Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper

Zhanibek Kozhirbayev

National Laboratory Astana, Nazarbayev University, Kazakhstan
Email: zhanibek.kozhirbayev@nu.edu.kz

***Abstract*—In recent years, the progress made in neural models trained on extensive multilingual text or speech data has shown great potential for improving the status of underresourced languages. This paper focuses on experimenting with three state-of-the-art speech recognition models, namely Facebook's Wav2Vec2.0 and Wav2Vec2-XLS-R, OpenAI's Whisper, on the Kazakh language. The objective of this research is to investigate the effectiveness of these models in transcribing Kazakh speech and to compare their performance with existing supervised Automatic Speech Recognition (ASR) systems. The study also aims to explore the possibility of using data from other languages for pre-training and to test whether fine-tuning the target language data can improve model performance. Thus, this work can provide insights into the effectiveness of using pretrained multilingual models in underresourced language settings. The wav2vec2.0 model achieved a Character Error Rate (CER) of 2.8 and a Word Error Rate (WER) of 8.7 on the test set, which closely matches the best result achieved by the end-to-end Transformer model. The large whisper model achieves a CER of approximately 4 on the test set. The results of this study can contribute to the development of robust and efficient ASR systems for the Kazakh language, benefiting various applications, including speech-to-text translation, voice assistants, and speech-based communication tools.**

***Keywords*—automatic speech recognition, Wav2Vec 2.0, Wav2Vec2-XLS-R, whisper, pretrained transformer models, speech representation models**

I. Introduction

In recent years, sequence-based models have demonstrated exceptional speech recognition performance compared to traditional automatic speech recognition frameworks. Sequence-based models use neural networks to learn speech-to-text mapping, which simplifies the modeling process. Among the sequence-based models, Transformer [1] is widely used and has shown remarkable success in building end-to-end speech recognition systems [2–4]. Although significant progress has been made in the development of ASR models, creating robust models for most languages other than English remains challenging. This is primarily due to the fact that state-of-the-art models typically require many hours of annotated speech for training to achieve satisfactory results. This is particularly true for Kazakh, a Turkic language spoken by more than 13 million people worldwide (according to the statistics of the Ethnologue website: https://www.ethnologue.com/language/kaz).

Recent advancements in self-supervised learning techniques have shown promise in addressing data scarcity problems for unresourced languages. Self-supervised training is a type of training for speech recognition systems that leverages the abundance of unlabeled speech data to learn useful representations of the speech signal. Unlike traditional supervised learning, where labeled data is required to train a model, self-supervised learning algorithms learn from the raw data without the need for explicit labels. In self-supervised training, the model is trained to perform tasks that are closely related to the ultimate task of speech recognition, but do not require labeled data. For example, one approach is to train the model to predict the next frame of a speech signal, given the previous frames. This is known as a Contrastive Predictive Coding (CPC) [5] task. Another approach is to train the model to differentiate between two different speech segments, such as discriminating between a pair of speech frames that were close in time versus those that were far apart. The power of self-supervised training for Kazakh speech recognition lies in the ability to leverage large amounts of unlabeled data. This is particularly useful in languages where labeled data is scarce. By training a speech recognition model on unlabeled data using self-supervised techniques, the model can learn to recognize key features of the speech signal, such as phonemes and acoustic variations that are critical for accurate transcription. In addition to its potential to improve the accuracy of speech recognition systems, self-supervised learning can also reduce the amount of labeled data required for training. This can significantly reduce the cost and time needed to develop a robust Kazakh speech recognition system, making it more accessible to researchers and developers. Recent advancements in self-supervised audio encoders, such as Wav2Vec2.0 [6], have successfully learned high-quality audio representations. However, their unsupervised pre-training approach poses a challenge when it comes to decoding these representations into practical outputs. This necessitates a fine-tuning phase to effectively apply these models for tasks like ASR. To address this limitation, OpenAI researchers have recently introduced "Whisper" [7], a fully supervised sequence-to-sequence transformer.

Manuscript received May 30 2023; revised June 19, 2023; accepted July 6, 2023; published December 14, 2023.

Based on the reviewed literature, there are two main paradigms for ASR: self-supervised models, such as Wav2Vec2.0, and fully supervised models, such as Whisper. This study aimed to compare these two approaches to determine their ability to perform robust ASR for the Kazakh language. The main contributions of this study are as follows:

1. In addition to the existing speech corpora available for the Kazakh language, we have gathered audio recordings with corresponding texts from open sources. The total amount of collected data was approximately 1000 hours. Each audio file corresponds to a separate text file containing the content of an audiobook. It is important to note that the audio and texts are not synchronized, meaning they are not aligned at the sentence or word level. To overcome this misalignment challenge, we utilized a segmentation approach based on the Connectionist Temporal Classification (CTC) [8] algorithm. This method allowed us to accurately extract audio-text alignments.
2. Several experiments were conducted with Wav2Vec2.0 base and XLRS-53 architectures, pretraining and finetuning it in various scenarios.
3. Several experiments were conducted with Whisper architectures and finetuning it in various scenarios.
4. Two of the most accurate neural-based ASR architectures to date were extensively compared: Wav2Vec2.0 and Whisper. In addition, they were compared with the baseline E2E Transformer model. This evaluation is valuable for other languages, and our contribution is open and applicable to other scenarios. To the best of our knowledge, this is the first study to utilize above mentioned ASR solutions and their ability to recognize Kazakh speech.

