

# Fine-tuning OpenAI's Whisper model for Kazakh speech recognition

Bakdaulet Kynabay<sup>1</sup>, Aimoldir Aldabergen<sup>2</sup>, Shirali Kadyrov<sup>3</sup>, Ardak Shalkarbay-uly<sup>4</sup>

<sup>1</sup>School of Information Technology and Applied Mathematics, SDU University, Almaty, Kazakhstan

<sup>2</sup>School of Artificial Intelligence, Qazaq AI Research University, Astana, Kazakhstan

<sup>3</sup>College of Humanities and Sciences, Ajman University, Ajman, United Arab Emirates

<sup>4</sup>School of Digital Technologies, Narxoz University, Almaty, Kazakhstan

## Article Info

### Article history:

Received Nov 25, 2025

Revised Jul 30, 2026

Accepted Aug 12, 2026

### Keywords:

Fine-tuning

Kazakh

Kazakh speech corpus

Low-resource languages

Speech recognition

Transfer learning

Whisper model

## ABSTRACT

Speech recognition remains a challenging task, especially for low-resource languages such as Kazakh. This study explores effectiveness of fine-tuning OpenAI's Whisper model on Kazakh speech data for improving speech recognition accuracy. The Kazakh speech corpus (KSC) is used as a benchmark dataset and to conduct experiments to evaluate how the adapted Whisper framework measures up against current leading speech recognition models, including deep neural network (DNN)-hidden Markov model (HMM), end-to-end long short-term memory (E2E-LSTM), and end-to-end transformer (E2E-transformer). Research outcomes indicate that the Whisper model outperforms all baselines after fine-tuning, showing a result with a word error rate (WER) of 12.5%, a character error rate (CER) of 3.15%, a character-level F1-score (chrF) of 93.72, and a bilingual evaluation understudy (BLEU) score of 76.05. These findings suggest that the Whisper model that is fine-tuned on Kazakh audio data can be a promising approach for improving Kazakh speech recognition accuracy. This work advances the development of speech recognition for the Kazakh language by demonstrating the effectiveness of transfer learning in low-resource languages. Subsequent research must prioritize collecting larger and more diverse speech data for Kazakh, including noisy and accented speech, to further improve speech recognition accuracy.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Aimoldir Aldabergen

School of Artificial Intelligence, Qazaq AI Research University

Astana, Kazakhstan

Email: aimoldir.alldabergen@gmail.com

## 1. INTRODUCTION

Speech recognition plays a key role in modern communication, allowing people to interact with devices using natural speech. Machines use acoustic processing and modeling algorithms to recognize human speech, creating a link between people and devices [1]. Despite the fact that there is much research done and many automatic speech recognition (ASR) systems are well developed for major languages, there is a lack of study of low-resource languages such as Kazakh, spoken by over 18 million people. The reasons for this situation are the lack of annotated datasets and the agglutinative nature of the Kazakh language [2]. Kazakh ASR improvement plays a big role in preserving the language, increasing its accessibility, and supporting cultural exchange.

During the past decades, the science of speech recognition has advanced remarkably. Initial ASR systems were able to recognize phonemes and acoustic patterns, but could process only small vocabularies.

One of the examples of an early system was a system for recognizing spoken digits on the telephone line [3]. In time, new approaches that relied on probability models and large corpora were introduced and improved the performance of the systems [4]–[6]. For instance, an experiment that compared pitch-shifting, time-stretching and volume control augmentations revealed a marked reduction in word error rate (WER) for the convolutional neural network (CNN) recognizer when used in conjunction with hidden Markov model (HMM), support vector machine (SVM), and CNN models [7].

The development of deep neural networks (DNNs) has enormously accelerated the research in ASR. DNNs are much better than older generations of Gaussian mixture model (GMM)-HMM models due to their ability to extract highly abstract and useful features from speech data [8]–[10]. Neural network models that are noise-resistant include CNNs [11]–[13] and long short-term memory (LSTM) networks [14], [15]. ASR systems have become increasingly simpler and achieved unprecedented levels of performance through end-to-end encoder-decoder models with attention [16]. Other unsupervised approaches such as generative adversarial network (GAN)-based learning and contrastive predictive coding (CPC) have further enhanced efficiency and noise tolerance [17].

