic similarity and relatedness generally lead to improved robustness of ASR models, particularly in resource-constrained

settings [16]. For example, linguistic relatedness and similarities have been made use of to build multilingual ASR models for lower-resourced Indian [17,18] and Ethiopian [19] languages and Arabic dialects [20]. The use of unrelated languages, however, generally results in a trade-off between quality and quantity, with models yielding performance comparable only with those of monolingual models [21] and with no significant improvement due to minimal linguistic overlap.

This work aims to make a contribution to the development of multilingual ASR for lower-resourced Turkic languages. To date, there have been studies conducted to develop multilingual models recognizing Turkic languages [21,22], but few Turkic languages were considered in the models or were recognized along with languages belonging to other language families (e.g., English, Persian, Russian, Swahili, etc.). In contrast, in this study, we exclusively focus on ten Turkic languages—namely, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek.

According to various sources, the ten languages under consideration are at present spoken by 125–150 million speakers [23,24]. Spread over the vast area of Eurasia, the languages fall into several branches (see Table 1). With the exception of Chuvash and Sakha, which have peculiarities stemming from the early detachment from Common Turkic of the former and the influence of the Tungusic languages on the latter [23], the languages are, on the whole, remarkably similar in terms of lexis, phonology, and morphology. This is reflected in a certain degree of mutual intelligibility across the languages, with some of the most frequent words in the Turkic languages being exactly alike [25]. We therefore hypothesize that utilizing such features common for the ten languages is more likely to result in a robust multilingual ASR model than when unrelated languages are used, with some of the lower-resourced Turkic languages (e.g., Azerbaijani, Chuvash, and Sakha) benefiting from other Turkic languages for which more training resources are available (e.g., Bashkir, Kazakh, and Uzbek).

Table 1. The languages and datasets used in the study.

Language	Code	Family	Branch	Script	Corpus	Validated Length (hr)	Utterances
Azerbaijani	az	Turkic	Oghuz	Latin	CVC	0.13	81
Bashkir	ba	Turkic	Kipchak	Cyrillic	CVC	232.37	189,970
Chuvash	ch	Turkic	Oghur	Cyrillic	CVC	11.90	8651
Kyrgyz	ky	Turkic	Kipchak	Cyrillic	CVC	18.58	14,599
Sakha	sa	Turkic	Siberian Turkic	Cyrillic	CVC	6.61	3975
Tatar	tt	Turkic	Kipchak	Cyrillic	CVC	25.07	24,109
Uyghur	ug	Turkic	Karluk	Arabic	CVC	35.61	21,282
Kazakh	kk	Turkic	Kipchak	Cyrillic	CVC KSC	1.60 332.60	1169 153,853
Turkish	tr	Turkic	Oghuz	Latin	CVC TSC	51.46 218.24	51,710 186,171
Uzbek	uz	Turkic	Karluk	Latin	CVC USC	94.24 104.90	77,220 108,387
English	en	Indo-European	West Germanic	Latin	CVC	344.74	217,968
Russian	ru	Indo-European	Slavic	Cyrillic	OpenSTT	338.30	235,148

To contribute to the development of multilingual ASR for Turkic languages,

1. We compare the results of multilingual models trained on the data of the ten Turkic languages with the results of monolingual models trained for each of the languages;
2. We compare the results of the multilingual models with the results of models trained on the data of the ten Turkic languages and two non-Turkic languages (English and Russian);

3. We create the largest open-source speech corpus for the Turkish language that contains 218.2 h of transcribed speech.

The remainder of the paper is organized as follows: Section 2 provides an overview of existing work on multilingual ASR, focusing on both related and unrelated languages. In Section 3, we provide a description of the datasets used in the study and the procedures adopted to pre-process and split the data, as well as the details of the experimental setup. Section 4 describes the results obtained and a discussion of these results. Section 5 concludes the paper.

2. Related Work

The proliferation of studies in the field of ASR in recent years can be attributed to several factors, including a reduction in training time thanks to the use of the GPUs in deep learning [2], publicly available datasets (e.g., LibriSpeech [4] and DiDiSpeech [7]), and regular speech recognition competitions (e.g., CHiME-6 Challenge [26]). Demonstrating a significant performance boost [27–31], reporting a word error rate (WER) as low as 2–3% on popular datasets [32], and even achieving human parity [6], ASR research may create the false impression that the task is almost solved. However, the vast majority of the research focuses on mainstream languages for which extensive resources (e.g., recorded speech and human-labeled speech corpora) are available. For example, the whole Corpus of Spontaneous Japanese [8] contains a speech signal of about 661 h; the DiDiSpeech corpus of Mandarin [7] and the LibriSpeech corpus of read English speech [4] consist of about 800 and 1000 h of data, respectively. Consequently, languages that suffer from lower data availability can hardly afford the development of high-quality ASR systems.

One of the proposed ways to get around this problem is the application of transfer-learning techniques [33]. Even though the original idea explores reusing the weights of a previously trained DNN for a new task, it can be extrapolated to the problem of data insufficiency. In [9], the use of transfer learning in adapting a neural network originally trained for English ASR to German resulted in faster training, lower resource requirements, and reduced costs. Some studies propose similar methods, where the core idea is to train an ASR model jointly on multiple languages with the expectation that the system will perform better than systems trained on a single specific language. This approach is commonly referred to as multilingual ASR [34].

The earlier experiments with multilingual ASR [35,36] mostly explored the cases with only a few languages at a time and did not produce meaningful results except in language identification (LID) tasks. Language identifiers (IDs) are used as an additional input signal when multiple languages are involved, proving to be useful in both code-switching [37,38] and multilingual ASR [12,39]. There are two common ways to incorporate language IDs: (1) using special LID tokens at the beginning of output [37,38], thus using one-hot vector representation as an additional feature [12], or (2) using an auxiliary classifier in a multi-task setting [39].

Some recent advances in multilingual ASR assume that the presence of higher-resourced languages in the training set positively affects the performance of a model for lower-resourced languages [14,15,17–20,40]. In [14], the scholars showed that it is possible to train a massive single ASR architecture for 51 different languages and more than 16,000 h of speech across them, which, in practice, is significantly less time-consuming to tune than developing 51 individual monolingual baselines. It was also reported that training ASR multilingual models can improve recognition performance for all the languages involved, with the lower-resource languages observing a more significant reduction of WER and character error rate (CER) for East Asian languages. In another study [19] exploring ASR for lower-resourced languages, multilingual systems for four Ethiopian languages were developed. One of the models trained with speech data from 22 languages other than the target languages achieved a WER of 15.79%. Furthermore, the inclusion of the speech of a closely related language (in terms of phonetic overlap) in multilingual model training resulted in a relative WER reduction of 51.41%.

Most of the studies on multilingual ASR conclude that the average increase in performance produced by multilingual models, as opposed to monolingual ones, is higher for languages with greater linguistic overlap. Moreover, the development of a unified end-to-end (E2E) solution for a large number of languages that can potentially outperform monolingual models has become one of the focal points of multilingual ASR. However, research consistently shows that a model trained on a random set of languages does not consistently outperform monolingual models, even at a very large scale where more than 40 languages are used in the training set [12,14,15]. The authors of [14] have demonstrated that this is the case for higher-resourced languages, as the multilingual model failed to beat the baseline WER and CER scores for all higher-resourced settings.

This has led to the realization that using a dataset of languages with high linguistic overlap between them might yield better results. One of the ways to select these languages is to draw upon the language families to which they belong, as it is clear that the linguistic overlap between these languages is much greater than for languages with no inherent linguistic connections [19]. As a result, several recent studies into multilingual ASR have been carried out at the level of language families [17–20,40].