The rest of the paper is organized as follows: Section II provides an in-depth discussion of the technical aspects of the Wav2Vec2.0, XLSR-53, and Whisper architectures in the context of Automatic Speech Recognition (ASR). Furthermore, this section offers an overview of recent advancements in Kazakh speech recognition. Detailed information about the dataset utilized in our experiments is described in Section III. The approaches of using Wav2Vec2.0, XLSR-53 and Whisper architectures as well as baseline end-to-end Transformer approach also presented in Section III. The results obtained from our conducted experiments are outlined in Section IV. Section V summarizes our findings and conclusions from the experiments conducted, and highlights potential areas for further research.

II. Background and Related Work

This section provides a brief overview of the related work to this article, which is categorized into four sections: Wav2Cec2.0, XLSR-53, Whisper and Kazakh ASR.

Wav2Vec 2.0. This model is designed to transcribe speech from audio signals, and it uses a self-supervised pre-training approach that allows it to learn from large amounts of unlabeled audio data. This particular model is an amalgamation of several earlier models, namely, Contrastive Predictive Coding (CPC) [5], Model Predictive Control (MPC) [9], wav2vec [10], and vqwav2vec [11]. Wav2Vec2 utilizes a combination of Convolutional Neural Networks (CNNs) and transformers, which enables it to capture both local and global patterns in audio data. The model utilizes a multi-layer convolutional feature encoder, denoted as $f: X \rightarrow Z$, to encode raw audio waveforms, X, into latent speech representations, $z_1, \ldots, z_T$, which are then fed into a transformer-masked network, denoted as $g: Z \rightarrow C$, that maps the representations from the latent space to a discrete set of outputs, $q_1, \ldots, q_T$, that represent targets in the self-supervised learning objective [6, 12]. The transformer module contextualizes the quantized representations using attention blocks, resulting in a set of discrete contextual representations, $c_1, \ldots, c_T$. The feature encoder is composed of seven convolutional blocks, each with 512 channels, kernel widths of {10, 3, 3, 3, 3, 2, 2}, and strides of {5, 2, 2, 2, 2, 2, 2}. On the other hand, the transformer network is made up of 24 blocks, with 1024 dimensions and inner dimensions numbering 4096. It also has a total of 16 attention heads.

This model has achieved impressive results on a number of benchmark datasets and has significantly advanced the state-of-the-art in speech recognition. An illustration of the model based on the one presented in [6] is shown in Fig. 1.

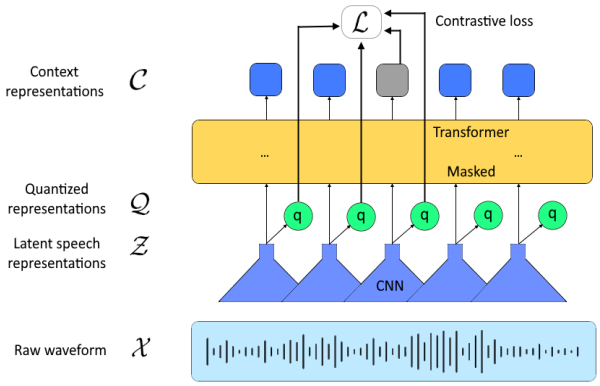

Figure 1. Wav2Vec2.0 architecture representation.

XLSR-53. It is a multilingual language model developed by Facebook AI Research [12]. It is an extension of the cross-lingual language model XLM-R, designed to handle multilingual and cross-lingual Natural Language Processing (NLP) tasks. Although XLSR-53 is built on the Wav2Vec 2.0 model, it can learn latent quantization that is spread across languages. XLSR-53 uses product quantization to select quantized representations from codebooks, which are then selected using the Gumbel-Softmax method in a completely distinguishable manner. XLSR-53's architecture is similar to that of Bidirectional Encoder Representations from Transformers (BERT) [13], with the exception that it has 53 language-specific embeddings, one for each language it supports. This means that the model can process data in different languages and understand their nuances, even when they have similar spellings or pronunciations. Additionally, XLSR-53 has 500 million parameters, making it one of the largest multilingual language models to date. The model is trained on a vast and diverse corpus

of speech text data from over 53 languages. XLSR-53's ability to understand multiple languages makes it particularly useful for cross-lingual transfer learning, where a model trained on one language can be adapted to another language with minimal additional training.