Nevertheless, low-resource languages like Kazakh face numerous challenges due to a lack of data, complex word structures, different accents, and very few language resources [18]. Recently, there was an increase in the number of papers dedicated to Kazakh ASR. Researchers explored different approaches, including HMM-based approaches, hybrid models, as well as deep learning models such as transformers and CTC-based approaches [19]. WERs for continuous speech recognition systems are 30% [20], while character error rate (CER) for transformer models were only 3.7% on clean data [2]. Crowdsourced datasets have increased training resources [21], and the Kazakh text-to-speech (TTS) corpus now supports both ASR and speech synthesis tasks [22].

The fine-tuning of the OpenAI Whisper model [23] has recently emerged as a promising direction for low-resource domain-specific ASR. Studies have shown improved performance when Whisper is fine-tuned in targeted lexical domains [24], Korean speech [25], Javanese [26], and atypical Italian speech [27]. These results show strong generalization and adaptability in languages and speaking conditions. This research is motivated by the need for accurate speech recognition technology for the Kazakh language, which has demonstrated a positive impact on language preservation and revitalization. Accurate speech recognition can help to create voice-controlled devices and software in Kazakh, which can be implemented in education, government, and business. Based on a review of the literature, this study is the first to apply the Whisper model to Kazakh speech recognition. The introduced approach uses recent progress in DNN-based speech recognition and adapts it for the Kazakh language. These results are expected to improve Kazakh ASR and stimulate further research in this area.

The article proceeds as follows. Section 2 outlines the Kazakh speech corpus (KSC), data pre-processing procedures, the Whisper model, and fine-tuning procedure. Section 3 gives experimental results. Section 4 concludes this study.

## 2. METHOD

The details of training a speech recognition model using the Whisper architecture and fine-tuning it using a KSC are described in this section. In particular, this section introduces a Kazakh language large-scale speech corpus and its possible applications. Then, the training approach via web-scale supervised pre-training for the Whisper model is discussed, including the data preprocessing steps, automated filtering methods for improving the transcript quality, and the model architecture. Finally, this study describes the fine-tuning of the Whisper model using the KSC and performance evaluation metrics for evaluating the accuracy of the model. Namely, the WER and CER.

### 2.1. Kazakh speech corpus

The open-source KSC is used as a comprehensive, publicly accessible large-scale dataset in Kazakh [21]. It includes roughly 332 hours of verified audio, which are split into 153,000 distinct phrases. To guarantee high quality, the corpus was compiled using crowdsourcing techniques and carefully checked by native Kazakh speakers. To ensure data integrity, the dataset was built through crowdsourcing and underwent rigorous validation by native Kazakh speakers. This resource was specifically developed to advance automated speech and language technologies for Kazakh, which is a Turkic language characterized by agglutination and vowel harmony. The foundational texts for the audio prompts were aggregated from a diverse array of materials, ranging from legislation and e-books to digital platforms like blogs, news outlets, and Wikipedia. A special text extraction tool was created for each website to extract text data from it. The text data collected was used to generate speech prompts, which were recorded by participants using a web interface.

## 2.2. Whisper model and fine-tuning

Developed by OpenAI, Whisper is a robust ASR model trained on a massive 680,000-hour collection of annotated audios. Within this dataset, 117,000 hours are dedicated to multilingual recordings over 96 languages, including Kazakh, with some annotations translated into English. Whisper model, after pre-training, shows strong performance in many languages, achieving a WER of 3% on the test-clean subset of the LibriSpeech ASR benchmark. Structurally, Whisper relies on a transformer-based encoder-decoder framework. The encoder translates log-mel spectrograms into vector sequences through several transformer layers. After that, the decoder functions as a language model, predicting subsequent tokens from these vectors to generate the final transcription (as depicted in Figure 1). Whisper model was fine-tuned and pre-trained on a specific Kazakh dataset in order to maximize transcription accuracy for the Kazakh language. Fine-tuning allowed us to adapt the model to handle the linguistic nuances of Kazakh better, further improving its transcription quality.

