Noise-Robust Multilingual Speech Recognition and the Tatar Speech Corpus

Saida Mussakhojayeva
Institute of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
saida.mussakhojayeva@nu.edu.kz

Rinat Gilmullin*, Bulat Khakimov†, and Mansur Galimov‡
Institute of Applied Semiotics
Tatarstan Academy of Sciences
Kazan, Russia
*rinatgilmullin@gmail.com, †khakeem@yandex.ru, ‡magl.galimov@gmail.com

Daniil Orel§, Adal Adilbekov¶, and Huseyin Atakan Varol‖
Institute of Smart Systems and AI
Nazarbayev University
Astana, Kazakhstan
§daniil.orel@nu.edu.kz, ¶adalabilbekov@gmail.com, ‖ahvarol@nu.edu.kz

***Abstract*—After focusing on individual languages for a long time, multilingual automatic speech recognition has recently become an active area of research. For instance, Whisper by OpenAI is capable of recognizing speech in 99 languages. However, the performance of Whisper is significantly lower for low-resource languages than for high-resource ones. In this work, we aim to address this and present a fine-tuning strategy for the pre-trained Whisper model so that its performance is improved for a low-resource language family while maintaining performance for a set of high-resource languages. Specifically, our Söyle model exhibited high performance for both the Turkic language family (11 languages) and the official languages of the United Nations. Our work also presents the first large open-source speech corpus for the Tatar language. We demonstrate that speech recognition performance for Tatar improves with the model trained using the new Tatar Speech Corpus (TatSC). Our model is also trained to be noise-robust. We open-source our model and TatSC to encourage further research. We envision that our fine-tuning approach will guide the creation multilingual speech recognition models for other low-resource language families.**

I. Introduction

Automatic Speech Recognition (ASR) has seen considerable advancement in recent years, becoming a cornerstone technology in various applications such as transcription services, voice assistants, and more [1]. However, initial approaches such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) often struggled with the complexity and variability of human speech [2]. In the last decade, however, there has been a significant paradigm shift with the advent of Deep Learning (DL).

With the capabilities of DL, researchers began to push the boundaries of what ASR systems could accomplish. One of these challenging tasks was multilingual speech recognition. Instead of building a separate model for each language, the focus shifted to constructing universal models that could handle a variety of languages [3]. Despite substantial progress, a major challenge remains in achieving comparable recognition accuracy between resource-rich and resource-poor languages [4]. The imbalance stems from the lack of extensive annotated data for low-resource languages, which has resulted in their under-representation in training data and, consequently, poor model performance [5].

This work aims to address the overarching question: How can we enhance the performance of multilingual speech recognition models for low-resource languages and language families? As a salient example, we consider the Turkic language group, which, despite its linguistic significance and large number of speakers, often falls into the low-resource category in ASR research due to the lack of annotated speech data.

To bridge this gap, our work adopts a multi-faceted strategy. First, we explore the use of linguistic commonalities within language families, such as the Turkic group, as a means of propagating knowledge and enhancing performance across related low-resource languages. Second, we focus on the robustness of these systems to ensure their reliability in diverse and challenging real-world environments. The noise robustness of ASR has traditionally been linked to the size and diversity of the datasets employed. However, the availability of massive datasets is a luxury that research in low-resource languages often cannot afford. Hence, this paper employs augmentation strategies designed to enable noise robustness even with relatively small datasets.

To this end, we introduce Söyle, a multilingual speech recognition model. Söyle employs a fine-tuning strategy to Whisper—a model developed by OpenAI and trained on 680,000 hours of data from 99 languages [6]. While Whisper is promising, especially in zero-shot settings for high-resource languages, its effectiveness deteriorates for low-resource languages. Our approach enhances Whisper by fine-tuning it on a combination of 11 Turkic languages and the six official languages of the United Nations (UN), preserving its global applicability.

Central to our methodology is the integration of the newly curated Tatar Speech Corpus (TatSC). Compiled through crowdsourcing and audiobook segmentation, TatSC comprises over 269 hours of audio-text pairs, representing a significant

enrichment of the speech resources available for Turkic languages. In addition to these methodological contributions, our work includes experiments on noise robustness using the Söyle model, fine-tuned on both Turkic and UN languages. These experiments shed light on the efficacy of pre-training and fine-tuning strategies in multilingual ASR and offer key insights for future research in this area.

