

Leveraging Wav2Vec2.0 for Kazakh Speech Recognition: An Experimental Study

Zhanibek Kozhirbayev(✉)

National Laboratory Astana, Nazarbayev University, Astana 01000, Kazakhstan
zhanibek.kozhirbayev@nu.edu.kz

Abstract. In the fast-growing world of neural networks, models trained on extensive multilingual text and speech data have shown great promise for improving the state of low-resource languages. This study focuses on the application of state-of-the-art speech recognition models, specifically Facebook's Wav2Vec2.0 and Wav2Vec2-XLSR, to the Kazakh language. The primary objective is to evaluate the performance of these models in transcribing spoken Kazakh content. Additionally, the research explores the possibility of using data from other languages for initial training and examines whether fine-tuning the model with target language data can improve its performance. More so, this work gives insights into how effective pre-trained multilingual models are when used on low-resource languages. The fine-tuned wav2vec2.0-XLSR model demonstrated impressive results, achieving a character error rate (CER) of 1.9 and a word error rate (WER) of 8.9 when tested against the test set of the Kazcorpus dataset. These findings may help create robustness in Automatic Speech Recognition (ASR) systems for Kazakh which could be used for various applications such as voice-activated assistants; speech-to-text translators among others.

Keywords: Automatic speech recognition · Kazakh language · Wav2Vec 2.0 · Wav2Vec2-XLSR · Pre-trained transformer models · Speech representation models

1 Introduction

Recently, sequence-based models have demonstrated remarkable advances in speech recognition compared to traditional automatic speech recognition systems. These sequence-based models use neural networks to convert speech into text, which simplifies the modeling process. Among them, Transformer [1] stands out as widely used and has demonstrated noteworthy achievements in the creation of end-to-end speech recognition systems [2–4]. Despite the substantial headway in the advancement of ASR models, the development of reliable models for languages other than English remains a challenge. This problem is mainly related to the fact that in order to obtain satisfactory results using advanced models, it usually takes many hours of work with annotated speech data for training. This is especially true of Kazakh, the Turkic language spoken by the global community of more than 13 million people.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14814, pp. 120–132, 2024.
https://doi.org/10.1007/978-3-031-64608-9_8

Recent advances in the development of self-learning methodologies have demonstrated the potential to address the limited availability of language data in the absence of sufficient resources. Self-supervised is a unique approach to improving speech recognition systems, allowing to use a huge amount of unlabeled speech data to obtain a valuable representation of speech signals. Unlike traditional supervised learning, in which labeled data is required to train the model, self-supervised learning algorithms extract knowledge from raw data without the need for explicit annotations. As part of self-learning, models are honed to perform tasks that are closely related to the main task of speech recognition, but do not require labeled input. An example involves training a model to predict the next frame of a speech signal based on previous frames—a concept called contrastive predictive coding (CPC) [5]. Another strategy involves training the model to distinguish between two different speech segments, for example, to distinguish a pair of speech frames that are close in time from those that are significantly separated from each other. The potential of self-supervised training to recognize Kazakh speech lies in its ability to benefit from significant amounts of unlabeled data, which is an important advantage for languages where annotated data is still scarce. Using self-supervised techniques to train a speech recognition model based on unlabeled data, the model can skillfully decode key aspects of the speech signal, including phonemes and acoustic nuances necessary for accurate transcription. In addition to the ability to improve the accuracy of speech recognition systems, self-supervised method promises to reduce the amount of labeled data needed for learning. This reduction can significantly reduce the costs and time spent on creating a reliable Kazakh speech recognition system, which will make it more accessible to researchers and developers. Recent advances in self-supervised audio encoders, such as Wav2Vec2.0 [6], have made it possible to effectively assimilate high-quality audio representations. However, their unsupervised approach to pre-learning makes it difficult to effectively translate these ideas into practical results. Therefore, a fine-tuning step is required to effectively apply these models to tasks such as automatic speech recognition (ASR). The motivation of my work can be outlined as follows:

– Bridging the language gap: This experimental study is motivated by the need to bridge this gap and improve ASR capabilities for languages with limited resources.
– Use of unlabeled data: This study builds on the potential of self-learning techniques to use valuable information about unlabeled speech data. Moreover, the key motivation for this study is to study the effectiveness of self-supervised learning in reducing dependence on labeled data for training ASR models.