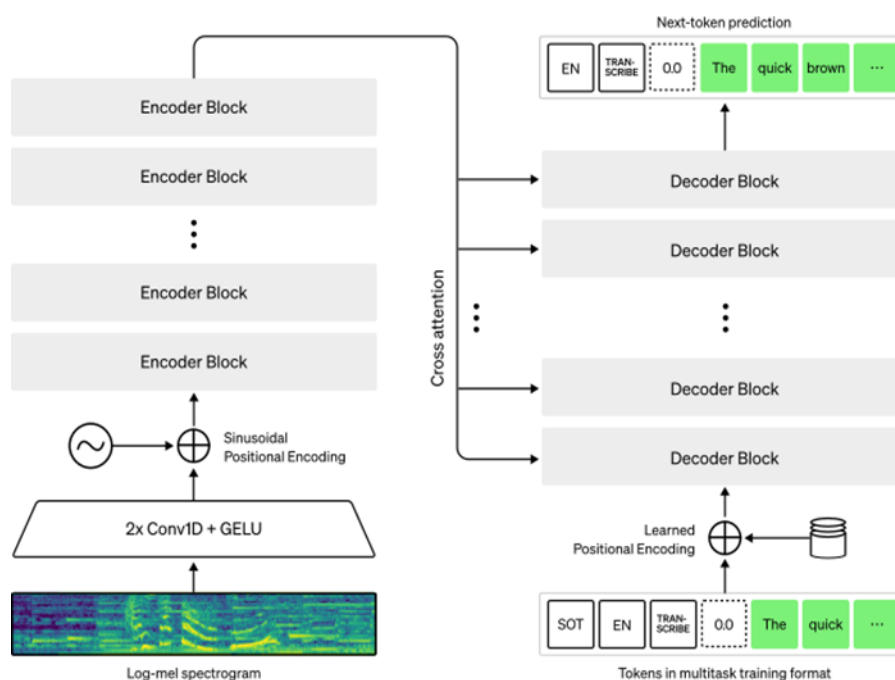


Figure 1. Whisper architecture

## 2.3. Dataset and data splitting

The KSC introduced by Khassanov *et al.* [21] was used for both model fine-tuning and evaluation. To maintain experimental integrity and allow direct baseline comparisons, this study strictly adhered to the corpus's original data separation strategy. The dataset was split with absolute speaker exclusivity for train, validation, and test sets, with no individual participant's voice overlapped across splits. The training set includes around 318 hours of audio sourced from anonymous individuals. Validation and test sets contain approximately 7.1 hours of audio recorded from identifiable participants balanced by gender, age group, and region. The test set was used exclusively for final evaluation, consistent with the experimental protocol reported in the KSC baseline study.

## 2.4. Preprocessing pipeline

Fine-tuning was carried out using a preprocessing pipeline consistent with the Whisper architecture and closely following the attached fine-tuning code template. Initially, all voice recordings were loaded and resampled to 16 kHz, which is the sampling rate expected by Whisper. Audio signals were converted to single-channel (mono) format where necessary. No explicit silence trimming or manual segmentation was applied beyond what is implicitly handled during corpus preparation, as KSC utterances are already pre-segmented and quality-checked.

In the next step, acoustic features were extracted using the Whisper feature extractor. Original audio signals were transformed into log-mel spectrograms by applying the exact configuration utilized during the model's initial pretraining. These log-mels form the encoder input and ensure consistency of architecture during pretraining and fine-tuning.

Subsequently, transcription texts were tokenized using the Whisper tokenizer. Ground truth transcriptions were tokenized into subword token ID sequences according to the multilingual vocabulary used by Whisper and supporting native Kazakh characters. No other text normalization, transliteration, or language-specific preprocessing was done besides that included in the Whisper tokenizer's logic. Mini-batch training was made possible by using a data collator to dynamically pad both the input spectrograms and output tokens in each mini-batch. The padding was only done at the mini-batch level, with attention masks appropriately used to ignore any contribution of the padding elements to the loss function. Table 1 provides the basic parameters for fine-tuning the Whisper used in the current study: model variant, dataset, speaker-disjoint data split and key preprocessing and evaluation settings. Table 1 gives an overview of the parameters described more thoroughly in subsections 2.2 to 2.5 and highlights the aspects most relevant for reproducibility and fair comparison with the KSC baseline results.