The authors of [18,20] developed E2E ASR systems for Indian and Arabic languages, respectively. Both papers report on average performance improvements over monolingual models, but were still unsuccessful in outperforming them in several languages. The findings were also consistent in the case of Ethiopian languages [19], where the scholars were able to obtain comparable results without having a target language in the training set. It is also important to note that the quality of training data may hinder the transfer learning capacity of the model, as was shown in [17]. The scholars were not able to achieve a significant improvement over monolingual experiments while using a dataset that contained systematic linguistic errors.

Most of the Turkic languages in our study are lower-resourced with few studies and datasets available. As can be seen from Table 1, these languages can be divided into five branches. Apart from Chuvash and Sakha, each belonging to a distinct subfamily, there are three major branches: Karluk, Kipchak, and Oghuz. To the best of our knowledge, while there are large open-source corpora for some of the languages belonging to the Karluk and Kipchak branches (e.g., the Bashkir set in Common Voice Corpus 10.0 (CVC) [3], Kazakh Speech Corpus (KSC) [41], and Uzbek Speech Corpus (USC) [42]), there are no similar or sufficiently large publicly available datasets for most of the languages under consideration. For example, in [43], a high-accuracy Tatar speech recognition system was trained on a proprietary dataset and the Tatar portion of CVC. Specifically, the model was trained on 328 h of unlabeled data and then finetuned on 129 h of annotated data, achieving a WER of 5.37% on the CVC test set. It should be noted that in this work, the ASR model was trained on a full Tatar CVC training set (28 h), which has 100% text overlap with the corresponding test set. Similarly, the authors of [44] developed an Uzbek ASR system trained on the Uzbek CVC (127 h) and Speechocean (https://en.speechocean.com/datacenter/details/1907.html (accessed on 22 January 2023)) (80 h) datasets and obtained a CER score of 5.41% on the Uzbek CVC test split. However, it is unclear whether the authors used part of the invalidated Uzbek CVC for training purposes, nor does the paper make mention of utterance overlap. In [45], different language models and acoustic training methodologies for the Azerbaijani language were investigated. Speech data of 80 h were collected from emergency calls. However, the data remain confidential, as they contain sensitive information about emergency cases.

As for the Turkish language, the corpus prepared by the Middle East Technical University (METU) [46,47] contains speech from 193 speakers (89 female and 104 male). Each speaker read 40 sentences that were selected randomly from a 2462-sentence set. Another Turkish speech corpus, containing broadcast news, was developed by Boğaziçi University [48] and has a total length of 194 h. The largest Turkish dataset [49] contains 350.27 h of validated speech. However, the data, which come from films and crowdsourcing, are

not publicly available. A detailed comparison between the existing Turkish ASR corpora and the Turkish Speech Corpus (TSC) can be found in Table 2.

Table 2. Turkish ASR datasets.

Corpus	Length (hr)	Utterances	Open-Source
METU [46,47]	5.6	N/A	-
Boğaziçi [48]	194	N/A	-
HS [49]	350.27	565,073	-
CVC 10.0 [3]	76	74,487	+
TSC (ours)	218.24	186,171	+

3. Materials and Methods

3.1. Data

3.1.1. Datasets

To build a multilingual dataset, we considered a total of 12 languages. The ten Turkic languages were the target languages. English and Russian were the control languages. Detailed information on the languages utilized in this work is given in Table 1. As regards the datasets, we used multiple sources of transcribed speech, including the CVC [3], the Russian Open Speech To Text Dataset (OpenSTT) (https://github.com/snakers4/open_stt (accessed on 22 January 2023)), the KSC, the USC, and a new TSC.

The choice of the CVC as the main component of the multilingual dataset was due to the fact that it is one of the largest publicly available multilingual datasets designed for ASR purposes, comprising transcribed speech for 98 languages, including the target languages. The KSC is a large open-source corpus for Kazakh ASR, containing approximately 332 h of transcribed speech data comprising more than 153,000 utterances. The USC is the first open-source Uzbek speech dataset and comprises a total of 105 h of transcribed audio recordings by 958 different speakers.

With respect to the TSC, to the best of our knowledge, it is the largest Turkish speech corpus in the public domain. The data were collected from open sources and embraced various domains, such as news, interviews, talk shows, and documentaries. To acquire audio recordings, we used a command-line program to download videos from YouTube called *youtube-dl* (https://github.com/ytdl-org/youtube-dl (accessed on 22 January 2023)). . The resulting audio files were transcribed by a team of language specialists to ensure quality and accuracy. The TSC contains a total of 186,171 utterances, which adds up to 218.2 h of recorded speech. The total number of unique words is 122,319.

The data for the control languages, utilized to evaluate the relative performance of the multilingual models, came from the English and Russian subsets of the CVC and the OpenSTT, respectively, used in [21]. Specifically, the English data were 344 h in length and consisted of validated recordings that received the highest number of upvotes (i.e., instances of verification of correctness) from contributors. For evaluation purposes, we randomly extracted seven-hour subsets from the original CVC validation and test sets. For Russian, we used 338 h of speech, with associated transcripts previously corrected by native Russian speakers to ensure accuracy. The data embrace the domains of books and YouTube. A seven-hour subset of the data was selected for the development set. For the test set, we used the official validation sets of OpenSTT from both domains. Detailed statistics for each data source used in the study are given in Tables 1 and 3.

Table 3. Statistics for the training (Train), development (Dev), and test (Test) sets of CVC, OpenSTT , KSC, TSC, and USC.

Language	Corpus	Length (hr)			Utterances		
		Train	Dev	Test	Train	Dev	Test
az	CVC	0.05	0.04	0.04	39	20	22
ba	CVC	193.15	19.39	19.83	160,885	14,559	14,526
ch	CVC	8.47	1.54	1.89	6244	1140	1267
ky	CVC	14.25	2.14	2.19	11,373	1613	1613
sa	CVC	2.72	1.67	2.22	1643	1083	1249
tt	CVC	16.28	3.05	5.74	15,928	3062	5119
ug	CVC	26.27	4.54	4.80	15,787	2748	2747
kk	CVC	0.57	0.49	0.54	406	379	384
	KSC	318.40	7.13	7.07	147,326	3283	3334
tr	CVC	31.83	9.23	10.40	33,491	9095	9124
	TSC	209.6	4.26	4.38	179,259	3428	3484
uz	CVC	61.50	14.91	17.83	53,409	11,569	12,242
	USC	96.40	4.00	4.50	100,767	3783	3837
en	CVC	330	7.4	7.3	208,976	4346	4646
ru	OpenSTT	324.3	7.0	7.0	222,643	4776	7729

3.1.2. Data Pre-Processing

For each language, audio scripts were lowercased; character encodings were normalized, and punctuation marks were filtered out. For all the scripts (Arabic, Cyrillic, and Latin), we used the character-level encoding. We filtered the training data for each CVC dataset to keep the overlap between training and evaluation (development and test) sets below 40% to avoid memorization by the language models. The filtered utterances, as well as the script used to remove punctuation marks, can be found in our GitHub repository (https://github.com/IS2AI/TurkicASR (accessed on 22 January 2023)) for reproducibility purposes.

Four special tokens (i.e., $<blank>$, $<unk>$, $<space>$, and $<sos/eos>$) were used. One of the twelve language IDs ($[az]$, $[ba]$, $[ch]$, $[en]$, $[kk]$, $[ky]$, $[ru]$, $[sa]$, $[tr]$, $[tt]$, $[ug]$, and $[uz]$) was prepended to each utterance. The character set size equaled 137 characters. Table 4 lists the 124 letters and symbols currently used in the alphabets of the languages under consideration. The remaining 13 characters (i.e., ئ, ا, ه, و, ۆ, ۇ, ۋ, ى, ې, â, î, ò, ҥ) can be encountered in words recorded using outdated or previous alphabets of the languages (e.g., Uyghur), as well as in loanwords or proper nouns containing symbols that do not originate in a particular language (e.g., diacritics). For reasons of efficiency, we removed utterances with more than 256 characters and audio recordings with a duration of more than 20 s.