Whisper. OpenAI has introduced a new ASR system called Whisper [7]. Unlike Wav2Vec2.0, Whisper is trained using fully supervised methods, which involves using up to 680,000 h of labeled speech data from various sources. Thanks to the massive database and the training techniques that they used, the model can be a multilingual and multitask ASR system. The model was enhanced to add the multitask training format using a set of special tokens that serve as task specifiers or classification targets. As part of development, a sample of the previously transcribed text was fed back into the model so that it would learn from the context that accompanies the transcription. The structure of the model itself is not new, as it consists of an encoder-decoder Transformer that utilizes 80-channel log-Mel spectrograms. The encoder is composed of two convolution layers with a kernel size of 3, a sinusoidal positional encoding, and a series of stacked Transformer blocks. Meanwhile, the decoder utilizes learned positional embeddings and the same number of Transformer blocks as the encoder. An illustration of the Whisper architecture based on the one presented in [7] is shown in Fig. 2.

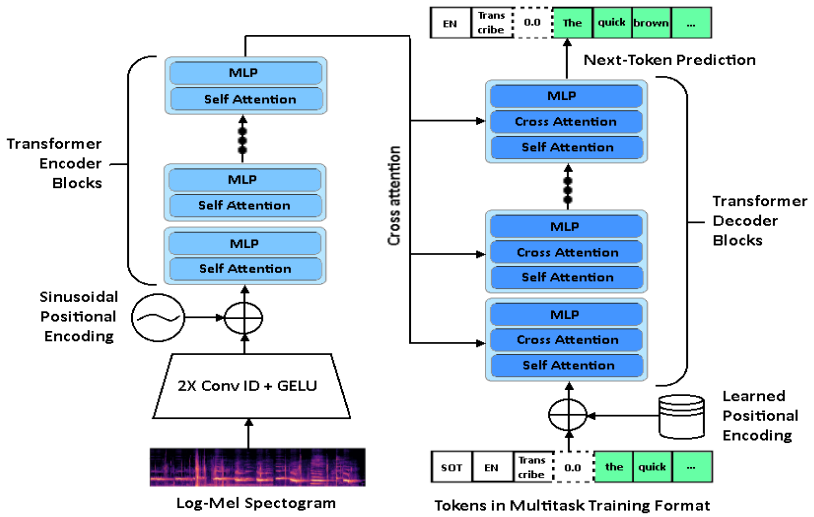

Figure 2. Whisper architecture representation.

Kazakh ASR. New developments in ASR have brought about innovative end-to-end structures, which demonstrate remarkable precision when provided with adequate datasets. The fundamental concept behind these end-to-end models is to map the speech signal input directly to character sequences. This simplifies the processes of training, fine-tuning, and inference making. Researchers in the field of ASR typically employ two distinct methods for training ASR systems: fully supervised or self-supervised models.

Regarding the first group, Yessenbayev *et al.* [14] conducted a comprehensive study to address the challenge of automatic, speaker-independent recognition of continuous Kazakh speech on a specific vocabulary basis in the presence of noise. According to the author, the proposed system achieved successful results in tasks such as phonetic recognition of English speech and recognition of continuous Kazakh speech, with a relative improvement in the recognition quality of up to 20%. Specifically, the recognition quality of Kazakh speech was 94.5%. Overall, this work serves as a starting point for the development of more advanced systems for continuous Kazakh speech recognition.

Mamyrbayev *et al*. [15] introduce stream speech recognition using the RNN-T model in their study. The architecture of the model is constructed using neural networks such as LSTM and BLSTM, and it was trained using over 300 h of prepared (reading) and spontaneous speech data. The study's findings show that the RNN-T model can achieve a CER of 10.6. In another study, Mamyrbayev *et al*. [16] introduce a combined Transformer + CTC LM model, which was trained on a 400-hour speech dataset. The study's findings indicate that the model achieved a CER of 3.7 and WER of 8.3.

It is also worth noting the joint work of researchers at the Center for Speech Technologies, St. Petersburg National Research University of Information Technologies, Mechanics and Optics and Kostanay State University named after Baitursynov [17]. The authors carried out work on the recognition and synthesis of the bilingual (Kazakh-Russian) language.

Khassanov *et al*. [18] have introduced the first comprehensive Kazakh database, KSC 1, which provides an open benchmark for Kazakh speech recognition research. The database comprises approximately 332 hours of transcribed audio, consisting of over 153,000 utterances spoken by individuals from various age groups, regions, and genders. According to the authors, a Transformer-based End-to-End (E2E) model yielded 2.8% CER and 8.7% WER on this dataset. Mussakhojayeva *et al*. [19] extended KSC to 1,128 h. Additional data from diverse sources, such as television news, television and

radio programs, parliament speeches, and podcasts were incorporated into the database. The authors specified the corpus specifications and verified its usefulness by employing a Transformer-based ASR model, which produced promising results. Specifically, the overall WER on the validation and test sets were 15.1% and 15.6%, respectively.