Table 1. Essential Whisper fine-tuning setup on KSC

| Aspect                                  | Specification                               |
|-----------------------------------------|---------------------------------------------|
| Model                                   | Whisper-small (pre-trained)                 |
| Dataset                                 | KSC                                         |
| Data split                              | Speaker-disjoint (train/validation/test)    |
| Training data/validation data/test data | 318 h/7.1 h/7.1 h                           |
| Sampling rate                           | 16 kHz                                      |
| Audio format                            | Mono                                        |
| Padding strategy                        | Dynamic padding via data collator           |
| Maximum audio length                    | Whisper default (long utterances truncated) |
| Loss function                           | Cross-entropy loss                          |
| Optimization                            | Supervised fine-tuning                      |
| Input features                          | Log-mel spectrograms                        |
| Tokenizer                               | Whisper multilingual tokenizer              |

## 2.5. Fine-tuning procedure

Supervised learning of the pre-trained Whisper-small model was carried out on the KSC training dataset. The model parameters were updated through minimizing the cross-entropy loss between the ground truth transcription and the predicted token sequences. The hyperparameters such as learning rate, batch size, and number of epochs were chosen based on the configuration specified in the fine-tuning script and tuned with respect to performance on the KSC validation dataset. The fine-tuned model was evaluated only on the KSC test set after training convergence, following the speaker-independent evaluation protocol as in the original KSC baseline experiments. This followed the speaker-independent testing protocol used for the baseline experiments with the KSC system. Thus, performance comparison between Whisper and ASR models previously trained on the KSC dataset is valid from a methodological viewpoint. Overall, this approach to preprocessing and fine-tuning allowed Whisper to adjust its acoustic and linguistic representation to the phonetic structure, morphology, and orthographic peculiarities of the Kazakh language.

## 2.6. Performance evaluation metrics: character error rate and word error rate

In this study, four evaluation metrics are used in order to compare different speech recognition models' performance: CER, WER, character-level F1-score (chrF), and bilingual evaluation understudy (BLEU). CER shows the percentage of incorrectly recognized transcript characters. It is calculated using (1).

$$CER = \frac{S+D+I}{N} \quad (1)$$

Where  $S$  stands for the number of characters that are recognized incorrectly (substitutions),  $D$  stands for number of missing characters (deletions),  $I$  stands for the number of extra characters (insertions), and  $N$  is total number of characters in the text. WER is the percentage of incorrectly recognized words in the context. It is calculated using (2).

$$WER = \frac{S+D+I}{N} \quad (2)$$

Where  $S$  represents substituted or misrecognized words,  $D$  denotes omitted words,  $I$  stand for falsely added words, and  $N$  indicates the overall word count of the ground-truth text. To evaluate the alignment of character-level n-grams between the generated output and the true transcript, we utilize the chrF. This metric calculates the harmonic mean of recall ( $R$ ) and precision ( $P$ ) at the character level defined by (3).

$$chrF = \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \quad (3)$$

Here,  $P$  denotes the fraction of generated character  $n$ -grams matching the ground truth, while  $R$  specifies the fraction of reference  $n$ -grams successfully captured by the output. The parameter  $\beta$  acts as a balancing weight assigned a value of 2 for these experiments to place greater importance on recall. The BLEU score itself is computed by combining a brevity penalty (BP) with the weighted geometric mean of  $n$ -gram precisions, as in (4).

$$BLEU = BP \cdot \exp(\sum_{n=1}^N \omega_n \log p_n), \quad (4)$$

Where  $BP$  is a BP applied when the hypothesis is shorter than the reference,  $N$  is the maximum  $n$ -gram order,  $\omega_n$  is the weight of each  $n$ -gram and  $p_n$  is the precision of  $n$ -grams.

Both CER and WER provide a quantitative measure of recognition errors. Meanwhile, chrF and BLEU capture overlap-based similarity between predicted and reference transcripts. Lower CER and WER values indicate higher transcription accuracy, whereas higher chrF and BLEU scores indicate better correspondence with the reference.

### 3. RESULTS AND DISCUSSION

Table 2 demonstrates the comparative analysis of the performance of different models on the KSC dataset designed for speech recognition. The considered models are the traditional DNN-HMM model, two end-to-end models (E2E-LSTM and E2E-transformer), the original Whisper-small model, and the proposed fine-tuned Whisper-small model. The metrics are CER, WER, chrF, and BLEU. It is evident from the obtained results that the Whisper-small model remarkably outperforms all other models by metrics of evaluation: CER of 3.15, a WER of 12.5, a chrF of 93.72, and a BLEU score of 76.05. This demonstrates that the fine-tuned model is substantially more accurate in recognizing Kazakh speech compared to both traditional and end-to-end models.