Two iterations of Wav2Vec2.0 pre-trained models exist: the first is solely pre-trained for a single language, and the second, is a multilingual pre-trained (XLSR-53) model. This study was undertaken to contrast these two approaches to ascertain their efficacy in facilitating robust ASR for the Kazakh language. The primary contributions of this study encompass:

1. Supplementary to the available Kazakh language speech corpora, we procured audio recordings along with their corresponding texts from publicly accessible sources. The aggregate data amassed equated to around 1000 h. Each audio file corresponds to an individual text file containing audiobook content. Notably, it is imperative to

acknowledge that synchronization was absent between the audio and text, implying a lack of alignment at the sentence or word level. To address this challenge of misalignment, we employed a segmentation technique grounded in the connectionist temporal classification (CTC) [7] algorithm. This approach facilitated the precise extraction of audio-text alignments.
2. A series of experiments were conducted utilizing the Wav2Vec2.0 base and XLSR-53 architectures, encompassing various pre-training and fine-tuning scenarios.
3. An extensive comparative analysis was performed between two methodologies: Wav2Vec2.0 and Wav2Vec2.0-XLSR.

This paper is structured as follows: Sect. 2 delves into a comprehensive examination of the technical nuances underlying the Wav2Vec2.0 and XLSR-53 architectures within the context of ASR. Additionally, this section provides insights into recent advancements in the domain of Kazakh speech recognition. The particulars concerning the dataset employed in our experimentation are detailed in Sect. 3. The strategies involving Wav2Vec2.0 and XLSR-53 architectures are expounded upon in the same section. Section 4 outlines the outcomes derived from the conducted experiments. Finally, Sect. 5 encapsulates our discoveries and conclusions from the undertaken experiments, shedding light on potential avenues for future exploration.

2 Background and Related Work

This section furnishes a concise outline of the pertinent literature related to this article, classified into two subsections: Wav2Vec2.0 and XLSR-53, and Kazakh ASR.

2.1 Wav2Vec 2.0 and XLSR-53

Wav2Vec 2.0. is meticulously crafted for transcribing speech embedded within audio signals, leveraging a self-supervised pre-training methodology that imbibes insights from vast volumes of unlabeled audio data. It amalgamates principles from several precursor models, namely, Contrastive Predictive Coding (CPC) [5], Model Predictive Control (MPC) [8], wav2vec [9], and vqwav2vec [10]. The architecture of Wav2Vec2 harmonizes convolutional neural networks (CNNs) and transformers, enabling it to apprehend both local nuances and overarching patterns within 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 16 attention heads.

Demonstrating its prowess, this model has achieved remarkable feats across various benchmark datasets, substantially propelling the boundaries of speech recognition technology.

XLSR-53, a product of Facebook AI Research [11], is a multilingual language model with a foundation in cross-lingual language model XLM-R, tailored to cater to multilingual and cross-lingual natural language processing (NLP) tasks. Building upon the Wav2Vec 2.0 model, XLSR-53 possesses the capacity to acquire latent quantization spanning various languages. This is accomplished through the utilization of product quantization to cherry-pick quantized representations from codebooks. The selection process employs the Gumbel-Softmax technique, ensuring complete distinction. The architecture of XLSR-53 bears resemblance to that of Bidirectional Encoder Representations from Transformers (BERT) [12], with a notable deviation: it encompasses 53 language-specific embeddings, catering to each of the supported languages. This intricate design empowers the model to process diverse languages, capturing their subtleties even in cases of similar spellings or pronunciations. Furthermore, XLSR-53 boasts an expansive parameter count of 500 million, positioning it among the largest multilingual language models available. This model is trained on an extensive and diverse corpus comprising speech text data extracted from over 53 languages. The inherent ability of XLSR-53 to comprehend multiple languages renders it exceptionally advantageous for cross-lingual transfer learning. This entails adapting a model trained in one language to perform well in another language, requiring only minimal supplementary training.