Table 4. Characters of the considered languages.

#	char	ug
1	ئا	+
2	ئە	+
3	ب	+
4	پ	+
5	ت	+
6	ج	+
7	چ	+
8	خ	+
9	د	+
10	ر	+
11	ز	+
12	ژ	+
13	س	+
14	ش	+
15	غ	+
16	ف	+
17	ق	+
18	ك	+
19	گ	+
20	ڭ	+
21	ل	+
22	م	+
23	ن	+
24	ھ	+
25	ئو	+
26	ئۇ	+
27	ئۆ	+
28	ئۈ	+
29	ۋ	+
30	ئې	+
31	ئى	+
32	ي	+
		32

#	char	az	tr	uz	en
1	a	+	+	+	+
2	ə	+	-	-	-
3	b	+	+	+	+
4	c	+	+	-	+
5	ç	+	+	-	-
6	d	+	+	+	+
7	e	+	+	+	+
8	f	+	+	+	+
9	g	+	+	+	+
10	ğ	+	+	-	-
11	h	+	+	+	+
12	i	+	+	+	+
13	ı	+	+	-	-
14	j	+	+	+	+
15	k	+	+	+	+
16	l	+	+	+	+
17	m	+	+	+	+
18	n	+	+	+	+
19	o	+	+	+	+
20	ö	+	+	-	-
21	p	+	+	+	+
22	q	+	-	+	+
23	r	+	+	+	+
24	s	+	+	+	+
25	ş	+	+	-	-
26	t	+	+	+	+
27	u	+	+	+	+
28	ü	+	+	-	-
29	vs.	+	+	+	+
30	w	-	-	-	+
31	x	+	-	+	+
32	y	+	+	+	+
33	z	+	+	+	+
34	oʻ	-	-	+	-
35	gʻ	-	-	+	-
36	sh	-	-	+	-
37	ch	-	-	+	-
38	ng	-	-	+	-
39	ʼ	-	-	+	-
		32	**29**	**30**	**26**

#	char	ba	ch	kk	ky	sa	tt	ru
1	а	+	+	+	+	+	+	+
2	ă	-	+	-	-	-	-	-
3	ә	+	-	+	-	-	+	-
4	б	+	+	+	+	+	+	+
5	в	+	+	+	+	+	+	+
6	г	+	+	+	+	+	+	+
7	ғ	+	-	+	-	-	-	-
8	ҕ	-	-	-	-	+	-	-
9	д	+	+	+	+	+	+	+
10	дь	-	-	-	-	+	-	-
11	е	+	+	+	+	+	+	+
12	ё	+	+	+	+	+	+	+
13	ĕ	-	+	-	-	-	-	-
14	ж	+	+	+	+	+	+	+
15	җ	-	-	-	-	-	+	-
16	з	+	+	+	+	+	+	+
17	ҙ	+	-	-	-	-	-	-
18	и	+	+	+	+	+	+	+
19	й	+	+	+	+	+	+	+
20	к	+	+	+	+	+	+	+
21	қ	+	-	-	-	-	-	-
22	ҟ	-	-	+	-	-	-	-
23	л	+	+	+	+	+	+	+
24	м	+	+	+	+	+	+	+
25	н	+	+	+	+	+	+	+
26	ң	+	-	+	+	-	+	-
27	ҥ	-	-	-	-	+	-	-
28	нь	-	-	-	-	+	-	-
29	о	+	+	+	+	+	+	+
30	ө	+	-	+	+	+	+	-
31	п	+	+	+	+	+	+	+
32	р	+	+	+	+	+	+	+
33	с	+	+	+	+	+	+	+
34	ҫ	+	+	-	-	-	-	-
35	т	+	+	+	+	+	+	+
36	у	+	+	+	+	+	+	+
37	ӳ	-	+	-	-	-	-	-
38	ҷ	-	-	+	-	-	-	-
39	ү	+	-	+	+	+	+	-
40	ф	+	+	+	+	+	+	+
41	х	+	+	+	+	+	+	+
42	һ	+	-	+	-	+	+	-
43	ц	+	+	+	+	+	+	+
44	ч	+	+	+	+	+	+	+
45	ш	+	+	+	+	+	+	+
46	щ	+	+	+	+	+	+	+
47	ъ	+	+	+	+	+	+	+
48	ы	+	+	+	+	+	+	+
49	і	-	-	+	-	-	-	-
50	ь	+	+	+	+	+	+	+
51	э	+	+	+	+	+	+	+
52	ю	+	+	+	+	+	+	+
53	я	+	+	+	+	+	+	+
		42	**37**	**42**	**36**	**40**	**39**	**33**

Note. The green and red shading indicates characters that are present in and do not belong to a specific language, respectively.

3.1.3. Data Augmentation

We applied speed perturbation [10] with factors of 0.9, 1.0, and 1.1 to the training sets. During training, we applied spectral augmentation [50] on-the-fly to the feature inputs of the encoder. Both data augmentation techniques are standard procedures regularly employed in ASR [17,20,21].

3.2. Experimental Setup

We first trained monolingual ASR models for each Turkic language on the CVC and then multilingual models. A total of 22 (13 monolingual and 9 multilingual) models were developed. A complete list of the models and the datasets on which they were trained can be found in Table 5. All of the models were trained on the training sets. Hyper-parameters were tuned using the development sets. The final models were evaluated on the test sets. Detailed information regarding the sets can be found in Table 3.

Table 5. A list of the models and the datasets used in training.

Model			Corpus														
			CVC														
Type	**#**	**Name**	**az**	**ba**	**ch**	**kk**	**ky**	**sa**	**tt**	**tr**	**ug**	**uz**	**en**	**Open STT**	**KSC**	**TSC**	**USC**
monolingual	1	az_cvc	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2	ba_cvc	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
	3	ch_cvc	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
	4	kk_cvc	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
	5	ky_cvc	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-
	6	sa_cvc	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
	7	tt_cvc	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
	8	tr_cvc	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
	9	ug_cvc	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
	10	uz_cvc	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
	11	kk_ksc	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
	12	tr_tsc	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
	13	uz_usc	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
multilingual	14	turkic	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
	15	ksc_turkic	+	+	+	+	+	+	+	+	+	+	-	-	+	-	-
	16	tsc_turkic	+	+	+	+	+	+	+	+	+	+	-	-	-	+	-
	17	usc_turkic	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+
	18	en_turkic	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-
	19	ru_turkic	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-
	20	en_ru_turkic	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
	21	all_turkic	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+
	22	all_languages	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Note. The green shading indicates the datasets used in training a specific model.

3.2.1. Acoustic Models

We trained all models in Pytorch [51] using the ESPnet framework [52] and primarily followed the procedure described in the CVC recipe [52]. ESPnet is an end-to-end neural network toolkit that is a widely used open-source standard providing a complete setup for various speech processing tasks. We followed the latest conformer architecture of the CVC recipe when developing both monolingual and multilingual models. Specifically, the recipe for monolingual models is identical to the CVC recipe, while, for multilingual models, we increased the following hyperparameters:

- attention heads: $4 \rightarrow 8$
- encoder output dimension: $256 \rightarrow 512$
- convolutional kernel size: $15 \rightarrow 31$
- number of batch bins: $10^7 \rightarrow 4 \times 10^7$

These changes accommodated the increased amount of data used in training multilingual models and helped prevent overfitting.