The structure of this paper is organized as follows: Section II offers a review of related work relevant to our study. In Section III, we describe the datasets employed, including the newly introduced Tatar Speech Corpus, and provide details of the fine-tuning techniques. Experimental findings and their subsequent discussions are presented in Section IV. Section V concludes the paper.

II. Related Work

A. Multilingual Speech Recognition

Historically, the focus of ASR systems has been predominantly on individual languages. However, with the emergence of DL, this perspective started to shift towards the development of shared models that can handle multiple languages [7], [8]. In recent years, multilingual end-to-end models have been proposed [3], [9], [10]. They do not rely on language-specific characteristics, but instead learn to map speech directly to text across a variety of languages. A common technique for building robust end-to-end multilingual ASR models is to use transfer learning [6], [11], [12], which uses resource-rich languages to maximize the performance of resource-poor languages. Several works have approached the problem of multilingual speech recognition by incorporating a language identification (LID) module into the model architecture [13], [14]. On the other hand, some works (e.g., [9]) did not utilize the LID tag during inference but instead predicted a language ID as well as text input.

B. ASR for Low-Resource Languages

Early ASR approaches for low-resource languages included Connectionist Temporal Classification (CTC) networks [15] and Recurrent Neural Network (RNN) encoder-decoders with attention [16]. Alternative studies have explored two-pass systems with Monotonic Chunkwise Attention (MoChA) [17] and multi-task learning-based transformer models [18]. Research has also delved into self-supervised models trained on unlabeled data to produce representations beneficial for low-resource languages [19]. Utilizing pre-trained wav2vec 2.0 [20], experiments showed a relative reduction in the Word Error Rate (WER) for Indian languages.

Several datasets and resources have been introduced to address the challenges posed by low-resource languages. Examples include the Kazakh Speech Corpus (KSC) [21] and its updated version [22], the THUYG-20 database for Uyghur [23], the Uzbek Speech Corpus (USC) [24], and the Turkish Speech Corpus (TSC) [10]. Kazakh, Kyrgyz, and Uyghur are also present in larger multilingual corpora (e.g., M2ASR [25]).

C. Noise Robust Speech Recognition

In ASR systems, remarkable progress has been achieved in achieving high accuracy for both mono- and multilingual scenarios. However, the presence of noise poses significant challenges in real-world applications. One potential solution is to separate the original speech and the accompanying noise. The separation can be based on classical methods [26] or DL. There are a variety of ways of DNN application for speech enhancement: ensembled perceptrons [27], CNNs [28], and generative models [29]. These approaches demonstrate a significant improvement in audio quality but often result in artifacts limiting the further use of the data for ASR.

The robustness of ASR models can be enhanced by incorporating noisy audio samples in the training process. In practice, however, it is challenging to obtain a sufficient amount of diverse and real-world noisy datasets. To address this limitation, modern corpora for robust speech recognition include simulated audio data [30]. This approach involves assembling an extensive dataset of ”clean”, noise-free audio recordings. Subsequently, these audio samples are augmented with real-world or white noise to simulate acoustic interference that mimics real-world conditions [31].

D. Whisper ASR Model

Recognizing the limitations of unsupervised pre-training and the increased robustness that supervised pre-training provides across many datasets and domains [32], OpenAI's Whisper model [6] has sought to bridge this gap. Whisper is based on a large-scale, weakly supervised pre-training approach that scales to 680,000 hours of labeled audio data and supports 99 languages. Whisper extends the scope of weakly supervised pre-training beyond English-only speech recognition to be both multilingual and multitasking. For further details on the pre-training procedures and the specific hyper-parameter values of the Whisper model, readers are referred to [6].

III. Methods

A. Datasets

In this study, we considered a total of 17 languages, including 11 Turkic languages along with the six official languages of the UN—Arabic, Chinese, English, French, Russian, and Spanish. A detailed description of these languages and their respective classifications is presented in Table I. The inclusion of the Turkic languages enables us to explore how multilingual ASR models can be adapted for low-resource language families. At the same time, the six official languages of the UN were included to better assess the generalizability of the model. These languages represent a significant portion of the world's population and are used in diverse geopolitical, cultural, and socio-economic contexts.