2.2 Kazakh ASR

Recent advances in ASR have led to the emergence of new end-to-end architectures that provide impressive accuracy when sufficient data sets are available.. The basic principle underlying these end-to-end models is the direct conversion of input speech signals into sequences of characters. This streamlined approach simplifies learning, fine-tuning, and logical inference procedures. In the field of ASR research, experts predominantly gravitate towards two distinct methodologies for training ASR systems: fully supervised and self-supervised models.

In the context of the first category, Yessenbayev et al. [13] undertook an extensive study aimed at surmounting the challenge of automatic, speaker-independent recognition of continuous Kazakh speech, focusing on a specific vocabulary foundation, while contending with noisy environments. The authors' proposed system exhibited commendable outcomes in diverse tasks, encompassing phonetic recognition of English speech as well as the recognition of continuous Kazakh speech. Notably, the system displayed a relative enhancement in recognition quality of up to 20%. Noteworthy is the achievement of a recognition quality of 94.5% for Kazakh speech. In essence, this research stands as a foundational step, laying the groundwork for the subsequent development of more advanced frameworks catering to the domain of continuous Kazakh speech recognition.

Kozhirbayev [14] 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. This work can provide insights into the effectiveness of using pretrained multilingual models in under esourced 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. Kozhirbayev and Islamgozhayev [15] first created a dedicated speech translation dataset ST-kk-ru based on the ISSAI Corpus to develop the cascade speech translation system. The ST-kk-ru dataset comprises a large collection of Kazakh speech recordings along with their corresponding Russian translations. The automatic speech recognition (ASR) module of the system utilizes deep learning techniques to convert spoken Kazakh input into text. The machine translation (MT) module employs state-of-the-art neural machine translation methods, leveraging the parallel Kazakh-Russian translations available in the dataset to generate accurate translations.

In their research, Mamyrbayev et al. [16] delve into stream speech recognition through the implementation of the RNN-T model. This architecture is structured with neural networks like LSTM and BLSTM, leveraging a training dataset of over 300 h comprising both prepared (reading) and spontaneous speech recordings. The outcomes of the study underscore the RNN-T model's capability to attain a CER of 10.6.

In a distinct study by the same authors, Mamyrbayev et al. [17] introduce a hybrid Transformer + CTC (Connectionist Temporal Classification) model. It was honed using a speech dataset spanning 400 h. Notably, the study's findings spotlight the model's performance, registering a CER of 3.7 and a WER of 8.3.

Certainly, it's noteworthy to highlight the collaborative efforts of researchers from the Center for Speech Technologies at St. Petersburg National Research University of Information Technologies, Mechanics, and Optics, along with Kostanay State University named after A. Baitursynov [18]. These researchers embarked on a project centered around the recognition and synthesis of the bilingual (Kazakh-Russian) language. Their endeavors focused on advancing the field of bilingual speech processing, thereby contributing to the broader spectrum of language technology research.