Monolingual and multilingual models were trained on one and four NVIDIA DGX A100 (40 GB) GPUs, respectively. The combinations of datasets used for training the multilingual models are provided in Table 5. The monolingual model names are given

in the lowercase format `language code_dataset name` (e.g., `az_cvc` for the monolingual model trained on the Azerbaijani Common Voice Corpus). The multilingual model names are given in a lowercase format where

- `turkic` refers to models whose training data included the CVC datasets for the target Turkic languages,
- `ksc`, `tsc`, and `usc` refer to models whose training data included the Kazakh, Turkish, and Uzbek Speech Corpora, respectively,
- `en` and `ru` refer to models whose training data included the English and Russian datasets,
- `all_turkic` refers to the model whose training data included all datasets for the target Turkic languages,
- `all_languages` refers to the model whose training data included all datasets in the study.

As a baseline, the monolingual models were trained for each data source using the conformer architecture [27] with 42.98×10^6 trainable parameters. The decoder consisted of 6 transformer blocks, with a dropout rate set to 0.1. The same decoder configurations were also used for the multilingual architecture. For optimization, we used the Adam optimizer [53], with an initial learning rate of 4.0 and 2.5×10^5 warm-up steps.

In total, nine multilingual models were trained. All models had the same model configurations and were trained for 60 epochs based on the conformer architecture. We used 12 conformer encoder blocks with an output dimensionality of 512, 8 attention heads [30], a convolution kernel size of 31, and a dropout rate of 0.1. To optimize the training process, we used the Adam optimizer with the initial learning rate set at 25×10^{-4} and 3×10^5 warm-up steps. The gradient clipping was set to 5 and the gradient accumulation to 4, while CTC [54] loss and label smoothing weights were set to 0.3 and 0.1, respectively. During inference, we used a beam size of 10 and set the CTC decoding weight to 0.6. The contribution weight of the language model was set to 0.3. All of the multilingual models had the same number of parameters, 108.68×10^6.

3.2.2. Language Models

For language models, we chose the transformer architecture [30]. We used sequential positional encoding, as the length of any utterance did not exceed 256 characters. Therefore, no sophisticated positional encoding methods were employed. Each of the language models had 16 transformer blocks, an embedding size of 128, with 8 attention heads, each with a dimensionality of 512. The dropout rate was set to 0.1 [55]. We trained the models for 30 epochs using the Adam optimizer. The initial learning rate was set to 0.001, with gradient clipping of 5.0 and gradient accumulation of 1. Similar to the acoustic models, we used 2.5×10^4 warm-up steps with batch bins set to 10^6.

3.2.3. Performance Evaluation

The WER and CER metrics are the most common performance measures for ASR [56–58]. Even though the WER metric is usually preferred for most of the monolingual cases, calculating errors on the character level would convey the multilingual model performance in a more precise manner. The CER/WER of the predicted sequence is computed by dividing the sum of all substitutions, insertions, and deletions by the total number of characters/words in a reference transcription. The percentage of characters/words that have been inaccurately predicted is frequently related to CER/WER. However, CER/WER can exceed 100%, particularly when there are too many insertions. For example, the CER for a reference transcription 'fan' and a longer predicted sequence 'fantastic' is 200%, which is calculated by dividing the sum of substitutions (0), insertions (6 in 'tastic'), and deletions (0) in the predicted sequence ('fantastic') by the total number of characters in the reference transcription (3 in 'fan'). The performance of an ASR system improves as CER/WER decreases, with a value of 0% denoting the ideal result. In our study, CER/WER scores were transformed into percentages and displayed as such.

4. Results and Discussion

The performance of the models on the test sets is given in Tables 6 and 7. Considering the uncomparable distribution of data across the training, development, and test sets for some of the languages for which more than one dataset was available (i.e., Kazakh, Turkish, and Uzbek), we considered it fair and reasonable to evaluate the developed multilingual models separately on the CVC and the KSC, TSC, and USC test sets. While Table 6 provides the results obtained by the models on the CVC test sets exclusively, Table 7 contains the CER and WER scores for the models evaluated on the KSC, TSC, and USC test sets only. For readability, the dashed line separates the monolingual baselines from the multilingual models, and the green shading indicates the best results.

Table 6. The CER (%) | WER (%) results, average boost (AB, %) over the monolingual baseline, and training time (TT, day) of the models on the CVC test sets.

Model	Language										AB	TT
	az	ba	ch	kk	ky	sa	tt	tr	ug	uz		
`_cvc`	107.6 | 325.7	1.7 | 5.5	15.5 | 46.2	69.9 | 101.2	13.6 | 36.7	35.3 | 82.9	13.6 | 37.9	7.3 | 20.1	6.5 | 24.0	4.2 | 14.6	-	-
`turkic`	36.5 | 91.1	2.1 | 6.1	7.0 | 22.0	39.3 | 83.2	8.7 | 21.1	18.4 | 49.9	6.7 | 19.5	6.5 | 17.1	6.0 | 15.0	4.7 | 14.4	33.5 | 34.9	0.8
`ksc_turkic`	29.9 | 81.7	1.6 | 5.0	5.7 | 18.7	12.9 | 32.5	6.3 | 16.3	16.8 | 46.1	7.3 | 23.0	5.4 | 15.1	5.7 | 13.7	4.8 | 14.1	43.7 | 44.7	1.5
`tsc_turkic`	30.7 | 83.8	1.6 | 5.3	6.2 | 20.4	34.3 | 74.2	6.3 | 17.0	17.0 | 48.0	6.8 | 22.0	3.5 | 9.5	4.3 | 12.1	4.3 | 13.4	46.5 | 43.5	1.2
`usc_turkic`	34.0 | 86.4	1.8 | 5.8	6.8 | 21.8	37.7 | 84.0	7.7 | 18.9	18.6 | 50.2	5.8 | 17.6	5.6 | 15.2	4.8 | 12.6	3.9 | 12.3	41.6 | 39.7	1.1
`en_turkic`	34.9 | 85.3	1.7 | 5.3	6.0 | 18.9	36.8 | 81.6	7.8 | 18.7	19.6 | 50.9	6.2 | 18.6	5.9 | 14.7	6.2 | 15.8	5.5 | 14.7	35.8 | 38.4	1.7
`ru_turkic`	35.2 | 83.2	2.0 | 5.7	6.2 | 19.5	38.6 | 80.6	6.6 | 16.4	18.6 | 48.2	7.0 | 20.8	5.5 | 15.0	4.9 | 13.0	4.4 | 13.8	38.6 | 39.4	1.8
`en_ru_turkic`	31.5 | 85.3	1.6 | 5.1	5.7 | 18.0	32.0 | 75.0	6.2 | 16.6	17.9 | 49.4	6.0 | 18.6	5.7 | 16.1	5.2 | 13.7	4.6 | 13.8	42.3 | 40.8	3.2
`all_turkic`	26.7 | 75.9	1.5 | 4.9	4.9 | 17.2	11.7 | 29.0	4.9 | 13.1	15.7 | 45.0	5.6 | 18.1	3.3 | 9.0	4.1 | 11.0	3.0 | 10.3	56.7 | 54.3	2.4
`all_languages`	29.9 | 82.2	1.9 | 5.6	5.4 | 18.7	11.9 | 28.6	5.4 | 13.9	16.0 | 44.8	5.5 | 16.5	2.9 | 8.7	4.7 | 12.3	2.8 | 10.2	53.7 | 52.6	4.7

Note. The green shading indicates the best results.

Table 7. The CER (%) | WER (%) results of eight models on the KSC, TSC, and USC test sets.