We used several data sources to compile the training corpora, including Common Voice (CVC) [33] version 13.0, KSC2 [22], TSC [10], USC [24], and FLEURS [34]. Among these, CVC stands out as one of the largest publicly available multilingual datasets that encompasses a wide variety of accents, demographics, and recording environments. The

TABLE I
THE LANGUAGES AND DATASETS IN THE STUDY

Language	Code	Family	Script	Corpus	Duration (hr)
Azerbaijani	az	Turkic	Latin	CVC	0.13
				FLEURS	13.99
Bashkir	ba	Turkic	Cyrillic	CVC	239.66
Chuvash	cv	Turkic	Cyrillic	CVC	13.19
Kazakh	kk	Turkic	Cyrillic	CVC	1.69
				KSC2	1,194.16
				FLEURS	3.85
Kyrgyz	ky	Turkic	Cyrillic	CVC	19.28
				FLEURS	3.27
Sakha	sah	Turkic	Cyrillic	CVC	6.78
Turkmen	tk	Turkic	Latin	CVC	1.30
Turkish	tr	Turkic	Latin	CVC	72.24
				TSC	218.23
				FLEURS	2.62
Tatar	tt	Turkic	Cyrillic	CVC	26.55
				TatSC	269.15
Uyghur	ug	Turkic	Arabic	CVC	62.63
Uzbek	uz	Turkic	Latin	CVC	102.30
				USC	104.91
				FLEURS	2.86
Arabic	ar	Afroasiatic	Arabic	CVC	81.82
				FLEURS	1.31
Chinese	zh	Sino-Tibetan	Chinese	CVC	248.20
				FLEURS	3.10
English	en	Indo-Eur.	Latin	CVC	2,430.18
				FLEURS	1.79
French	fr	Indo-Eur.	Latin	CVC	944.28
				FLEURS	1.97
Russian	ru	Indo-Eur.	Cyrillic	CVC	165.09
				FLEURS	2.52
Spanish	es	Indo-Eur.	Latin	CVC	480.31
				FLEURS	3.11

CVC dataset, which comprises approximately 18,000 validated hours in 112 languages, was collected via a crowdsourcing platform. In addition, this study makes use of FLEURS, a parallel speech dataset that includes 102 languages and was built based on the machine translation FLoRes-101 benchmark [35]. Of the 17 languages considered in our research, all of the six official UN languages are included in FLEURS, whereas only 5 of the 11 Turkic languages are present in this dataset (Azerbaijani, Kazakh, Kyrgyz, Turkish, and Uzbek). Specifically, for Azerbaijani, the inclusion of FLEURS was necessary because of the limited amount of data in CVC, which consists of only 8 minutes of speech. For other languages examined in this study that are present in FLEURS, only the testing data were utilized. This allowed for a more nuanced assessment of the zero-shot generalization of the model across different linguistic contexts.

B. Tatar Speech Corpus

The Tatar Speech Corpus (TatSC) is the first large open-source speech corpus for Tatar. There are no Tatar speech corpora of sufficient size to develop modern ASR models, while the only available open-source corpus is provided on the Common Voice platform [33] and its total duration barely exceeds 31 hours. TatSC consists of texts narrated by recruited speakers, crowdsourced data collected using a social media bot, and audiobooks (see Table II).

1) Website Sentences: The first part of the corpus consists of sentences obtained from various Tatar websites. Speakers were then recruited to narrate these sentences. To ensure the quality of the recordings, the recording sessions were conducted using laptops and headsets in a noise-free environment. A total of 99.5 hours of data were collected, which included 87,425 utterances.

TABLE II
TATAR SPEECH CORPUS SPECIFICATIONS

Source	Duration (hr)	Utterances	Words	Unique words
Web	99.5	87,425	540,584	50,719
Telegram	146.1	110,683	881,168	12,957
Audiobooks	23.5	19,806	171,117	28,214
Total	**269.1**	**217,914**	**1,592,869**	**68,623**

2) Telegram Bot: To diversify the speech corpus to include a range of voices and background noises, a Telegram bot was utilized. The bot facilitated the collection of speech from speakers of varying ages and genders. The balance and representativeness of the dataset were achieved by exporting sentences from the Tatar National Corpus "Tugan Tel". The corpus was analyzed in depth to ensure linguistic representativeness through N-gram analysis and frequency lists of lemmas and inflections. The bot is equipped with two functions: narration and evaluation. The user starts the narration process by first providing age and gender information. Then the user is presented with a random sentence to narrate. The user has the option to listen to the recorded audio, record it again, or skip the sentence. In total, the dataset included 667 speakers, including 437 females and 230 males, resulting in 146.1 hours of speech with 110,683 utterances.