Although the models being developed have demonstrated exceptional performance, many of them rely on supervised training techniques, which demand a considerable amount of labeled data. Unfortunately, the process of data labeling and annotation is labor-intensive, expensive, and time-consuming, often requiring manual effort. Furthermore, there are situations where acquiring such data is not feasible due to limitations or inaccessibility. In contrast to fully supervised models, recent research has concentrated on employing big acoustic models trained through self-supervised learning techniques and a vast volume of unlabeled data. Meng and Yolwas [20] employed unsupervised pre-training using Wav2Vec2.0 and integrated a Factorized TDNN layer to preserve the relationship between voice and time steps, thereby enhancing speech recognition for the Kazakh language. Additionally, they utilized multi-language pre-training and speech synthesis to further improve performance. The results of the experiments indicated that incorporating unlabeled data from non-target languages and using data enhancement with speech synthesis significantly reduced word error rates on test sets.

This work compares self-supervised (Wav2Vec2.0) and fully supervised (Whisper) models for ASR in the Kazakh language. It explores various pretraining and finetuning scenarios, providing valuable insights for ASR in low-resource languages like Kazakh.

III. Materials and Method

This section focuses on datasets specifically designed for speech recognition in the Kazakh language, along with the approaches employed to develop accurate speech recognition modules.

A. Dataset

ISSAI KSC. The ISSAI KSC is the largest publicly accessible database created to support Kazakh speech and language processing applications [18]. It consists of over 332 h of data collected through a web-based speech recording platform, which invited volunteers to read sentences from various sources, including books, laws, Wikipedia, news portals, and blogs. The KSC dataset is diverse, featuring speakers and audio recordings from different regions of Kazakhstan and made using different devices such as smartphones, tablets, and laptops. The speakers are from five different regions, and the validation and test sets contain 51.7% female and 48.3% male speakers.

M2ASRKazakh-78. The speech corpus [21] created by Xinjiang University and made available through the M2ASR Free Data Program consists of 78 h of recordings obtained from 96 students using a variety of recording equipment. The recordings were made in quiet environments, and the speakers read the sentences in a scripted, reading style. The transcriptions are written using Latin characters and follow the rules established by the authors. However, there are some limitations to this corpus. All the speakers in the corpus share similar characteristics such as age group, social status, and region, which restricts its overall applicability. Moreover, the corpus only contains 4,000 sentences, and most of the participants recorded the same transcriptions, leading to a lack of diversity in linguistic content and speaker traits.

Kazcorpus. The kazcorpus acoustic corpus [22] consists of two independent subcorpuses—kazspeechdb and kazmedia. The kazspeechdb corpus was used as a starting point in creating the corpus for broadcast news. It is a set of speech fragments—12,675 sentences in Kazakh, voiced in studio conditions by speakers of different sexes, ages, from different regions of Kazakhstan. The size of the subcorpus is 22 h of speech. The subcorpus consists of 169 speakers, from them 73 male and 96 female voices. Each speaker read 75 sentences. The body of KazMedia is audio and text data collected from the official websites of the television news agencies, namely “Khabar”, “Astana TV” and “Channel 31”. Text data is the text of all news in the Kazakh language, published on the official websites of three TV channels. Audio data is wav files representing audio tracks extracted from a number of video news of these three TV channels in the Kazakh language. The total audio duration is 21 h of speech.

KazLibriSpeech. We have collected audio recordings with relevant texts from open sources. The amount of data collected was approximately 1,000 h. Each audio file corresponds to one common file with the text of the audiobook, i.e., audio and texts are not aligned either by sentences or by words. Therefore, the next task is to segment the audio file into smaller intervals (words, phrases or sentences). Further, each such interval must be compared with the corresponding text that was voiced in this interval. Alignment and segmentation can be a complex task, but this method allows to create large datasets from various sources and domains at the lowest possible cost.

The quality of the collected audiobooks varies and needs to be cleaned up and normalized to make sure that the alignment of subsequent segments is not broken. Cleaning and normalization of the text was carried out, including such processes as noise removal, work with homoglyphs, transliterator, removal of unreproducible fragments, replacement of acronyms and abbreviations with their full equivalents, normalization of numbers, replacement of characters with their verbal equivalents, initial segmentation at the chapter level and division of the source text into short sentences based on punctuation at the end of a sentence. The musical accompaniment at the beginning and at the end of the read audiobooks has been removed.