Model	Language		
	kk	tr	uz
`kk_ksc`	2.0 | 6.8	-	-
`tr_tsc`	-	3.8 | 12.6	-
`uz_usc`	-	-	5.0 | 16.8
`ksc_turkic`	1.5 | 5.7	-	-
`tsc_turkic`	-	2.9 | 9.6	-
`usc_turkic`	-	-	3.2 | 10.8
`all_turkic`	1.5 | 6.0	2.9 | 10.6	2.7 | 9.5
`all_languages`	1.5 | 5.9	3.0 | 10.8	2.9 | 10.2

Note. The green shading indicates the best results.

As can be seen from Table 6, for the CVC test sets, the `all_turkic` model, trained on the datasets of the Turkic languages, performed best, achieving the lowest CER and WER scores for six out of the ten target languages. The `all_languages` model, trained on all the 15 datasets in the study (with the addition of English and Russian), produced the lowest CER and WER scores for Tatar, Turkish, and Uzbek. Of note is Kazakh, for which the lowest CER score was achieved by `all_turkic`, while the lowest WER score was obtained by `all_languages`. However, the difference between the scores was negligibly small.

What stands out in Table 7 is that, when evaluated on the KSC, TSC, and USC test sets, the `all_turkic` and `all_languages` models mostly produced second best CER/WER scores, yielding to `ksc_turkic` and `tsc_turkic`, although not considerably. Nevertheless, `all_turkic` was able to achieve even lower CER/WER scores for Uzbek than `all_languages`, evaluated on the corresponding CVC test set.

4.1. Monolingual versus Multilingual Models

In Table 6, it is noticeable that all the monolingual models were outperformed by the multilingual models. To better illustrate how the ten Turkic languages were recognized by the monolingual models and the best performing `all_turkic` model, we present some of the decoded samples in Table 8.

From Table 6, we can see that improvement was at its peak for the lowest-resourced language in the study, Azerbaijani. With only a 0.13-hour-long dataset available, a significant CER/WER reduction from 107.6% to 26.7% and from 325.7% to 75.9%, respectively, was observed for this language. In Table 8, the monolingual `az_cvc` model appears to have output the same sequence *də* based on the likelihood model and thus did not produce correct results. In comparison, the `all_turkic` model generated both an intelligible and a comprehensible text with respect to the reference text, although it systematically failed to correctly predict words with the character ə, representing the /e/ sound, in the Azerbaijani utterance. Presumably, due to the lack of Azerbaijani training data, the model therefore proposed similar-sounding words that only slightly differed in spelling, originating from Turkish (*illerde*, *faaliyetine*) and Uzbek (*muxtalif*).

Table 8. Sample ASR results for monolingual models and the `all_turkic` model (R: reference, P_mono: prediction of a monolingual model, P_all_turkic: prediction of `all_turkic`).

Lang	Type	Text								CER	WER
az	R	*müxtəlif*	*illərdə*	*fərqli*	*sahələrdə*	*iş*	*fəaliyyətinə*	*başlayır*		0.0	0.0
	P_mono	***	*də*	*də*	*də*	*də*	*də*	*ır*		171.7	>100
	P_all_turkic	*müxtalif*	*illerde*	*fergili*	*sohalarda*	*iş*	*faaliyetine*	*başlayır*		24.1	71.4
ba	R	*лотерея*	*билеты*	*һымаҡтыр*	*инде*	*ул*				0.0	0.0
	P_mono	*нафария*	*пивиста*	*һымаҡтыр*	*инде*	*ул*				30.3	33.3
	P_all_turkic	*лотерея*	*билеты*	*һымаҡтыр*	*инде*	*ул*				0.0	0.0
ch	R	*заведующий*	*çемçе*	*диван*	*холодильник*	*микрохумлă*	*кăмака*	*ыйтнă*		0.0	0.0
	P_mono	*хĕветувĕççи*	*çемçе*	*тиван*	*халакельн*е*	*микра*ăнлă*	*кăмата*	*ыйтнă*		38.3	100
	P_all_turkic	*совету***	*çемçе*	*тиван*	*холодильник*	*микрохумлă*	*камака*	*ыйтнă*		15.0	50.0
kk	R	*өз*	*елімнің*	*басы*	*болмасам*	*да*	*сайының*	*тасы*	*болайын*	0.0	0.0
	P_mono	*көзі жерген*	*жайдың*	*болан*	*боламан*	*жаланын*	*байдын*	*болан*	*байды*	84.0	>100
	P_all_turkic	*үз*	*елемнің*	*басы*	*болмасам*	*да*	*сайының*	*тасы*	*болайын*	2.1	25.0
ky	R	*исак*	*өзү*	*айткандай*	*анын*	*нанын*	*бүт*	*көчөдөгүлөр*	*алчу*	0.0	0.0
	P_mono	*исак*	*өз*	*айткандай*	*анын*	*анын*	*бир*	*көчөдөгүлөр*	*болчу*	73.1	50.0
	P_all_turkic	*исак*	*өзү*	*айткандай*	*анын*	*наанын*	*бүт*	*көчөдөгүлөр*	*алчу*	4.0	12.5
sa	R	*оо*	*онтон*	*ону*	*баран*	*хаһан*	*ылыахха*	*буоллаҕа*	*дии*	0.0	0.0
	P_mono	*оо*	*онтон*	*ому*	*байан*	*хаһан*	*ыҥыах*	*тоҕуоллаҕа*	*duhu*	19.1	75.0
	P_all_turkic	*оо*	*онтон*	*уну*	*баран*	*хаан*	*ылыакка*	*буолду*	*дии*	17.8	47.6
tt	R	*азанны*	*иң*	*оста*	*әйтүче*	*рөстәм*	*ибатуллин*	*булып*	*чыкты*	0.0	0.0
	P_mono	*узанны*	*иң*	*оста*	*итү*	*чәрстән*	*батыр*	*булып*	*чыкты*	25.0	50.0
	P_all_turkic	*азанны*	*иң*	*оста*	*итүче*	*рөстәм*	*ибатуллин*	*булып*	*чыкты*	21.2	12.5
tr	R	*ormanın*	*bütün*	*dalları*	*bütün*	*yaprakları*	*ötüyor*	*haykırıyordu*		0.0	0.0
	P_mono	*ormanın*	*bütün*	*damları*	*bütün*	*yaprakları*	*atıyor*	*aykılıyordu*		10.0	42.9
	P_all_turkic	*ormanın*	*bütün*	*dalları*	*bütün*	*yaprakları*	*ötüyor*	*haykırıyordu*		0.0	0.0
ug	R	ھېسابلىناتتى	ماشىنا	ئېسىل	شەھرىمىزدە	چاغدا	ئەينى	ماشىنىسى	ئۇنىڭ	0.0	0.0
	P_mono	ھېسابلىناتتى	مۇشۇنى	شەئېسىل	شەھرىمزدە	چاغدا	ئەينى	ماشىنىسى	ئۇنىڭ	14.1	62.5
	P_all_turkic	ھېسابلىناتتى	ماشىنا	ئېسىل	شەھرىمزدە	چاغدا	ئەينى	ماشىنىسى	ئۇنىڭ	0.0	0.0
uz	R	*biroq*	*o*	*sha*	*vaziyatda*	*bunga*	*jur*	*at*	*etolmadi*	0.0	0.0
	P_mono	*biroq*	*o*	*sha*	*vaziyatda*	*bundan*	*jur*	*at*	*etolmadim*	6.7	25.0
	P_all_turkic	*biroq*	*o*	*sha*	*vaziyatda*	*bundan*	*jur*	*at*	*etolmadi*	4.7	12.5

Note. The green and red shading indicates correctly and incorrectly predicted words, respectively. A WER of >100 refers to the presence of insertion errors, which were not presented for visualization purposes. The asterisk signs (*) refer to deletion errors by a model.

