ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№2
2026

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

2 (358)

APRIL – JUNE 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).
Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.
Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*
Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*

Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**

Тематическая направленность: *информационно-коммуникационные технологии.*

В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*

Периодичность: *4 раза в год.*

http://www.physico-mathematical.kz/index.php/en/

CONTENTS

COMPUTER SCIENCE

МАЗМҰНЫ

КОМПЬЮТЕРЛІК ҒЫЛЫМДАР

СОДЕРЖАНИЕ

КОМПЬЮТЕРНЫЕ НАУКИ

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 2.
Namber 358 (2026). 317–334

https://doi.org/10.32014/2026.2518-1726.441

IRSTI 20.53.15
UDC 004.934

© Rakhimova D.[1], Duisenbekkyzy Zh.[1*], Karibayeva A. [1], Eşref A.[2], Ilessova B.[1], 2026.
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2] Istanbul Technical University, Istanbul, Turkey.
E-mail: duisenbekkyzy.zh@kaznu.kz

IMPROVING THE VOICE RECOGNITION SYSTEM FOR CHILDREN IN KAZAKH THROUGH ADDITIONAL TRAINING (FINE-TUNING)

Rakhimova Diana — PhD, Associate Professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Zhansaya Duisenbekkyzy — Senior Lecturer, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: 1695507zh@gmail.com, https://orcid.org/0000-0001-7743-0795;
Eşref Adali — Professor, Istanbul Technical University, Istanbul, Turkey,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255;
Karibayeva Aidana — PhD, Acting Associate Professor; Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: a.s.karibayeva@gmail.com, https://orcid.org/0000-0002-2023-1573;
Ilessova Bakytgul — Senior Lecturer, Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: b.ilessova@gmail.com, https://orcid.org/0000-0001-8357-519X.

Abstract. This article presents a comprehensive method for automatic speech recognition (ASR) of Kazakh children's speech, with a particular focus on children with speech disorders. The study introduces the first annotated acoustic dataset for Kazakh children's speech, comprising 8,594 audio recordings from children aged 2–11 years, collected via a Telegram bot, controlled dictaphone sessions, and naturalistic recordings. A three-stage experimental framework was employed: (1) zero-shot evaluation of six ASR models (Whisper large-v2, MMS, TurkicASR, ElevenLabs, Vosk, GenSec); (2) fine-tuning of Whisper large-v2, MMS, and TurkicASR on the children's corpus; and (3) application of a novel LLM-based post-editing pipeline (PhonNorm → MorphNorm → LLM). In the zero-shot regime, all models achieved WER of 0.70–0.90, confirming that adult-trained models are inadequate for Kazakh children's speech. After fine-tuning, Whisper large-v2 achieved WER = 0.30 and Accuracy = 58%. The proposed post-editing pipeline further reduced WER to 0.18 and CER to 0.065, with Accuracy reaching 70% — a 74.3% overall reduction from the baseline WER of 0.70. Practical validation at Kindergarten No. 154 in Almaty (48 children, aged 3–8) confirmed the system's applicability for speech disorder screening, with results showing high agreement with expert speech-language pathologist assessments. The proposed method is applicable in educational platforms, logopedic diagnostics, and mobile speech-assistance systems for low-resource languages.

Keywords: Whisper large-v2, fine-tuning, Kazakh language, automatic speech recognition (ASR), children's speech, speech disorders, LLM-based post-editing

***Financing**. This study was funded by the grant project «Study of automatic generation of parallel speech corpora of Turkic languages and their use for neural models» (grant number IRN AP23488624) of the Ministry of Science and Higher Education of the Republic of Kazakhstan.*

For citations: Rakhimova D., Duisenbekkyzy Zh., Karibayeva A., Eşref A., Ilessova B. Improving the voice recognition system for children in Kazakh through additional training (fine-tuning). Academic Scientific Journal of Computer Science, 2026. — No.2. – P. 317–334. DOI https://doi.org/10.32014/2026.2518-1726.441

© Рахимова Д.[1], Дүйсенбекқызы Ж.[1*], Кәрібаева А.[1], Eşref A.[2], Ілесова Б.[1], 2026.

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Стамбул техникалық университеті, Стамбул, Түркия.
E-mail: duisenbekkyzy.zh@kaznu.kz

ҚАЗАҚ ТІЛІНДЕГІ БАЛАЛАР ДАУЫСЫН ТАНУ ЖҮЙЕСІН ҚОСЫМША ОҚЫТУ (FINE-TUNING) АРҚЫЛЫ ЖЕТІЛДІРУ

Рахимова Диана — PhD, қауымдастырылған профессор, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: drakhimova060@gmail.com, https://orcid.org/0000-0003-1427-198X;
Дүйсенбеккызы Жансая — аға оқытушы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: 1695507zh@gmail.com, https://orcid.org/0000-0001-7743-0795;
Eşref Adali — PhD, профессор, Стамбул техникалық университет. Стамбул, Түркия,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255;
Кәрібаева Айдана — PhD, доцент м.а., әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: a.s.karibayeva@gmail.com, https://orcid.org/0000-0002-2023-1573;
Ілесова Бақытгүл — аға оқытушы, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
E-mail: b.ilessova@gmail.com, https://orcid.org/0000-0001-8357-519X.