3) Audiobooks: The final segment of TatSC consists of audiobooks in Tatar with a combined duration of 23.5 hours, comprising 19,806 utterances and a vocabulary of 28,214 unique words. In order to secure a broad and representative collection, we referred to the Tatar Book Publishers' collection[1] and other online platforms offering open-access audiobooks. The audiobooks were professionally recorded in a studio by female and male narrators, largely theatre actors. The extensive preparation process involved several crucial steps, including the selection of relevant audiobooks, the alignment and manual verification of sentences, and the creation of the final database. This process was facilitated by the open-source platform, Label Studio[2], and carried out by native speakers with linguistic expertise.

The final TatSC dataset was divided into training and evaluation sets. The evaluation sets include the development set, which was used to fine-tune the training process, and the test set, which was used to report metric results. Each evaluation set consisted of around seven hours of data. In total, we collected 269.1 hours of data with the corresponding 217,914 utterances.

[1]https://tatkniga.ru/
[2]https://labelstud.io/

C. Noise Robustness

To make our model robust to noise, we additionally fine-tuned it on noise-augmented audios. For this purpose, we simulated noisy audio recordings using augmentation. As a source for the noises, we utilized Audio Set, a large-scale manually-annotated collection of audio events [36]. The key benefit of this data source is that it contains more than two million audio materials that can be used as noises. However, due to the size of the dataset, not all audio categories are relevant and of high quality. Therefore, we only used the categories that were rated by human evaluators with over 90% consistency. As a result, we collected 54,260 noise samples from 20 different categories.

These noise samples were used to tune our pre-trained model. Specifically, we trained our Whisper Medium model for an additional four epochs with noise-augmented samples. Half of the utterances within the original dataset underwent augmentation by the introduction of noise, employing specific signal-to-noise ratios (SNRs) chosen from the set [100, 75, 50, 40, 30, 25].

D. ASR Model Training

We named our model "Söyle", a word translated as "speak" in most Turkic languages. In our study, we developed two versions of the Söyle model: 1) `söyle-trc` was fine-tuned exclusively on the Turkic languages, and 2) `söyle` was fine-tuned on both the Turkic and official UN languages.

Following the Whisper guidelines, we used the Byte-Pair Encoding (BPE) text tokenizer [37]. The batch size was set to 16 per graphics processing unit (GPU), making an effective batch size of 64. The learning rate was 1×10^{-5} with 500 warmup steps. Our models were initialized with Whisper-medium weights and underwent training for six epochs, a duration deemed adequate for training purposes. The training process for each model was conducted on four Nvidia DGX A100 (40 GB) GPUs.

For benchmarking, we first evaluated all languages in our study using the pre-trained Whisper model, which we did not train ourselves but used the off-the-shelf implementation from Hugging Face [38]. We excluded Chuvash, Kyrgyz, Sakha, and Uyghur, because their LID tags were not included in the off-the-shelf model. As an additional reference point, we recreated the state-of-the-art models for Turkic languages developed in [10], which we will further refer to as `baseline-turkic` and `baseline`. These two models are built upon the ESPNet framework [39]. Accordingly, we employed three models for comparison with our Söyle models: 1) `whisper`: the original Whisper-medium model with 769 million trainable parameters, which was used for evaluation only, 2) `baseline-trc`: an ESPNet-based model trained with datasets for the Turkic languages only, and 3) `baseline`: an ESPNet-based model trained with all datasets in Table I (i.e., both Turkic and UN languages).