Khassanov et al. [19] have introduced a significant contribution to Kazakh speech recognition research in the form of the KSC 1 database. This comprehensive database serves as an open benchmark, constituting around 332 h of transcribed audio. The database encompasses over 153,000 utterances spoken by individuals from diverse age groups, regions, and genders. The authors employed a Transformer-based end-to-end (E2E) model, achieving a Character Error Rate (CER) of 2.8% and a Word Error Rate (WER) of 8.7% on this dataset. In a related development, Mussakhojayeva et al. [20] extended the KSC database to a staggering 1128 h. This augmentation involved the incorporation of additional data from diverse sources, including television news, television and radio programs, parliamentary speeches, and podcasts. The authors meticulously defined the corpus specifications and substantiated its utility by employing a Transformer-based ASR model. This model yielded promising outcomes, with an overall Word Error Rate (WER) of 15.1% on the validation set and 15.6% on the test set. These efforts significantly bolster the resources available for advancing Kazakh speech recognition technology.

Although the ongoing development of models has demonstrated remarkable performance, a significant portion of them rely heavily on supervised learning techniques, which requires significant amounts of annotated data. Unfortunately, the process of collecting and annotating data is resource-intensive, expensive and time-consuming, often

requiring manual intervention. Moreover, there may be circumstances in which obtaining such data becomes impractical due to limitations or unavailability. Unlike fully controlled models, recent research has focused on the use of powerful acoustic models prepared using self-supervised methods and extensive amount of unlabeled data. An example is the work of Meng and Yolvas [21], who implemented unsupervised pretraining using Wav2Vec2.0. They integrated the factorized TDNN level to maintain the connection between voice and time steps, thereby increasing the effectiveness of speech recognition in the Kazakh language. In addition, they used multilingual pre-training and speech synthesis techniques to further enhance productivity. The results of their experiments highlighted the benefits of assimilating unlabeled data from languages other than the target language and using data enhancement techniques using speech synthesis. These approaches, in particular, have led to a significant reduction in the frequency of errors in words in test suites. This study indicates significant progress in optimizing speech recognition systems while reducing dependence on extensive labeled datasets.

3 Materials and Method

This section is dedicated to data sets specially designed for speech recognition in the Kazakh language. It also examines the methodologies used to develop accurate speech recognition modules designed for a given linguistic context.

3.1 Dataset

In these experiments, several sets of data on the Kazakh language were used. Some of them were used for pre-training phase, while others were used for fine tuning phase. ISSAI KSC [19] and Kazcorpus [22] were selected for fine-tuning and evaluation due to their comprehensive coverage, diverse linguistic content and suitability for solving specific tasks. They offer a rich variety of linguistic data, which makes them ideal for fine-tuning models to effectively solve target tasks. On the other hand, KazLibriSpeech was chosen for pre-training because of its large-scale audio data, which can facilitate the study of high-level acoustic characteristics and improve the performance of the model in speech-related tasks.

ISSAI KSC. The ISSAI KSC database is the most extensive open resource created to support Kazakh speech and language data processing applications. This substantial dataset contains over 332 h of content collected using a web-based platform designed for speech recording. this platform invited volunteers to formulate suggestions taken from various sources, including books, laws, Wikipedia, news portals and blogs. The KSC dataset is diverse and includes speakers and audio recordings from various regions of Kazakhstan, which use a variety of devices such as smartphones, tablets and laptops. The speaker group consists of representatives from five different regions, with 51.7% of female speakers and 48.3% of male speakers in the screening and testing kits.

Kazcorpus. The kazcorpus acoustic corpus includes two separate subcorpuses: kazspeechdb and kazmedia. The foundation for the creation of the broadcast news corpus was laid using the kazspeechdb corpus. This subcorpus consists of fragments of speech,

namely 12,675 sentences spoken in Kazakh, recorded in controlled studio conditions. The speakers represent different genders, ages and regions of Kazakhstan. In total, the performance of the subgroup lasted 22 h, 169 speakers took part in it, including 73 male and 96 female voices. Each speaker delivered 75 sentences.

On the other hand, the KazMedia subcorporation combines audio and text data collected from the official websites of television news agencies, in particular Khabar, Astana TV and Channel 31. Text data represents the entire set of news materials in the Kazakh language published on the official websites of these channels. Audio data consists of WAV format files, which are audio tracks extracted from various news videos broadcast on these channels in the Kazakh language. Collectively, this subcorpus includes 21 h of speech.