The CER/WER reduction trend held for another two lower-resourced languages in the study. The multilingual models for Chuvash and Sakha—the only representatives of their branches—were able to notably decrease CER/WER for both languages, despite their considerable deviation from standard Turkic forms. The `all_turkic` model produced scores of 4.9%/17.2% and 15.7%/45.0% for Chuvash and Sakha, respectively, which is more than twice as low as the scores obtained by the corresponding monolingual models.

With respect to three Kipchak Turkic languages—namely, Bashkir, Kyrgyz, and Tatar—there was also a reduction in CER/WER observed, although to a different degree and thanks to different models. While the scores of 13.6%/37.9% by the Tatar monolingual model were reduced to 5.5%/16.5% by `all_languages`, it was `all_turkic` again that took the Kyrgyz baseline scores down to 4.9%/13.1%. That said, the monolingual model for Bashkir—the Turkic language whose CVC data were over 230 h in length—yielded CER/WER scores that were not considerably higher than the lowest scores by `all_turkic`, 1.7%/5.5% and 1.5%/4.9%, respectively. These observations seem to suggest that CER/WER reduction is

more notable for languages with lower amounts of (CVC) data (e.g., Azerbaijani, Chuvash, Kazakh, and Sakha) and less evident for languages with a higher number of resources (e.g., Bashkir and Uzbek). Despite the less remarkable CER/WER improvement for Bashkir than for the lower-resourced languages, it can be clearly seen in Table 8 that, in contrast to the monolingual `ba_cvc` model, the `all_turkic` model was successful in recognizing loanwords, especially those taken from Russian and instantly familiar to most people in the former Soviet countries (*лотерея*, *билеты*). Similary, the `all_turkic` model outperformed the monolingual Chuvash model in predicting loanwords, recognizing some completely correctly (*холодильник*) and others to varying degrees (*заведующий* → *совету****, *диван* → *тиван*).

ASR for Uyghur, a language of the Karluk branch, also seems to have notably benefited from the development of multilingual models. One can see a steady decrement in CER/WER as the data of other Turkic languages were added to the training set. The joint use of data of all the Turkic languages in the `all_turkic` model resulted in scores of 4.1%/11.0%.

In the case of Kazakh, Turkish, and Uzbek—the three languages in the study for which in addition to the CVC there was another speech corpus used for model development—the data in Tables 6 and 7 appear to suggest that the results may vary depending on the training and test sets used. To begin with, the Kazakh and Turkish monolingual models trained on the CVC data produced notably higher CER/WER results than the monolingual models trained on the KSC and the TSC. This can probably be attributed to the marked difference in the size of the training data. It is especially the case for Kazakh, for which the total amount of the CVC data was as little as 1.60 h as opposed to the hefty 332.60 h in the KSC. Thus, it seems nothing but expected that `kk_ksc` and `tr_tsc` achieved the remarkable 2.0%/6.8% and 3.8%/12.6%, respectively, as compared to the 69.9%/101.2% of `kk_cvc` and the 7.3%/20.1% of `tr_cvc`. For example, the scores of `uz_cvc` and `uz_usc` were quite similar—although slightly lower for the former (4.2%/14.6% and 5.0%/16.8%, respectively), for the two Uzbek datasets were comparable in size.

As regards the multilingual models for Kazakh, Turkish, and Uzbek, when evaluated on the CVC test sets, the best performance was achieved by the `all_languages` model. While the CER score for Turkish and Uzbek was approximately 2.9%, the WER score held in the range of 8.7% to 10.2%. For Kazakh, the model produced the lowest WER score (28.6%), but achieved the second best CER score of 11.9%, yielding to `all_turkic` with 11.7%.

On the other hand, in the evaluation of the multilingual models on the KSC, TSC, and USC test sets, the best CER and WER results of 1.5% and 5.7%, respectively, in Kazakh ASR were produced by the `ksc_turkic` model. Such low scores are likely to have been achieved owing to the sufficient amount of data in the training and test sets for the model to learn from and test its hypotheses on. The CER scores produced by `all_turkic` and `all_languages` were identical to that of `ksc_turkic`, with the WER scores being only negligibly higher. For Turkish, the lowest scores were achieved by `tsc_turkic`, 2.9%/9.6%. Although the model exhibited a CER result lower than that obtained on the CVC test set, the WER score was still slightly higher. Looking at the scores for Turkish ASR in Tables 6 and 7, it is apparent that the multilingual models evaluated both on the CVC and TSC test sets produced somewhat similar results. In the case of the Uzbek language, the result of 2.7%/9.5% achieved by `all_turkic` was the lowest in the evaluation of the multilingual models on both test sets.

4.2. Multilingual ASR versus Transfer Learning

For comparison purposes, we conducted additional experiments using transfer learning. We pre-trained monolingual models for the two highest-resourced Turkic languages in the study (i.e., Kazakh and Turkish). Then, we finetuned the models on the three lowest-resourced Turkic languages (i.e., Azerbaijani, Chuvash, and Sakha).

As can be seen from Table 9, the two models built using transfer learning (`tl_ksc` and `tl_tsc`) produced considerably lower CER/WER scores than those of the monolingual baselines. That said, they were still higher than the scores of the multilingual `all_turkic`

model. Of note was the CER score for Sakha produced by the model that was pre-trained on Kazakh data, proving the best CER score in the study for this language.

Table 9. The CER (%) | WER (%) results of monolingual models, models built using transfer learning, and `all_turkic` for Azerbaijani, Chuvash, and Sakha.

Model	Model Type	Language		
		az	**ch**	**sa**
`_cvc`	monolingual baseline	107.6 | 325.7	15.5 | 46.2	35.3 | 82.9
`tl_ksc`	transfer learning	87.1 | 329.8	7.6 | 30.2	15.3 | 54.6
`tl_tsc`	transfer learning	86.7 | 290.1	8.4 | 31.7	17.4 | 59.1
`all_turkic`	multilingual	26.7 | 75.9	4.9 | 17.2	15.7 | 45.0

Note. The green shading indicates the best results.

4.3. Turkic versus Non-Turkic

The experiments clearly show that ASR for the Turkic languages appears to have benefited more from multilingual models trained jointly on the data of other related (Turkic) languages (e.g., `all_turkic`) than from models developed using data from non-Turkic (control) languages (i.e., English and Russian). We attribute this to essential linguistic features shared by Turkic languages [23,24].

Nevertheless, looking at Table 6, one cannot but admit that the CER/WER scores for six out of the ten Turkic languages produced by the `en_ru_turkic` model are appealing. The joint use of English and Russian training data led to a remarkable CER/WER reduction from 107.6%/325.7% to 31.5%/85.3.6% for Azerbaijani and an approximately twofold CER/WER decrease for Chuvash, Kazakh, Kyrgyz, Sakha, and Tatar. Of note are also the results of `ru_turkic` for Turkish and Uyghur. Although the scores were higher than those of `all_turkic` and `all_languages`, they were still lower than those of the corresponding monolingual baselines.

While the results of `en_ru_turkic` can be attributed to the likely presence of international words found (phonetically almost unchanged) in many languages (e.g., alcohol, Internet, computer, etc.) and Russian loanwords widely used in the six languages, the reduction in CER/WER for Turkish and Uyghur could be explained by the findings of [59]. The researchers found that the amount of source language data was more important than the relatedness of the source language to the target language, yielding greater performance. In other words, training a model on more than 300 h of transcribed speech in an unrelated language is more likely to result in CER/WER reduction than developing a model trained on data of a related and similar language, but of a smaller size.