Table 2. Comparison of model performance (%)

| Reference | Model                      | CER   | WER    | chrF  | BLEU  |
|-----------|----------------------------|-------|--------|-------|-------|
| [23]      | DNN-HMM                    | 4.6   | 13.7   | -     | -     |
| [23]      | E2E-LSTM                   | 8.7   | 28.8   | -     | -     |
| [23]      | E2E-transformer            | 4.9   | 18.8   | -     | -     |
| [25]      | Original whisper small     | 73.34 | 118.57 | 29.36 | 1.30  |
| This work | Fine-tuned whisper (small) | 3.15  | 12.5   | 93.72 | 76.05 |

The DNN-HMM model showed the results with 4.6 of CER and 13.7 of WER, while the E2E-LSTM model was shown to have the CER of 8.7 and WER of 28.8. In contrast, the original Whisper-small model performed poorly on this dataset, with a CER of 73.34, a WER of 118.57, a chrF of 29.36, and a BLEU score of 1.30. These results indicate that, although end-to-end models are promising, they still require further adaptation to accurately recognize Kazakh speech, whereas the Whisper model shows state-of-the-art performance after being fine-tuned. To further understand where Whisper adapts during fine-tuning, this study investigated the parameter changes in the encoder layers. For each layer  $j$ , the relative L2 change was computed as (5).

$$\Delta_j = \frac{\|\theta_j^{(ft)} - \theta_j^{(0)}\|_2}{\|\theta_j^{(0)}\|_2} \quad (5)$$

Where  $\theta_j^{(0)}$  and  $\theta_j^{(ft)}$  denote the pre-trained and fine-tuned parameters of layer  $j$ , respectively.

Figure 2 shows the resulting layer-wise adaptation map. The most substantial parameter updates occur in mid-level encoder layers (layers 3–7,  $\Delta_j = 0.034 - 0.036$ ), while early and late layers exhibit smaller changes (e.g., layer 0: 0.0255; layer 11: 0.0166). This pattern suggests that fine-tuning primarily refines intermediate acoustic–phonetic representations, while preserving low-level spectral processing and higher-level linguistic abstraction learned during large-scale multilingual pre-training.

The results of this research have proven that fine-tuning OpenAI’s Whisper-small model on the KSC improves the accuracy of Kazakh speech recognition remarkably, by achieving state-of-the-art performance across CER, WER, chrF, and BLEU. These results have shown the perspectives of applying this

model for practical purposes within voice-operated systems in different fields, namely, education, government, and business spheres. Moreover, the layer-wise adaptation analysis suggests that fine-tuning primarily refines intermediate acoustic–phonetic representations while preserving lower-level spectral and higher-level linguistic knowledge acquired during large-scale multilingual pre-training. Future research should investigate the deployment of the fine-tuned Whisper model in real-world scenarios and explore strategies to further enhance the performance of speech recognition systems for low-resource languages such as Kazakh.

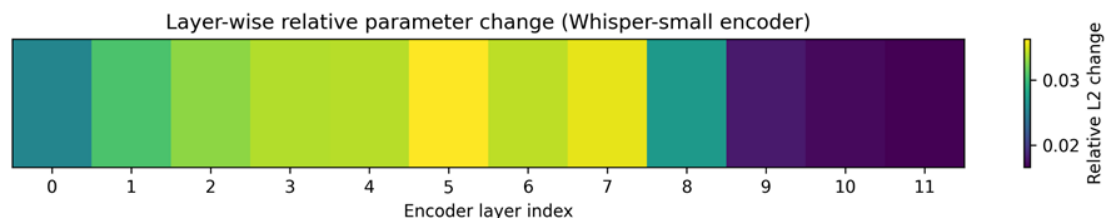


Figure 2. Layer-wise relative L2 parameter change after fine-tuning, with mid-level layers showing the strongest adaptation

#### 4. CONCLUSION

The current study has demonstrated the possibility of fine-tuning the Whisper model developed by OpenAI for Kazakh speech recognition and creating a new state-of-the-art benchmark for the language. Tested against the baseline models (DNN-HMM, E2E-LSTM, and E2E-transformer) based on the KSC, the Whisper model showed the best results with CER of 3.15%, WER of 12.5%, chrF of 93.72, and BLEU score of 76.05 after being fine-tuned. The obtained results have proved the efficiency of transfer learning as an approach that allows improving the performance of ASR systems for low-resource languages such as Kazakh. Using the transfer learning approach to adapt a pre-trained model is an effective way to improve speech recognition in low-resource languages like Kazakh. This study has also provided a benchmark for future research in the field of Kazakh ASR. Future research should be aimed at collecting a larger and more varied corpus, including noisy, spontaneous, and accented speech, and using metrics like chrF and BLEU to evaluate the models.