KazLibriSpeech. We have collected audio recordings in combination with relevant texts from open sources, accumulating a total of about 1,000 h of data. Each audio file corresponds to the corresponding text contained in the audiobook, although there is no alignment at the sentence or word level. Therefore, the subsequent attempt is to segment these audio files into smaller intervals, whether they are words, phrases, or sentences. Then each such segment must be compared with the corresponding voiced text that was created during the same interval. Although the alignment and segmentation process can be complex, this method allows to create extensive datasets covering various sources and areas with minimal cost.

Given that the quality of the collected audiobooks varies, the cleaning and normalization process has been launched. This included removing noise, working with homoglyphs, transliteration, extracting non-reproducible fragments, replacing abbreviations and abbreviations with their full forms, normalizing numerical composition, replacing characters with their phonetic counterparts, initial chapter-level segmentation and dividing the source text into compressed sentences using punctuation marks. In addition, all the musical elements accompanying the audiobook at the beginning and at the end were removed. Our segmentation approach involved the use of the CTC algorithm, which ensures accurate alignment of the audio text, even if the audio recording contains incomprehensible fragments of speech at the beginning or end. Our method is to train an end-to-end network based on pre-aligned data using a CTC/attention based ASR system. CTC, as a mechanism for neural network inference and scoring, plays an important role in training recurrent neural networks to solve sequence-based tasks, taking into account variable time. This mechanism does not depend on the basic structure of the neural network. In our context, this model identifies speech segments in audio files at the sentence level. The ASR model requisite for segmentation was honed using the ISSAI KSC dataset and the Espnet tool [23].

For a comprehensive overview of the corpora designed for speech recognition in the Kazakh language, please consult Table 1. This table presents the dataset structures, the division of the dataset into training, testing, and development sets, and the methodologies employed for dataset acquisition, the number of WAV files, the aggregate duration of WAV files in hours.

Table 1. 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**	Crowdsourced recordings	**153853**	**332.6**
1.1	Train		147236	318.4
1.2	Dev		3283	7.1
1.3	Test		3334	7.1
2	**Kazcorpus**	Mixed type: studio recordings, prepared speech + spontaneous speech in different acoustic conditions	**13425**	**44.16**
2.1	kazspeechdb		12675	22.61
2.1.1	Train		11175	19.92
2.1.2	Dev		750	1.36
2.1.3	Test		750	1.34
2.2	KazMedia		740	21.55
2.2.1	Train		561	18.04
2.2.2	Dev		49	1.00
2.2.3	Test		130	2.51
3	**KazLibriSpeech**	Audio books	**575243**	**992**

3.2 Methods

In the context of this research, an evaluation was conducted on two variants of wav2vec: (1) Wav2Vec 2.0, pre-trained and fine-tuned exclusively for the Kazakh language, and (2) XLSR-53, initially pre-trained across 53 languages, followed by continuous pre-training and fine-tuning for the Kazakh language.

Wav2Vec 2.0. The experimentation was carried out utilizing the Fairseq platform [24]. The Wav2Vec 2.0 base model underwent pre-training using varied configurations, encompassing an encoder layerdrop set to 0.05, dropout_input, dropout_features, feature_grad_mult set to 0.1, and an encoder_embed_dim established at 768. Training hyperparameters encompassed a learning rate of 5×10^{-4}, with a warm-up phase in the initial 10% of the training duration. The number of updates was specified as 800,000, while the maximum token count was set at 1,200,000. Consistently, the Adam optimizer was applied, in line with the original approach.

For fine-tuning, conventional procedures were followed, with parameters set as follows: the number of updates reached 160,000, while the maximum token count stood at 2,800,000. Similar to pre-training, the Adam optimizer was engaged, utilizing a learning rate of 3×10^{-5} and a gradient accumulation of 12 steps. The training batch size was dynamically determined by the framework, factoring in the predefined token maximum. Throughout training, the optimal model was chosen based on the lowest WER achieved on the validation set.