We employed a segmentation approach based on Connectionist Temporal Classification (CTC) algorithm to extract accurate audio-text alignments even when the audio recording includes unknown speech sections at the beginning or end. Our method employs an end-to-end

network trained on pre-aligned data using a CTC/attention ASR system. CTC is a type of neural network inference and associated scoring function for training recurrent neural networks to solve sequencing problems where time is a variable. CTC refers to results and scoring and is independent of the underlying structure of the neural network. In our case, this model defines speech segments in audio files in sentence level. The speech recognition model required for segmentation was trained using the ISSAI KSC dataset in Espnet tool [23].

More details of the available corpora for the speech recognition in the Kazakh language are shown in Table I.

TABLE I. THE STRUCTURE OF THE CORPORA FOR THE KAZAKH LANGUAGE

Structure	Name of the corpus/sets	Data type	Amount of wav-files	Overall duration of wav-files in hour
1	**ISSAI KSC**		**153853**	**332.6**
1.1	Train	Crowdsourced recordings	147236	318.4
1.2	Dev		3283	7.1
1.3	Test		3334	7.1
2	**M2ASRKazakh-78**		**37892**	**86**
2.1	Train	Reading-style recordings	34392	78
2.2	Test		3500	8
3	**Kazcorpus**		**13425**	**44.16**
3.1	kazspeechdb	Studio recordings, prepared speech	12675	22.61
3.1.1	Train		11175	19.92
3.1.2	Dev		750	1.36
3.1.3	Test		750	1.34
3.2	KazMedia	Mixed type: studio recordings, prepared speech + spontaneous speech in different acoustic conditions	740	21.55
3.2.1	Train		561	18.04
3.2.2	Dev		49	1.00
3.2.3	Test		130	2.51
4	**KazLibriSpeech**	Audio books	**575243**	**992**

B. Methods

In this study, two of the most precise neural-based ASR architectures currently available were evaluated: (1) Wav2Vec2.0, which is trained using a self-supervised paradigm; (2) Whisper, which is trained using a fully supervised strategy. Additionally, a baseline E2E Transformer; and (3) model was trained. Further information about each model can be found in the following sub-sections.

Baseline E2E Transformer. An encoder-decoder architecture based on the Transformer was trained using the ESPnet framework [23], jointly trained with the Connectionist Temporal Classification (CTC) objective function. The input speech was represented using 80-dimensional filterbank features with pitch computed every 10 ms over a 25 ms window. To process the acoustic features for the E2E architecture, a few initial blocks of VGG network were used. The E2E-Transformer ASR system was comprised of 12 encoder and 6 decoder blocks, with 4 heads in the self-attention layer, hidden states of 256-dimension, and feed-forward network dimensions of 2,048. The model was trained for 100 epochs using the Adam optimizer with an initial learning rate of 10 and warmup-steps of 30,000, with a dropout rate and label smoothing of 0.1 set. To assist decoding, a two-layer RNN with 650 Long Short-Term Memory (LSTM) units each was utilized to construct a character-level LM, based on the transcripts of the training set. Additionally, the LSTM LM was used during the decoding stage.

Wav2Vec 2.0. and XLSR-53. The experiments were conducted using the fairseq platform. The Wav2Vec 2.0 base model was pre-trained with unlabeled speech data using various configurations, such as encoder layerdrop set to 0.05, dropout_input, dropout_features, and feature_grad_mult set to 0.1, and encoder_embed_dim set to 768. The training hyperparameters included a learning rate of 5×10^{-4} and warmed up in the first 10% of the training time. The number of updates was set to 800,000, and the maximum quantity of tokens was set to 1,200,000. Additionally, the Adam optimizer was used, as in the original work. Standard fine-tuning procedures were used, with fine-tuning parameters defined using the following configurations: the number of updates was set to 160,000, and the maximum quantity of tokens was set to 2,800,000. The Adam optimizer was used, and other parameters included a learning rate of 3×10^{-5} and a gradient accumulation of 12 steps. The batch size during training was defined automatically by the framework, depending on the maximum quantity of predefined tokens. During training, the best model was selected based on the lowest WER obtained on the validation set.

The XLSR model was pre-trained using the same configurations as the Wav2Vec large model. The encoder block consisted of 24 layers with a dimension of 1024, and 16 attention blocks were used with no dropout. Fine-tuning parameters were defined using the same configurations used in the original Wav2Vec 2.0 experiment with XLSR.

After fine-tuning the model, decoding is performed using a 3-gram language model, which was trained using Kenlm on the KSC LM corpus. During decoding, a beam search decoder is used, with the beam size set to 1,500.

Whisper. The five configurations of Whisper checkpoints are available in varying model sizes. The smaller four are trained on either English-only or multilingual data. The largest checkpoint is trained exclusively on multilingual data. The multilingual version of the "small" checkpoint with 244 M parameters and the "large" checkpoint with 1,550 M parameters were fine-tuned with a warmup of 10% of the steps and a learning rate of 10^{-5}. The "small" model trained for 40,000 steps, while the "large" model trained for 30,000 steps.