#### FUNDING INFORMATION

The Ministry of Science and Higher Education of the Republic of Kazakhstan provided this research with financial support under project grant AP23487777.

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author       | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|----------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Bakdaulet Kynabay    | ✓ | ✓ | ✓  | ✓  | ✓  | ✓ |   | ✓ | ✓ | ✓ |    | ✓  | ✓ |    |
| Aimoldir Aldabergen  | ✓ | ✓ | ✓  |    | ✓  | ✓ |   |   | ✓ | ✓ | ✓  |    |   |    |
| Shirali Kadyrov      |   | ✓ |    | ✓  | ✓  | ✓ | ✓ |   |   | ✓ |    | ✓  | ✓ | ✓  |
| Ardak Shalkarbay-uly |   | ✓ | ✓  | ✓  |    |   | ✓ |   | ✓ |   |    |    |   | ✓  |

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**dit

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

#### CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

This section not applicable, as this study did not involve human participants or personally identifiable information.

## ETHICAL APPROVAL

This section not applicable, as this study did not involve human or animal subjects and used only non-sensitive computational or public data.

## DATA AVAILABILITY

The data used for the findings of this study are available from a third party. The Kazakh Speech Corpus 2 developed by Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, is openly available at <https://doi.org/10.48342/m90y-aj02> and can be downloaded from the ISSAI website at <https://issai.nu.edu.kz/kz-speech-corpus/>.

## REFERENCES

- [1] H. Ibrahim and A. Varol, "A study on automatic speech recognition systems," in *8th International Symposium on Digital Forensics and Security, ISDFS 2020*, 2020, pp. 1–6, doi: 10.1109/ISDFS49300.2020.9116286.
- [2] M. Orken, O. Dina, A. Keylan, T. Tolganay, and O. Mohamed, "A study of transformer-based end-to-end speech recognition system for Kazakh language," *Scientific Reports*, vol. 12, no. 1, May 2022, doi: 10.1038/s41598-022-12260-y.
- [3] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952, doi: 10.1121/1.1906946.
- [4] C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994, doi: 10.1109/89.279278.
- [5] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007, doi: 10.1561/20000000004.
- [6] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 69–88, 2002, doi: 10.1006/csla.2001.0184.
- [7] J. Galić and Đ. Grozdić, "Exploring the impact of data augmentation techniques on automatic speech recognition system development: a comparative study," *Advances in Electrical and Computer Engineering*, vol. 23, no. 3, pp. 3–12, 2023, doi: 10.4316/AECE.2023.03001.
- [8] D. Fohr, O. Mella, and I. Illina, "New paradigm in speech recognition: deep neural networks," in *IEEE International Conference on Information Systems and Economic Intelligence*, Apr. 2017.
- [9] T. Grósz, "Training methods for deep neural network-based acoustic models in speech recognition," *Ph.D. thesis*, Department of Informatics, University of Szeged, Szeged, Hungary, 2018.
- [10] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks – studies on speech recognition tasks," in *International Conference on Learning Representations (ICLR 2013)*, 2013.
- [11] O. A.-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Annual Conference of the International Speech Communication Association, Interspeech*, 2013, pp. 3366–3370, doi: 10.21437/interspeech.2013-744.
- [12] O. A.-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014, doi: 10.1109/TASLP.2014.2339736.
- [13] T. N. Sainath *et al.*, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015, doi: 10.1016/j.neunet.2014.08.005.
- [14] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.
- [15] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Annual Conference of the International Speech Communication Association, Interspeech*, 2014, pp. 338–342, doi: 10.21437/interspeech.2014-80.
- [16] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.
- [17] A. Baevski and M. Auli, "Effectiveness of self-supervised pre-training for speech recognition," 2020, *arXiv:1911.03912*.
- [18] A. Mehri, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, 2023, doi: 10.1016/j.inffus.2023.101869.
- [19] A. Karibayeva, V. Karyukin, B. Abduali, and D. Amirova, "Speech recognition and synthesis models and platforms for the Kazakh language," *Information*, vol. 16, no. 10, 2025, doi: 10.3390/info16100879.
- [20] O. Mamyrbayev *et al.*, "Continuous speech recognition of Kazakh language," *ITM Web of Conferences*, vol. 24, 2019, doi: 10.1051/itmconf/20192401012.
- [21] Y. Khassanov, S. Mussakhoyayeva, A. Mirzakhmetov, A. Adiyev, M. Nurpeiissov, and H. A. Varol, "A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline," in *16th Conference of the European Chapter of the Association for Computational Linguistics*, 2020, pp. 697–706, doi: 10.18653/v1/2021.eacl-main.58.
- [22] S. Mussakhoyayeva, Y. Khassanov, and H. A. Varol, "KazakhTTS2: extending the open-source Kazakh TTS corpus with more data, speakers, and topics," in *2022 Language Resources and Evaluation Conference, LREC 2022*, 2022, pp. 5404–5411, doi: 10.63317/2592g9pbifyft.
- [23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, *arXiv:2212.04356*.