XLSR-53. The XLSR model underwent pre-training with identical configurations as those employed for the Wav2Vec large model. The encoder block encompassed 24 layers, each with a dimension of 1024, and a set of 16 attention blocks was utilized, without incorporating dropout. Fine-tuning parameters were determined in line with the configurations applied in the original XLSR experiment with Wav2Vec 2.0.

Language Model. Subsequent to the fine-tuning process, the model undergoes decoding, facilitated by a 3-g language model [25]. The language model was trained using the transcriptions from all the accessible datasets outlined in Table 1, utilizing the Kenlm toolkit [26]. For decoding purposes, a beam search decoder is employed, with the beam size configured to 1500.

4 Result and Discussion

The evaluation of Wav2Vec2.0 models utilized the datasets described in Sect. 3.1. The ASR systems' performance, as indicated by WER and CER scores, is presented in Table 2. Various training scenarios were employed for each architecture with different parameters. The experiments took place on the NVIDIA DGX-1 server, equipped with 8 V100 GPUs.

Table 2 displays the character error rate and word error rate scores of the fine-tuned Wav2Vec 2.0-base and XLSR-53 models. During the pre-training phase, the KazLibriSpeech corpus was exclusively employed, while the ISSAI KSC1 and Kazcorpus were used for fine-tuning. Results highlight the exceptional performance of the pre-trained XLSR-53 model. After undergoing pre-training with the KazLibriSpeech corpus and fine-tuning with the Kazcorpus (train + dev) data, it achieves a CER of 1.9 and a WER of 8.9 on the test set. These figures are notably higher by 28.2% and 26.9% compared to the Wav2Vec 2.0-base model in terms of WER and CER, respectively. These findings underscore the significant enhancement in model performance due to pre-training, where the dataset size used for pre-training plays a pivotal role. Conversely, the Wav2Vec 2.0-base model, having undergone pre-training with the KazLibriSpeech corpus and fine-tuning with the ISSAI KSC1 (train + dev) data, exhibits superior results compared to the XLSR-53 model. This occurrence might have been triggered due to the fact that the initial model underwent pre-training using a dataset containing audiobook content.

Table 2. Wav2Vec 2.0 models performance.

ID	Initial model	Pretrain dataset	Finetune dataset	Evaluation set	LM dataset	Test	
						WER	CER
1	Wav2Vec 2.0 Base	KazLibriSpeech	Kazcorpus (train + dev)	Kazcorpus (test)	3-g KenLM	12.4	2.6
2	Wav2Vec 2.0 Base	KazLibriSpeech	ISSAI KSC1 (train + dev)	ISSAI KSC1 (test)	3-g KenLM	10.1	2.8
3	XLSR-53	KazLibriSpeech	Kazcorpus (train + dev)	Kazcorpus (test)	3-g KenLM	8.9	1.9
4	XLSR-53	KazLibriSpeech	ISSAI KSC1 (train + dev)	ISSAI KSC1 (test)	3-g KenLM	15.1	4.8

The training progression during the pre-training and fine-tuning phases is illustrated in Figs. 1 and 2. These visuals depict the changes in loss values throughout training, offering insights into the optimization process of the models.