4.4. Language Identification

Since we prepended language IDs to utterances, we were also able to evaluate the best-performing model, `all_turkic`, in terms of LID. The confusion matrix in Table 10 provides a clearer insight into the model performance and the errors made in predicting the language of an utterance. These results were obtained on all the target language test sets, which included the CVC, the KSC, the TSC, and the USC.

As can be seen from Table 10, the accuracy of the `all_turkic` model in LID was above 97% for seven out the ten Turkic languages. While the LID accuracy scores of 36.36% for Azerbaijani and 77.50% for Sakha may be explained by the insufficient amount of training data available for the languages (only 6.74 h in aggregate), the score of 80.86% for Tatar is the result of the frequent failure of the model to discriminate Tatar from Bashkir. Almost 16% of the Tatar utterances were identified as Bashkir, which should come as no surprise, given the close phonetic affinity of the languages, which differ mainly in their consonant systems [24]. When failing to unambiguously identify the language, this reliance of the model on phonetic similarities between languages—being particularly strong when they belong to the same branch—can be especially observed for Azerbaijani, Kyrgyz, and

Uyghur. Of 22 Azerbaijani utterances, ten were predicted to be Turkish, both languages being from the Oghuz branch. Most of the erroneously predicted Kyrgyz utterances were identified as either Bashkir or Kazakh (Kipchak languages). The second most likely language in the recognition of Uyghur utterances was Uzbek, the other language from the Karluk branch.

However, this should be taken with a grain of salt, for this observation also holds when the amount of training data of the actual language is lower than that of the closely related but falsely predicted language. That is, we can assume that Azerbaijani utterances were often misidentified as Turkish, Tatar utterances as Bashkir, and Uyghur speech as Uzbek, mainly because there were fewer training data for Azerbaijani, Tatar, and Uyghur than for their close relatives. A closer look at Table 10 reveals that Turkish, Bashkir, and Uzbek, when misidentified, were not necessarily recognized as languages that come from the same branch, but rather as languages with data of considerable size in the training set. The case of Sakha—the second lowest-resourced and one of the two languages with the greatest linguistic distance from the other languages in the study, the other being Chuvash—can serve as an example. It is apparent from the confusion matrix that the `all_turkic` model only minimally confused the other languages with Sakha. Overall, we can conclude that the amount of data of a language in the mixed dataset and language relatedness were probably the two most important factors influencing the ability of the `all_turkic` model to successfully identify languages.

Table 10. Language identification confusion matrix and accuracy (Acc, %) of the `all_turkic` model for different Turkic languages on the combined test sets (CVC, KSC, TSC, and USC).

		Predicted Language									
Actual Language	**Lang**	**az**	**ba**	**ch**	**kk**	**ky**	**sa**	**tr**	**tt**	**ug**	**uz**	**Acc**
	az	8	2	0	0	0	0	10	0	0	2	36.36
	ba	0	14,400	12	2	15	0	43	8	18	28	99.13
	ch	0	7	1,241	0	4	0	8	7	0	0	97.95
	kk	0	34	6	3634	23	0	2	0	8	11	97.74
	ky	2	17	11	22	1534	2	6	12	2	5	95.10
	sa	4	108	10	8	105	968	2	25	11	8	77.50
	tr	34	181	26	4	25	1	12,242	12	39	46	97.08
	tt	2	817	41	30	25	2	22	4139	12	29	80.86
	ug	1	9	2	1	17	0	2	2	2681	32	97.60
	uz	13	133	10	3	17	0	18	15	26	15,833	98.54

Note. The green shading indicates the instances where an actual class and a predicted class match.

5. Conclusions

This study set out to develop a multilingual ASR model for lower-resourced Turkic languages. Ten languages—namely, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek—were considered. A total of 22 models were developed, of which 13 were monolingual and 9 multilingual. The multilingual models outperformed the monolingual baselines, with the best performing model (i.e., `all_turkic`) achieving an average boost of 56.7%/54.3% in CER and WER reduction, respectively, for six out of the ten languages. The experiment results showed that CER and WER reduction was more likely to be observed when multilingual models were trained on the data of Turkic languages than when developed using data from such non-Turkic languages as English and Russian. The study also presented the TSC—an open-source speech corpus for the Turkish language. The corpus contains 218.2 h of transcribed speech comprising over 186,171 utterances and is the largest publicly available Turkish dataset of its kind. The datasets and codes used to train the models are available for download from https://github.com/IS2AI/TurkicASR (accessed on 22 January 2023). It is hoped that our work will stimulate further efforts in training ASR systems for Turkic languages. Presumably, the `all_turkic` model can serve as a springboard for transfer learning for monolingual ASR models for other Turkic

languages whose inclusion in our study proved challenging due to the lack of open-source corpora.

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

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The datasets and models can be found here: https://github.com/IS2AI/TurkicASR (accessed on 22 January 2023).

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

Abbreviations

The following abbreviations are used in this manuscript:

CER	character error rate
CTC	connectionist temporal classification
CVC	Common Voice Corpus 10.0
DNN	deep neural network
E2E	end-to-end
GPU	graphics processing unit
ID	identifier
KSC	Kazakh Speech Corpus
LID	language identification
METU	the Middle East Technical University
OpenSTT	Russian Open Speech To Text Dataset
TSC	Turkish Speech Corpus
USC	Uzbek Speech Corpus
WER	word error rate

References

1. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2009.
2. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
3. Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 4218–4222.
4. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
5. Godfrey, J.; Holliman, E.; McDaniel, J. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), San Francisco, CA, USA, 23–26 March 1992; Volume 1, pp. 517–520. [CrossRef]
6. Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.L.; Stolcke, A.; Yu, D.; Zweig, G. Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2410–2423. [CrossRef]
7. Guo, T.; Wen, C.; Jiang, D.; Luo, N.; Zhang, R.; Zhao, S.; Li, W.; Gong, C.; Zou, W.; Han, K.; et al. *DiDiSpeech: A Large Scale Mandarin Speech Corpus*; In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6968–6972.
8. Maekawa, K. Corpus of Spontaneous Japanese: Its design and evaluation. In Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, 13–16 April 2003.