E. Data Normalization

In normalizing the datasets used in this study, we primarily followed the text standardization procedure outlined in Whisper [6]. Specifically, we converted all letters to lowercase and removed all phrases enclosed in square brackets or parentheses. All markers, symbols, and punctuation characters that fell into the Unicode categories starting with M, S, or P in the NFKC-normalized string were replaced with a space. Consecutive whitespace characters were replaced with a single space. For languages that do not use spaces to separate words, such as Chinese, a space was inserted between each letter. This method is in line with the standardization process used in Whisper and helps to ensure a fair comparison between different models and languages.

F. Evaluation Metrics

When evaluating ASR systems, one of the major metrics is the word error rate (WER). It is calculated by dividing the sum of substitutions, deletions, and insertions required to match the transcribed text with the reference by the number of words in the reference. In essence, WER provides a percentage of the errors made in the transcription. While WER serves as an effective metric for most languages, its applicability to Chinese is limited because of the lack of spaces between words. In the assessment of Chinese ASR, spaces are inserted between characters. This adjustment shifts the evaluation metric from word-level accuracy to character-level accuracy, ultimately leading to the calculation of the character error rate (CER).

IV. Results & Discussion

Table III presents the performance of the models. The models are categorized based on their training data, distinguishing between those trained solely on Turkic languages (the `baseline-trc` and `söyle-trc` columns) and those trained on both Turkic and UN languages (the `baseline` and `söyle` columns). The best performance in each row is highlighted in bold for clarity. The `söyle` model, fine-tuned on a combination of the Turkic languages and the six UN languages, demonstrated exceptional results, achieving a WER of less than 30% WER across 17 of the 21 corpora considered. Notably, it outperformed the `baseline` model in all instances except two (Chuvash and Arabic CVC). The `baseline-trc` model exhibited superior performance for the Turkic languages, attaining the lowest WER for 10 out of the 15 Turkic test sets. However, it faced challenges in recognizing the UN languages. The performance of the `baseline` model, trained on both Turkic and UN languages, displayed a decline, suggesting potential limitations in the learning capacity of the ESPNet-based model architecture. Furthermore, it is noteworthy that the `söyle` model consistently outperformed the original `whisper` model across all test sets. This observation highlights the viability of fine-tuning-based customization within the Whisper architecture, particularly for low-resource language families.

The integration of the newly developed TatSC into the söyle model appears to have effectively improved the performance of the model for Tatar. In [10], the best WER for the Tatar CVC test set was reported to be 16.5%. In contrast, the söyle model, trained on TatSC, attained a substantially improved WER of 9.1% when evaluated on the same test set.

For five of the UN languages (English, Spanish, French, Russian, and Chinese), söyle exhibited superiority over the baseline model, and it even outperformed original whisper. This result underscores that our fine-tuning strategy is effective in maintaining and even improving ASR performance for high-resource languages. However, the relatively low performance observed in the Arabic CVC test set, both for the original whisper and our söyle models, underscores the need for larger and cleaner training datasets for this language.

Table IV presents the WER scores for models evaluated on the FLEURS dataset across several languages. It is important to note that these models were not trained on the FLEURS training set, with the exception of Azerbaijani, as the original CVC dataset did not contain a sufficient amount of data. Remarkably, in this zero-shot experiment, söyle achieved the lowest WER for 9 out of 11 test sets, which illustrates its capacity for generalization.

To measure the effect of noise-robust training, we conducted a comparative evaluation between the noise-robust (söyle-NR) model and the original model across major languages from three distinct corpora (TatSC, KSC2, and CVC). This evaluation involved augmenting the data with an unseen set of test noises. Noise robust model söyle-NR outperformed the original model across the whole range of SNRs, including SNR values not used during training (5 and 1).

At the same time, it is important to mention, that for high SNR values (100-25) there is no significant degradation in the performance of the original Söyle model. This robustness can be attributed to the diversity of the training data, as described in Section III-A. The training data for söyle contained spontaneous speech and recordings from various resources, such as TV and YouTube videos. These sources often feature speech with background noise, music, overlapping speech, and other acoustic elements that contribute to the noise robustness of the model. In general, we recommend noise augmentation as a necessary step to improve the performance of multilingual ASR models, provided adequate computational resources are available.