IV. RESULT AND DISCUSSION

The E2E Transformer, Wav2Vec2.0, and Whisper models were assessed using the corpora outlined in Section III.A. Tables II–IV present the word error rate results of the ASR systems. Each architecture was trained in different scenarios, employing various parameters. However, for validation and testing purposes, only the development and test sets of the ISSAI KSC1 corpus were used consistently across all cases. The experiments were conducted on the NVIDIA DGX-1 server, which is equipped with 8 V100 GPUs.

By incorporating the ISSAI KSC1 and Kazcorpus datasets into the training process, substantial enhancements were observed in the performance of the E2E-Transformer model (Table II). The best results for CER and WER achieved by the E2E-Transformer on the test set were 2.8 and 8.7, respectively. Notably, the utilization of an LSTM language model had a significant positive impact on the E2E-Transformer model, leading to improved performance. Additionally, the application of data augmentation techniques such as SpeedPerturb and SpecAugment proved to be highly effective in enhancing the Kazakh E2E ASR, resulting in further improvements.

TABLE II. E2E-TRANSFORMER MODELS PERFORMANCE

ID	Dataset	LM	Speed-Perturb	Spec-Augment	Valid		Test	
					WER	CER	WER	CER
1	ISSAI KSC1	No	Yes	Yes	15.9	6.2	15.2	5.5
2	ISSAI KSC1	Yes	Yes	Yes	10.0	3.3	8.8	2.8
3	ISSAI KSC1 + Kaz-corpus	No	Yes	Yes	13.0	5.7	12.4	5.1
4	ISSAI KSC1 + Kaz-corpus	Yes	Yes	Yes	10.0	3.2	**8.7**	**2.8**

TABLE III. WAV2VEC 2.0 MODELS PERFORMANCE

ID	Initial model	Pretrain dataset	Finetune dataset	Evaluation	LM dataset	Valid	Test	
						WER	WER	CER
1	Wav2Vec 2.0 Base		ISSAI KSC1 (train+dev)	ISSAI KSC1 (test)	KenLM (ISSAI KSC1 + Kazcorpus + KazLibriSpeech)	10.5	9.71	2.69
2	Wav2Vec 2.0 Base	ISSAI KSC1 (train)	ISSAI KSC1 (dev)	ISSAI KSC1 (test)	KenLM (ISSAI KSC1 + Kazcorpus + KazLibriSpeech)	14.4	31.1	8.6
3	Wav2Vec 2.0 Base	KazLibriSpeech + M2ASRKazakh-78	ISSAI KSC1 (train+dev)	ISSAI KSC1 (test)	KenLM (ISSAI KSC1 + Kazcorpus + KazLibriSpeech)	8.7	10.9	3.2
4	Wav2Vec 2.0 Base	KazLibriSpeech + M2ASRKazakh-78	ISSAI KSC1 (train+dev) + Kazcorpus	ISSAI KSC1 (test)	KenLM (ISSAI KSC1 + Kazcorpus + KazLibriSpeech)	8.5	**9.8**	**2.7**
5	XLSR-53	KazLibriSpeech + M2ASRKazakh-78	ISSAI KSC1 (train+dev)	ISSAI KSC1 (test)	KenLM (ISSAI KSC1 + Kazcorpus + KazLibriSpeech)	12.4	13.5	4.3

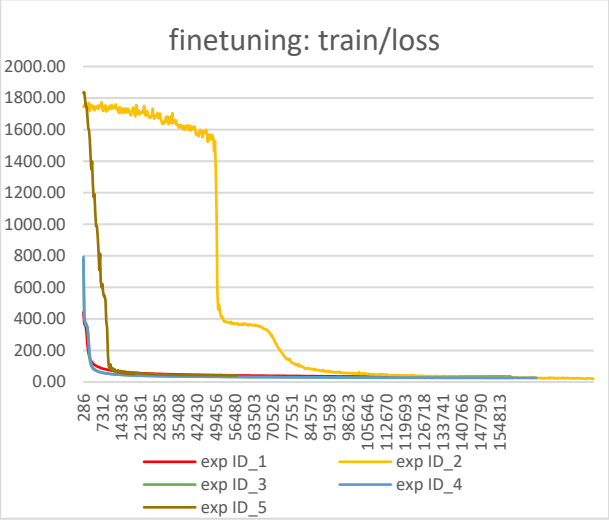

Figure 3. Wav2Vec 2.0 finetuning: train/loss (exp IDs refers to the ID column in Table III).