9. Kunze, J.; Kirsch, L.; Kurenkov, I.; Krug, A.; Johannsmeier, J.; Stober, S. Transfer Learning for Speech Recognition on a Budget. In Proceedings of the Workshop on Representation Learning for NLP, Vancouver, BC, Canada, 4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 168–177. [CrossRef]
10. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio Augmentation for Speech Recognition. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 3586–3589. [CrossRef]
11. Khare, S.; Mittal, A.; Diwan, A.; Sarawagi, S.; Jyothi, P.; Bharadwaj, S. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 1529–1533. [CrossRef]
12. Toshniwal, S.; Sainath, T.N.; Weiss, R.J.; Li, B.; Moreno, P.J.; Weinstein, E.; Rao, K. Multilingual Speech Recognition with a Single End-to-End Model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4904–4908.
13. Li, B.; Pang, R.; Sainath, T.N.; Gulati, A.; Zhang, Y.; Qin, J.; Haghani, P.; Huang, W.R.; Ma, M.; Bai, J. Scaling End-to-End Models for Large-Scale Multilingual ASR. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 15–17 December 2021; pp. 1011–1018. [CrossRef]
14. Pratap, V.; Sriram, A.; Tomasello, P.; Hannun, A.; Liptchinsky, V.; Synnaeve, G.; Collobert, R. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 4751–4755. [CrossRef]
15. Li, B.; Pang, R.; Zhang, Y.; Sainath, T.N.; Strohman, T.; Haghani, P.; Zhu, Y.; Farris, B.; Gaur, N.; Prasad, M. Massively Multilingual ASR: A Lifelong Learning Solution. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6397–6401. [CrossRef]
16. Wang, D.; Zheng, T.F. Transfer Learning for Speech and Language Processing. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, China, 16–19 December 2015; pp. 1225–1237.
17. Diwan, A.; Vaideeswaran, R.; Shah, S.; Singh, A.; Raghavan, S.; Khare, S.; Unni, V.; Vyas, S.; Rajpuria, A.; Yarra, C.; et al. MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 2446–2450. [CrossRef]
18. Sailor, H.; T, K.P.; Agrawal, V.; Jain, A.; Pandey, A. SRI-B End-to-End System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 2456–2460. [CrossRef]
19. Tachbelie, M.Y.; Abate, S.T.; Schultz, T. Development of Multilingual ASR Using GlobalPhone for Less-Resourced Languages: The Case of Ethiopian Languages. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1032–1036. [CrossRef]
20. Chowdhury, S.A.; Hussein, A.; Abdelali, A.; Ali, A. Towards One Model to Rule All: Multilingual Strategy for Dialectal Code-Switching Arabic ASR. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 2466–2470. [CrossRef]
21. Mussakhojayeva, S.; Khassanov, Y.; Varol, H.A. A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. In Proceedings of the International Conference on Speech and Computer, St. Petersburg, Russia, 27–30 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 448–459. ._41. [CrossRef]
22. Hou, W.; Dong, Y.; Zhuang, B.; Yang, L.; Shi, J.; Shinozaki, T. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1037–1041. [CrossRef]
23. Campbell, G.L.; King, G. *Compendium of the World's Languages*; Routledge: Abingdon, UK, 2020.
24. Johanson, Lars and Csató, Éva Á. *The Turkic Languages*, 2nd ed.; Routledge: Abingdon, UK, 2021. [CrossRef]
25. Altun, H.O. A Comparison of Modern Turkic languages (Turkish, Azerbaijani, Kazakh, Kyrgyz, Uzbek) in Terms of Most Frequently Used 1000 Words. *Acta Turc.* **2019**, *11*, 130–144.
26. Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E. CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings. *arXiv* **2020**, arXiv:abs/2004.09249.
27. 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. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 5036–5040. [CrossRef]
28. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778. [CrossRef]
29. Rao, K.; Sak, H.; Prabhavalkar, R. Exploring Architectures, Data and Units for Streaming End-to-End Speech Recognition with RNN-Transducer. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 193–199.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; 2017; Volume 30; pp. 1–11.

31. Zhang, Q.; Lu, H.; Sak, H.; Tripathi, A.; McDermott, E.; Koo, S.; Kumar, S. Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7829–7833.
32. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
33. Kamper, H.; Matusevych, Y.; Goldwater, S. Multilingual Acoustic Word Embedding Models for Processing Zero-Resource Languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6414–6418.
34. Li, X.; Dalmia, S.; Li, J.B.; Lee, M.R.; Littell, P.; Yao, J.; Anastasopoulos, A.; Mortensen, D.R.; Neubig, G.; Black, A.W.; et al. Universal Phone Recognition with a Multilingual Allophone System. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8249–8253.
35. Yadav, H.; Sitaram, S. A Survey of Multilingual Models for Automatic Speech Recognition. In Proceedings of the Conference on Language Resources and Evaluation (LREC), Marseille, France, 20–25 June 2022; European Language Resources Association (ELRA): Marseille, France, 2022; pp. 5071–5079.
36. Ma, B.; Guan, C.; Li, H.; Lee, C.H. Multilingual Speech Recognition with Language Identification. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, CO, USA, 16–20 September 2002; pp. 505–508. [CrossRef]
37. Seki, H.; Watanabe, S.; Hori, T.; Roux, J.L.; Hershey, J.R. An End-to-End Language-Tracking Speech Recognizer for Mixed-Language Speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–18 April 2018; pp. 4919–4923. [CrossRef]
38. Shan, C.; Weng, C.; Wang, G.; Su, D.; Luo, M.; Yu, D.; Xie, L. Investigating End-to-end Speech Recognition for Mandarin-English Code-Switching. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6056–6060. [CrossRef]
39. Watanabe, S.; Hori, T.; Hershey, J.R. Language Independent End-to-End Architecture for Joint Language Identification and Speech Recognition. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 265–271.
40. N, K.D.; Wang, P.; Bozza, B. Using Large Self-Supervised Models for Low-Resource Speech Recognition. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 2436–2440. [CrossRef]
41. Khassanov, Y.; Mussakhojayeva, S.; Mirzakhmetov, A.; Adiyev, A.; Nurpeiissov, M.; Varol, H.A. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline. In Proceedings of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 19–23 April 2021; pp. 697–706. [CrossRef]
42. Musaev, M.; Mussakhojayeva, S.; Khujayorov, I.; Khassanov, Y.; Ochilov, M.; Varol, H.A. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In Proceedings of the Speech and Computer, St. Petersburg, Russia, 27–30 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 437–447.
43. Khusainov, A.; Suleymanov, D.; Muhametzyanov, I. Incorporation of Iterative Self-Supervised Pre-Training in the Creation of the ASR System for the Tatar Language. In Proceedings of the International Conference on Text, Speech, and Dialogue, Brno, Czech Republic, 6–9 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 481–488. ._41. [CrossRef]
44. Mukhamadiyev, A.; Khujayarov, I.; Djuraev, O.; Cho, J. Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language. *Sensors* **2022**, *22*, 3683. [CrossRef] [PubMed]
45. Valizada, A.; Akhundova, N.; Rustamov, S. Development of Speech Recognition Systems in Emergency Call Centers. *Symmetry* **2021**, *13*, 634. [CrossRef]
46. Salor Durna, Ö.; Pellom, B.; Çiloğlu, T.; Hacıoğlu, K.; Demirekler, M. On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, CO, USA, 16–20 September 2002; Volume 1, pp. 349–352.
47. Salor, Ö.; Pellom, B.; Çiloğlu, T.; Demirekler, M. Turkish Speech Corpora and Recognition Tools Developed by Porting SONIC: Towards Multilingual Speech Recognition. *Comput. Speech Lang.* **2007**, *21*, 580–593. [CrossRef]
48. Arisoy, E.; Can, D.; Parlak, S.; Sak, H.; Saraclar, M. Turkish Broadcast News Transcription and Retrieval. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 874–883. [CrossRef]
49. Polat, H.; Oyucu, S. Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results. *Symmetry* **2020**, *12*, 290. [CrossRef]
50. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [CrossRef]
51. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
52. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Enrique Yalta Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2207–2211. [CrossRef]

53. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–13. [CrossRef]
54. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376. [CrossRef]
55. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
56. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition. In Proceedings of the Interspeech, Jeju Island, Korea, 4–8 October 2004; pp. 2765–2768. [CrossRef]
57. Wang, P.; Sun, R.; Zhao, H.; Yu, K. A New Word Language Model Evaluation Metric for Character Based Languages. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 315–324.
58. MacKenzie, I.S.; Soukoreff, R.W. A character-level error analysis technique for evaluating text entry methods. In Proceedings of the Nordic Conference on Human–Computer Interaction, Aarhus, Denmark, 19–23 October 2002; pp. 243–246.
59. Hjortnaes, N.; Partanen, N.; Rießler, M.; Tyers, F.M. The Relevance of the Source Language in Transfer Learning for ASR. In Proceedings of the Workshop on the Use of Computational Methods in the Study of Endangered Languages. Association for Computational Linguistics, Online, 2–3 March 2021; pp. 63–69.