Аннотация. Бұл мақалада қазақ тіліндегі сөйлеу бұзылысы бар балалардың сөйлеуін автоматты тануға арналған кешенді үш кезеңді әдіс ұсынылады. Зерттеу барысында 2–11 жас аралығындағы балалардан жиналған 8 594 аудиожазбадан тұратын қазақ тіліндегі балалар сөйлеуінің тұңғыш аннотацияланған акустикалық корпусы жасалды. Үш кезеңді схема қолданылды: (1) алты ASR моделін zero-shot режимінде бағалау; (2) таңдалған модельдерді балалар корпусында fine-tuning арқылы бейімдеу; (3) LLM негізіндегі постөңдеу тізбегін (PhonNorm → MorphNorm → LLM) қолдану. Zero-shot режимінде барлық модельдер WER = 0.70–0.90 нәтижесін берді, бұл ересектер деректерінде оқытылған модельдердің қазақ балалар сөйлеуіне жарамсыздығын дәлелдеді. Fine-tuning кейін Whisper large-v2 WER = 0.30, Accuracy = 58% нәтижесіне жетті. Ұсынылған постөңдеу тізбегі WER-ді 0.18-ге, CER-ді 0.065-ке дейін төмендетті, Accuracy 70%-ға жетті — базалық WER 0.70-тен жалпы 74.3% азайды. Алматы қаласындағы №154 балабақшада (48 бала, 3–8 жас) жүргізілген практикалық апробация жүйенің сөйлеу бұзылыстарын скринингтеуге жарамдылығын растады; нәтижелер логопед мамандарының бағалауымен жоғары сәйкестік көрсетті. Ұсынылған постөңдеу тізбегі WER-ді 0.18-ге, CER-ді 0.065-ке дейін төмендетті, Accuracy 70%-ға жетті — базалық WER 0.70-тен жалпы 74.3% азайды. Алматы қаласындағы №154 балабақшада (48 бала, 3–8 жас) жүргізілген практикалық апробация жүйенің сөйлеу бұзылыстарын скринингтеуге жарамдылығын растады; нәтижелер логопед мамандарының бағалауымен жоғары сәйкестік көрсетті. Ұсынылған әдіс білім беру платформаларында, логопедиялық диагностикада және аз ресурсты тілдерге арналған мобильді сөйлеу жүйелерінде қолданылуы мүмкін.

Түйін сөздер: Whisper large-v2, fine-tuning, қазақ тілі, автоматты сөйлеуді тану (ASR), балалар сөйлеуі, сөйлеу бұзылыстары, , LLM негізіндегі постөңдеу

© Рахимова Д.[1], Дуйсенбеккызы Ж.[1*], Карибаева А.[1], Eşref A.[2], Илесова Б.[1], 2026.

[1]Казахский Национальный университет имени аль-Фараби, Алматы, Казахстан;

[2]Стамбульский Технический Университет, Стамбул, Турция.

E-mail: duisenbekkyzy.zh@kaznu.kz

СОВЕРШЕНСТВОВАНИЕ СИСТЕМЫ РАСПОЗНАВАНИЯ ГОЛОСА ДЕТЕЙ НА КАЗАХСКОМ ЯЗЫКЕ ПУТЕМ ДОПОЛНИТЕЛЬНОГО ОБУЧЕНИЯ (FINE-TUNING)

Рахимова Диана — PhD, ассоциированный профессор, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: drakhimova060@gmail.com, ORCID: https://orcid.org/0000-0003-1427-198X;
Дуйсенбеккызы Жансая — старший преподаватель, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

E-mail: 1695507zh@gmail.com, ORCID: https://orcid.org/0000-0001-7743-0795;
Eşref Adali — PhD, профессор, Стамбульский технический университет, Стамбул, Турция,
E-mail: adali@itu.edu.tr, https://orcid.org/0000-0002-1561-8255;
Карибаева Айдана — PhD, и.о. доцента, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: a.s.karibayeva@gmail.com, https://orcid.org/0000-0002-2023-1573;
Илесова Бақытгүл — старший преподаватель, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: b.ilessova@gmail.com, https://orcid.org/0000-0001-8357-519X.