V. Conclusions

In this work, we have presented our approach to building a multilingual ASR model through fine-tuning a pre-trained Whisper model. Our söyle model not only maintains but also improves the performance of the Whisper model for the UN languages, while improving ASR results for low-resource languages within the Turkic language family. Furthermore, we introduced the first large-scale open-source speech corpus for the Tatar language. The söyle model, trained on this new corpus, achieved the lowest WER score reported for this language using the corresponding CVC test set. In addition, we presented our augmentation recipe for improving the noise robustness of ASR models. To encourage further research in this direction, we have made the datasets and codes used in this work publicly available at https://github.com/IS2AI/Soyle.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning (ICML)*, 2016, p. 173–182.

[2] M. Gales and S. Young, *Application of Hidden Markov Models in Speech Recognition*. Now Publishers Inc, 2008.

[3] O. Scharenborg, L. Besacier, A. Black *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 04 2018, pp. 4979–4983.

[4] B. Zoph, D. Yuret, J. May *et al.*, "Transfer learning for low-resource neural machine translation," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1568–1575. [Online]. Available: https://aclanthology.org/D16-1163

[5] A. Rosenberg, Y. Zhang, B. Ramabhadran *et al.*, "Speech recognition with augmented synthesized speech," in *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[6] A. Radford, J. W. Kim, T. Xu *et al.*, "Robust speech recognition via large-scale weak supervision," 2022.

[7] J.-T. Huang, J. Li, D. Yu *et al.*, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7304–7308.

[8] K. M. Knill, M. J. F. Gales, S. P. Rath *et al.*, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 138–143.

[9] C. Zhang, B. Li, T. Sainath *et al.*, "Streaming end-to-end multilingual speech recognition with joint language identification," in *Proc. Interspeech*, 2022.

[10] S. Mussakhojayeva, K. Dauletbek, R. Yeshpanov *et al.*, "Multilingual speech recognition for Turkic languages," *Information*, vol. 14, no. 2, p. 74, Jan 2023. [Online]. Available: http://dx.doi.org/10.3390/info14020074

[11] Y. Zhang, W. Han, J. Qin *et al.*, "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[12] D. Orel, R. Yeshpanov, and H. A. Varol, "Speech recognition for Turkic languages using cross-lingual transfer learning from kazakh," in *Proc. of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2023, pp. 174–182.

[13] S. Toshniwal, T. N. Sainath, R. J. Weiss *et al.*, "Multilingual speech recognition with a single end-to-end model," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.

[14] A. Waters, N. Gaur, P. Haghani *et al.*, "Leveraging language id in multilingual end-to-end speech recognition," in *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 928–935.

[15] A. Graves, S. Fernandez, F. Gomez *et al.*, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets," in *International Conference on Machine Learning (ICML)*, 2006.

[16] D. Bahdanau, J. Chorowski, D. Serdyuk *et al.*, "End-to-end attention-based large vocabulary speech recognition," 2015.

[17] J. Kim, M. Kumar, D. Gowda *et al.*, "Semi-supervised transfer learning for language expansion of end-to-end speech recognition models to low-resource languages," 2021.

[18] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[19] N. KrishnaD., P. Wang, and B. Bozza, "Using large self-supervised models for low-resource speech recognition," in *Proc. Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239705336

TABLE III
WER (%) RESULTS ON THE CVC, KSC, TSC, TATSC, AND USC TEST SETS.

Group	Code	Corpus	whisper	baseline-trc	söyle-trc	baseline	söyle
Turkic languages	az	CVC	60.0	66.8	32.3	49.1	**25.0**
	ba	CVC	131.3	**8.0**	8.6	15.5	8.2
	cv	CVC	-	**30.6**	39.7	37.6	38.5
	kk	CVC	81.1	28.7	18.2	22.3	**17.6**
		KSC2	87.8	**11.5**	12.6	12.7	12.2
	ky	CVC	-	**13.0**	20.4	20.0	19.4
	sah	CVC	-	**47.4**	48.3	52.0	48.2
	tk	CVC	156	82.8	63.9	73.0	**63.1**
	tr	CVC	25.3	10.6	**9.7**	21.1	**9.7**
		TSC	58.6	**11.8**	12.2	17.0	12.3
	tt	CVC	103.7	12.9	**8.7**	17.4	9.1
		TatSC	110.5	**6.2**	7.9	8.0	7.8
	ug	CVC	-	**10.9**	16.2	26.0	16.2
	uz	CVC	166.6	**12.3**	13.4	21.2	13.5
		USC	173.4	**11.1**	14.3	17.5	14.0
UN languages	ar	CVC	71.0	-	112.8	**62.0**	63.6
	en	CVC	19.5	-	113.3	12.5	**9.2**
	es	CVC	19.5	-	63.0	8.7	**5.3**
	fr	CVC	32.1	-	88.6	11.8	**8.7**
	ru	CVC	19.5	-	20.4	16.8	**7.8**
	zh	CVC	16.9	-	523	9.8	**5.9**