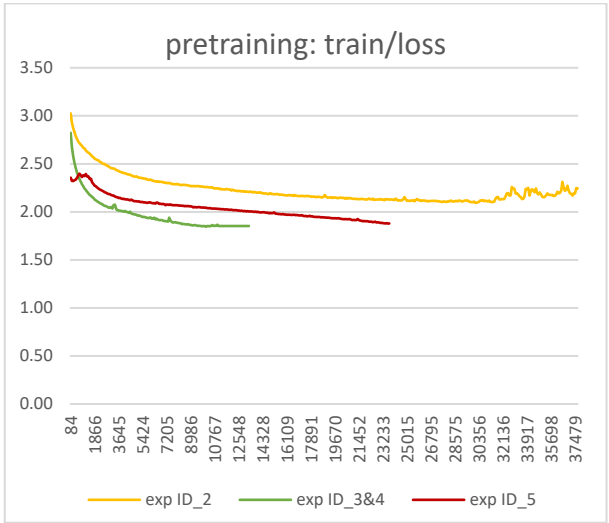

Figure 4. Wav2Vec 2.0 pretraining: train/loss (exp IDs refers to the ID column in Table III).

Table III presents the word error rate and character error rate scores of the fine-tuned Wav2Vec 2.0-base and XLS-R models. In the pretraining phase, the KazLibriSpeech and M2ASRKazakh-78 corpora were exclusively utilized, while the ISSAI KSC1 (train+dev) and Kazcorpus were employed for finetuning. The results demonstrate that the pre-trained Wav2Vec 2.0 Base model, which underwent pretraining using the KazLibriSpeech and M2ASRKazakh-78 corpora, followed by finetuning with the ISSAI KSC1 (train+dev) and Kazcorpus data, exhibits remarkable performance. It achieves a CER of 2.8 and a WER of 8.7 on the test set, which closely matches the best result achieved by the E2E-Transformer model. These findings indicate that pretraining significantly enhances

the model's performance, and the size of the dataset used for pretraining plays a crucial role. Moreover, the XLRS-53 model, trained on a vast and diverse corpus of speech text data from over 53 languages, also demonstrates competitive result. It is worth noting that the Language Model (LM) applied to all the models provides significant benefits, since it helps to refine the output of the models and reduce errors, leading to better results in terms of word error rate and character error rate. Figs. 3 and 4 depict the training loss progression during the pretraining and finetuning phases. These figures provide visual representations of how the loss values change over the course of training, offering insights into the optimization process of the models.

The pretrained Whisper models, utilizing the ISSAI KSC1 (train) and Kazcorpus datasets, exhibit competitive performance (Table IV). However, it is important to note that all models have a word error rate of at least 25, indicating that they do not transcribe Kazakh speech as accurately as desired without further adjustments. To improve the performance of the Whisper models, fine-tuning can be applied. When the small model is fine-tuned using only the ISSAI KSC1 (train) dataset, better results can be achieved. Specifically, a CER of 17.4 and a WER of 5.4 are obtained on the test set. Furthermore, the large model achieves a CER of approximately 4 on the test set. It is worth mentioning that the Language Model (LM) was not applied to the Whisper models. The absence of a language model may contribute to the lower transcription accuracy of the models. Fig. 5 illustrates the validation performance during the training, while Fig. 6 depicts the training loss progression.

TABLE IV. WHISPER MODELS PERFORMANCE

ID	Initial model	Finetune dataset	Validation	Evaluation	Valid	Test	
					WER	WER	CER
1	Multilingual small model	ISSAI KSC1 (train)	ISSAI KSC1 (dev)	ISSAI KSC1 (test)	16.4	**17.4**	**5.4**
2	Multilingual large model	ISSAI KSC1 (train)	ISSAI KSC1 (dev)	ISSAI KSC1 (test)	20.4	22.3	4.5
3	Multilingual large model	ISSAI KSC1 (train) + Kazcorpus	ISSAI KSC1 (dev)	ISSAI KSC1 (test)	18.1	19.8	4.1

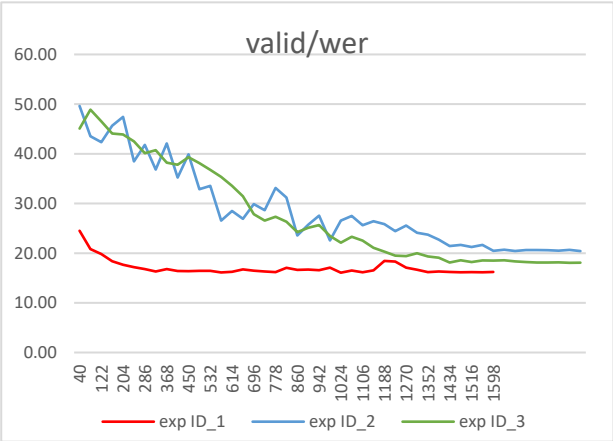

Figure 5. Whisper: valid/wer (exp IDs refers to the ID column in Table IV).

Figure 6. Whisper: train/loss (exp IDs refers to the ID column in Table IV).