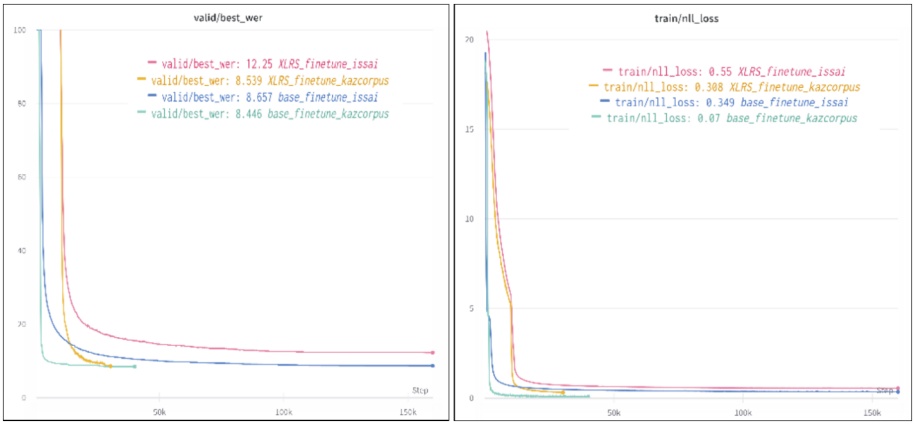

Fig. 1. Wav2Vec 2.0 fine-tuning: valid/best_wer (left) and train/loss (right).

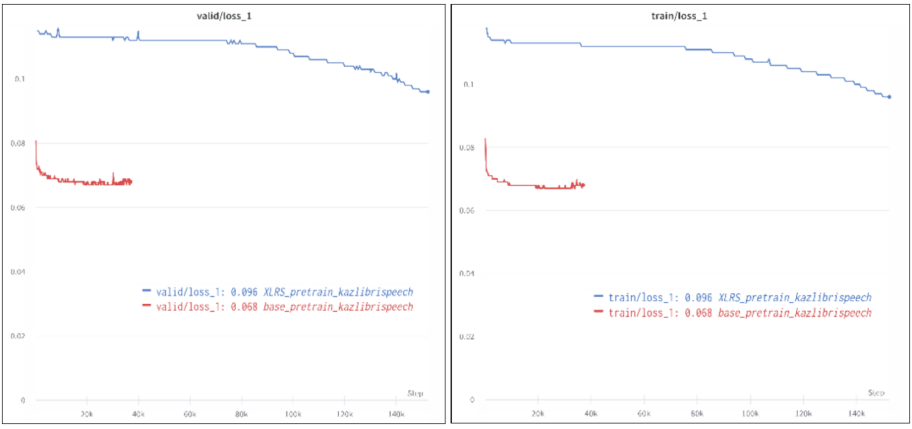

Fig. 2. Wav2Vec 2.0 pre-training: valid/loss (left) and train/loss (right).

5 Conclusion

The main purpose of this study is to evaluate the effectiveness of advanced speech recognition models in transcribing the Kazakh language, which belongs to the category of languages with a low level of resources. Our analysis includes a comprehensive comparison of these models, taking into account the nature and volume of data used during the pre-training and fine-tuning stages. In addition, the study aims to uncover the potential benefits of pre-training using data from other languages, followed by fine-tuning using data from the target language. The study includes a series of experiments using the Wav2Vec2.0 and Wav2Vec2-XLS-R architectures, during which various scenarios of pre-training and fine-tuning are studied. Thanks to these experiments, we not only get an idea of how these models work specifically for the Kazakh language, but also identify the consequences applicable to other languages and settings.

In fact, this study highlights the prospects for using advanced multilingual models and comparing self-supervised and fully supervised methods for reliable ASR in conditions of limited language resources. The findings and methodologies presented in this study have broader implications, including reducing the limitations of language resources and advancing the development of an ASR system for a wide range of languages.