TABLE IV
WER (%) RESULTS ON THE FLEURS TEST SET.

Model	Language										
	az	kk	ky	tr	uz	ar	en	es	fr	ru	zh
whisper	52.8	55.4	-	36.0	122.9	**34.0**	**5.4**	14.2	22.1	13.1	12.8
baseline	29.6	13.0	66.9	24.9	33.6	109.3	13.5	14.6	17.6	25.1	38.7
söyle	**23.7**	**11.8**	**23.6**	**12.5**	**19.8**	54.6	5.9	**6.0**	**9.6**	**11.6**	**9.5**

[20] S. Schneider, A. Baevski, R. Collobert *et al.*, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1873

[21] Y. Khassanov, S. Mussakhojayeva, A. Mirzakhmetov *et al.*, "A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline," in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, 04 2021, pp. 697–706.

[22] S. Mussakhojayeva, Y. Khassanov, and H. Atakan Varol, "KSC2: An industrial-scale open-source Kazakh speech corpus," in *Proc. Interspeech*, 2022, pp. 1367–1371.

[23] A. Rouzi, Y. Shi, Z. Zhiyong *et al.*, "Thuyg-20: A free uyghur speech database," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 2, p. 182, 2017. [Online]. Available: http://jst.tsinghuajournals.com/EN/abstract/article_149668.shtml

[24] M. Musaev, S. Mussakhojayeva, I. Khujayorov *et al.*, "Usc: An open-source uzbek speech corpus and initial speech recognition experiments," in *Speech and Computer*. Cham: Springer International Publishing, 2021, pp. 437–447.

[25] D. Wang, F. Zheng, Z. Tang *et al.*, "M2asr: Ambitions and first year progress," in *Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 11 2017, pp. 1–6.

[26] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Selective magnitude subtraction for speech enhancement," in *International Conference on High-Performance Computing in the Asia-Pacific Region*, vol. 2, 2000, pp. 692–695.

[27] P. Karjol, M. Kumar, and P. Ghosh, "Speech enhancement using multiple deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 04 2018, pp. 5049–5052.

[28] K. Kinoshita, T. Ochiai, M. Delcroix *et al.*, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," 2020.

[29] S. Pascual, J. Serrà, and A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech Communication*, vol. 114, p. 10–21, nov 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.09.001

[30] J. Barker, E. Vincent, N. Ma *et al.*, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, p. 621–633, 05 2013.

[31] D. Orel and H. A. Varol, "Noise-robust automatic speech recognition for industrial and urban environments," in *Proc. of the Conference of the IEEE Industrial Electronics Society (IECON)*, 2023.

[32] A. Narayanan, A. Misra, K. Sim *et al.*, "Toward domain-invariant speech recognition via large scale training," in *IEEE Spoken Language Technology Workshop (SLT)*, 12 2018, pp. 441–447.

[33] R. Ardila, M. Branson, K. Davis *et al.*, "Common voice: A massively-multilingual speech corpus," in *International Conference on Language Resources and Evaluation (LREC)*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209376338

[34] A. Conneau, M. Ma, S. Khanuja *et al.*, "Fleurs: Few-shot learning evaluation of universal representations of speech," 2022.

[35] N. Goyal, C. Gao, V. Chaudhary *et al.*, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.30

[36] J. F. Gemmeke, D. P. W. Ellis, D. Freedman *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[37] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 08 2016, pp. 1715–1725.

[38] T. Wolf, L. Debut, V. Sanh *et al.*, "Transformers: State-of-the-art natural language processing," in *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 10 2020, pp. 38–45.

[39] S. Watanabe, T. Hori, S. Karita *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456