V. CONCLUSION

This work focuses on evaluating the effectiveness of state-of-the-art speech recognition models in transcribing the Kazakh language, which is considered a low-resourced language. The models examined in the study include Facebook's Wav2Vec2.0 and Wav2Vec2-XLS-R, as well as OpenAI's Whisper. We compared the performance of these models with existing supervised Automatic Speech Recognition (ASR) systems and investigate the potential benefits of pre-training with data from other languages and fine-tuning with target language data. This study conducted various experiments with the Wav2Vec2.0 and Wav2Vec2-XLS-R architectures, exploring different pre-training and fine-tuning scenarios. Additionally, experiments were carried out with the Whisper architecture, focusing on fine-tuning it in various scenarios. One of the primary contributions of this study is the extensive comparison of two highly accurate neural-based ASR architectures, Wav2Vec2.0 and Whisper, with the baseline E2E Transformer model. This evaluation not only provides insights into the performance of these models for the Kazakh language but also holds value for other languages and scenarios. Notably, this study is the first to utilize the aforementioned ASR solutions to recognize Kazakh speech.

Overall, this research sheds light on the potential of utilizing advanced multilingual models and comparing self-supervised and fully supervised approaches for robust ASR in underresourced language settings. The findings and methodologies presented in this study can have broader applications in addressing language resource limitations and advancing the development of ASR systems for diverse languages.

CONFLICT OF INTEREST

The author declares no conflict of interest.

FUNDING

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP13068635).

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. Soplin, R. Yamamoto, X. Wang, and S. Watanabe, "A comparative study on transformer vs RNN in speech applications," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[3] T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *Proc. Interspeech 2019*, 2019, pp. 1408–1412. doi: 10.21437/Interspeech.2019-1938

[4] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[5] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint, arXiv:1807.03748, 2018.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[7] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint, arXiv:2212.04356, 2022.

[8] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for German end-to-end speech recognition," in *Proc. the 22nd International Conference on Speech and Computer, SPECOM 2020*, St. Petersburg, Russia, October 7–9, 2020, pp. 267–278.

[9] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," arXiv preprint, arXiv:1910.09932, 2019.

[10] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint, arXiv:1904.05862, 2019.

[11] A. Baevski, S. Schneider, and M. Auli, "VQ-wav2vec: Self-supervised learning of discrete speech representations," arXiv preprint, arXiv:1910.05453, 2019.

[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," arXiv preprint, arXiv:2006.13979, 2020.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018.

[14] Z. Yessenbayev, M. Karabalayeva, and F. Shamayeva, "Large vocabulary continuous speech recognition for Kazakh," in *Proc. the International Conference on Computer Processing of Turkic Languages*, Astana, 2013, pp. 217–221.

[15] O. Mamyrbayev, D. Oralbekova, A. Kydyrbekova, T. Turdalykyzy, and A. Bekarystankyzy, "End-to-end model based on RNN-T for Kazakh speech recognition," in *Proc. the 2021 3rd International Conference on Computer Communication and the Internet (ICCCI)*, 2021, pp. 163–167.

[16] O. Mamyrbayev, D. Oralbekova, K. Alimhan, and B. Nuranbayeva, "Hybrid end-to-end model for Kazakh speech recognition," *International Journal of Speech Technology*, pp. 1–10, 2022.

[17] O. Khomitsevich, V. Mendelev, N. Tomashenko, S. Rybin, I. Medennikov, and S. Kudubayeva, "A bilingual Kazakh-Russian system for automatic speech recognition and synthesis," in *Proc. the 17th International Conference on Speech and Computer, SPECOM 2015*, Athens, Greece, September 20–24, 2015, pp. 25–33.

[18] Y. Khassanov, S. Mussakhojayeva, A. Mirzakhmetov, A. Adiyev, M. Nurpeiissov, and H. Varol, "A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline," arXiv preprint, arXiv:2009.10334, 2020.

[19] S. Mussakhojayeva, Y. Khassanov, and H. Varol, "KSC2: An industrial-scale open-source Kazakh speech corpus," in *Proc. the Interspeech*, Incheon, Republic of Korea, 2015, pp. 18–22.

[20] W. Meng and N. Yolwas, "A study of speech recognition for Kazakh based on unsupervised pre-training," *Sensors*, vol. 23, no. 2, 870, 2023.

[21] Y. Shi, A. Hamdullah, Z. Tang, D. Wang, and T. Zheng, "A free Kazakh speech database and a speech recognition baseline," in *Proc. the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 745–748.

[22] O. Makhambetov, A. Makazhanov, Z. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, and A. Sharafudinov, "Assembling the Kazakh language corpus," in *Proc. the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1022–1031.

[23] S. Watanabe, T. Hori, S, Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, and A. Renduchintala, "EspNet: End-to-end speech processing toolkit," arXiv preprint, arXiv:1804.00015, 2018.