- [24] J. Chang and H. Nam, "Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: a case study on Whisper model and KsponSpeech dataset," *Phonetics and Speech Sciences*, vol. 15, no. 3, pp. 83–88, 2023, doi: 10.13064/ksss.2023.15.3.083.
- [25] C. Oh, C. Kim, and K. Park, "Building robust Korean speech recognition model by fine-tuning large pretrained model," *Phonetics and Speech Sciences*, vol. 15, no. 3, pp. 75–82, 2023, doi: 10.13064/ksss.2023.15.3.075.
- [26] R. S. A. Pratama and A. Amrullah, "Analysis of Whisper automatic speech recognition performance on low resource language," *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 1, pp. 1–8, 2024, doi: 10.33480/pilar.v20i1.4633.
- [27] D. Mulfari and M. Villari, "A voice user interface on the edge for people with speech impairments," *Electronics*, vol. 13, no. 7, 2024, doi: 10.3390/electronics13071389.

## BIOGRAPHIES OF AUTHORS



**Bakdaulet Kynabay**    received his bachelor's degree in Computer Science from SDU University, Kazakhstan, in 2018, and completed his master's degree in Computer Science at the same university in 2020. His research interests include deep learning, machine learning, natural language processing, and inclusive technologies, with a particular focus on speech recognition and language modeling for low-resource languages. In this work, he led the design and development of models for fine-tuning AI-based speech recognition systems. He contributed to model training and comparative evaluation against baseline systems. He also played a key role in conceptualizing the study, supervising the research activities, and managing project administration. He can be contacted at email: bakdaulet.kynabay@gmail.com.



**Aimoldir Aldabergen**    received her bachelor's degree in Computer Science from SDU University, Kazakhstan, in 2018. She received her master's degree in Computer Science at the same institution in 2020. Currently, she is pursuing her Ph.D. in Computer Science at SDU University and works as a senior lecturer at Qazaq AI Research University, Astana, Kazakhstan. Her research area includes artificial intelligence, speech and language technologies, and algorithm design to support inclusive education. In the present study, she contributed to the development, fine-tuning and evaluation of an AI-based speech recognition model for the Kazakh language. Her work focuses on applying transfer learning using OpenAI's Whisper model to enhance recognition accuracy in low-resource language settings. She can be contacted at email: aimoldir.alldabergen@qairu.edu.kz.



**Shirali Kadyrov**    earned his B.Sc. from Bogazici University in Turkey (2004) before completing both his M.Sc. (2009) and Ph.D. (2010) at The Ohio State University, United States. Currently serving as an associate professor at College of Humanities and Sciences, Ajman University. He has held a number of international academic positions. These include working as an EPSRC postdoctoral researcher at the University of Bristol, a research assistant at ETH Zurich, a full-time faculty member at Nazarbayev University, a full professor at SDU University, and the dean of Faculty at Oxus University. His primary research focuses on ergodic theory, dynamical systems, data science and number theory. For this specific project, he directed the experimental design and analytical methods, applying his background in mathematical modeling to guarantee the scientific validity and rigor of the speech recognition evaluation. He can be contacted at email: s.kadyrov@ajman.ac.ae.



**Ardak Shalkarbay-uly**    holds both a master's (2010) and a bachelor's (2004) degree in Computer Science from SDU University in Kazakhstan. He currently works as a senior lecturer at Narxoz University, where his primary focus is on natural language processing (NLP) and artificial intelligence transformation. Within the broader scope of machine learning, his specific research interests involve deep neural networks and automated grammar error correction. For this publication, he leveraged his algorithmic background to manage the design, implementation, and fine-tuning of the ASR models. Additionally, he was responsible for successfully securing the grant funding that supported this research. He can be contacted at email: ardak.shalkar@gmail.com.