Acknowledgments. 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. Vaswani, A., et al.: Attention is all you need. In: Advances in neural information processing systems, vol. 30 (2017)
2. Karita, S., et al.: A comparative study on transformer vs rnn in speech applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 449–456 (2019). https://doi.org/10.1109/ASRU46091.2019.9003750

3. Nakatani, T.: Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In: Proceedings of Interspeech (2019). https://doi.org/10.21437/Interspeech.2019-1938
4. Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5888 (2018). https://doi.org/ https://doi.org/10.1109/ICASSP.2018.8462506
5. Oord, A. V. D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
6. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020)
7. Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., Rigoll, G.: CTC-segmentation of large corpora for german end-to-end speech recognition. In: Proceedings of Speech and Computer: 22nd International Conference, SPECOM 2020, pp. 267–278 (2020). https://doi.org/10.1007/978-3-030-60276-5_27
8. Jiang, D., et al.: Improving transformer-based speech recognition using unsupervised pre-training, arXiv preprint arXiv:1910.09932 (2019)
9. Schneider, S., Baevski, A., Collobert, R., Auli, M: wav2vec: unsupervised pre-training for speech recognition. In: Proceedings of Interspeech, pp. 3465–3469 (2019). https://doi.org/10.21437/Interspeech.2019-1873
10. Baevski, A., Schneider, S., Auli, M.: vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv preprint arXiv:1910.05453 (2019)
11. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. In: Proceedings of Interspeech, pp. 2426–2430 (2021). https://doi.org/10.21437/Interspeech.2021-329
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019). https://doi.org/10.21437/10.18653/v1/N19-1423
13. Yessenbayev, Z., Karabalayeva, M., Shamayeva, F.: Large vocabulary continuous speech recognition for Kazakh. In: Proceedings of the I International Conference on Computer processing of Turkic Languages, Astana, pp. 217–221 (2013)
14. Kozhirbayev, Z.: Kazakh speech recognition: Wav2vec 2.0 vs. Whisper. J. Adv. Inform. Technol. **14**(6), 1382–1389 (2023). https://doi.org/10.21437/10.12720/jait.14.6.1382-1389
15. Kozhirbayev, Z., Islamgozhayev, T.: Cascade speech translation for the kazakh language. Appl. Sci. **13**(15), 8900 (2023). https://doi.org/10.3390/app13158900
16. Mamyrbayev, O., Oralbekova, D., Kydyrbekova, A., Turdalykyzy, T., Bekarystankyzy, A.: End-to-end model based on RNN-T for Kazakh speech recognition. In: 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), pp. 163–167 (2021). https://doi.org/10.1109/ICCCI51764.2021.9486811
17. Mamyrbayev, O., Oralbekova, D., Alimhan, K., Nuranbayeva, B.: Hybrid end-to-end model for Kazakh speech recognition. Inter. J. Speech Technol., 1–10 (2022). https://doi.org/10.1007/s10772-022-09983-8
18. Khomitsevich, O., Mendelev, V., Tomashenko, N., Rybin, S., Medennikov, I., Kudubayeva, S.: A bilingual Kazakh-Russian system for automatic speech recognition and synthesis. In: Proceedings of Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, 20–24 September, pp. 25–33 (2015). https://doi.org/10.1007/978-3-319-23132-7_3

19. Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M., Varol, H.: A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 697–706 (2021). https://doi.org/10.18653/v1/2021.eacl-main.58
20. Mussakhojayeva, S., Khassanov, Y., Varol, H.: KSC2: an industrial-scale open-source Kazakh speech corpus. In: Proceedings of the INTERSPEECH, Incheon, Republic of Korea, pp. 18–22 (2015). https://doi.org/10.21437/Interspeech.2022-421
21. Meng, W., Yolwas, N.: A study of speech recognition for kazakh based on unsupervised pre-training. Sensors **23**(2), 870 (2023). https://doi.org/10.3390/s23020870
22. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A.: Assembling the kazakh language corpus. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1022–1031 (2013)
23. Watanabe, S., et al.: Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015 (2018)
24. Ott, M., et al.: fairseq: a fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 48–53 (2019)
25. Myrzakhmetov, B., Kozhirbayev, Z.: Extended language modeling experiments for Kazakh. In: Proceedings of 2018 International Workshop on Computational Models in Language and Speech, p. 42 (2018)
26. Heafield, K.: KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197 (2011)