Аннотация: *Актуальность.* Автоматическое распознавание речи детей на казахском языке является актуальной задачей для образовательных технологий, логопедической диагностики и мобильных речевых систем. Особую сложность представляет распознавание речи детей раннего возраста и детей с речевыми нарушениями, поскольку такие данные характеризуются высокой вариативностью произношения, фонетическими и морфологическими отклонениями, а также ограниченностью доступных размеченных корпусов для казахского языка. *Цель.* Разработать и апробировать комплексный трехэтапный метод автоматического распознавания речи казахских детей, включая детей с речевыми нарушениями, на основе дополнительного обучения ASR-моделей и постредактирования с использованием больших языковых моделей. *Методы.* В ходе исследования создан первый аннотированный акустический корпус казахской детской речи, включающий 8 594 аудиозаписи детей 2–11 лет, собранные через Telegram-бот, диктофонные сессии и натуралистические записи. Методика исследования включает три этапа: zero-shot оценку шести ASR-моделей, включая Whisper large-v2, MMS, TurkicASR, ElevenLabs, Vosk и GenSec; дообучение моделей Whisper large-v2, MMS и TurkicASR на детском корпусе; постредактирование результатов распознавания на основе LLM-конвейера PhonNorm → MorphNorm → LLM. Обучение проводилось в среде Google Colab с использованием GPU NVIDIA T4, оптимизатора AdamW и функции потерь cross-entropy. Эффективность моделей оценивалась по метрикам Word Error Rate (WER), Character Error Rate (CER) и Accuracy. *Результаты и выводы.* В режиме zero-shot все модели показали высокий уровень ошибок: WER находился в диапазоне 0,70–0,90. После fine-tuning модель Whisper large-v2 достигла WER = 0,30 и Accuracy = 58%. После применения LLM-постредактирования показатели улучшились до WER = 0,18, CER = 0,065 и Accuracy = 70%, что соответствует снижению WER на 74,3% относительно базового уровня. Практическая апробация системы в детском саду №154 города Алматы с участием 48 детей в возрасте 3–8 лет подтвердила применимость предложенного подхода для скрининга речевых нарушений и показала высокое соответствие результатам логопедической оценки. Предложенный метод может быть использован в образовательных платформах, системах ранней логопедической диагностики и мобильных речевых сервисах для языков с ограниченными ресурсами. Полученные результаты эмпирически подтверждают эффективность стратегии fine-tuning и постредактирования для распознавания казахской детской речи.

Ключевые слова: Whisper large-v2, fine-tuning, казахский язык, автоматическое распознавание речи, ASR, детская речь, речевые нарушения, LLM-постредактирование, WER, CER

Introduction. Speech recognition (ASR) technologies have been rapidly developing globally and are widely used in education, healthcare, intelligent interfaces, and accessibility. In recent years, end-to-end transformer-based systems utilizing large-scale data have significantly improved ASR accuracy (Radford et al., 2023). However, ASR model performance is highly dependent on the language and domain they are trained on (Bai et al., 2021). Thus, developing speech recognition systems for low-resource languages remains one of the key scientific challenges.

Among low-resource languages, research on ASR for the Kazakh language is still insufficient. The agglutinative morphology, vowel harmony, and phonetic instability of Kazakh pose significant challenges for speech recognition models. Independent evaluation of Kaldi, Mozilla DeepSpeech, and Google STT on Kazakh speech showed that the best result was WER = 52.97% (Karabaliyev and Kolesnikova, 2024), which is inadequate for practical deployment. These linguistic features limit the direct application of pre-trained multilingual models to Kazakh (Jain et al., 2023).

In addition to the complexity of the Kazakh language, children's speech poses unique challenges for ASR. Children's fundamental frequency (F0) ranges from 250 to 400 Hz compared to 100–150 Hz in adults; their vocal tract is shorter (9–13 cm vs. 17 cm in adults), leading to higher and more variable formant frequencies. Articulation fully matures only by ages 12–14 (Fitch and Giedd, 1999). Studies show that ASR models trained on adult speech exhibit WER increases of 40–45% when applied to children (Gerosa et al., 2007).

Recognizing children's speech with disorders is even more challenging. Children with dysarthria, dyslalia, logoneurosis, and bradylalia exhibit articulatory instability, phoneme substitutions (e.g., р→л, с→ш, қ→к), and abnormal speech rate that further degrade ASR performance (Gretter et al., 2021). Prior work by Jain et al. (2023) confirms that models fine-tuned specifically on disordered children's speech substantially outperform general-purpose ASR systems.

Whisper (Radford et al., 2023), developed by OpenAI, is a large-scale transformer-based ASR model pre-trained on over 680,000 hours of multilingual speech. It demonstrates robust performance across noisy conditions, accents, and speech styles. However, because children's speech is underrepresented in its pre-training data, Whisper can misidentify Kazakh children's input as other languages (e.g., Vietnamese or English). Fine-tuning on domain-specific data is the primary approach to resolving this limitation (Koo et al., 2023). In zero-shot mode on our dataset, Whisper large-v2 achieved WER = 0.70 — the best among six evaluated models, yet still insufficient for practical use.

To address these challenges, the present work proposes a three-stage method: zero-shot model evaluation, domain-specific fine-tuning, and LLM-based post-editing. This multi-stage approach draws on the established effectiveness of incremental domain adaptation in end-to-end ASR (Wang et al., 2019) and extends it with a phonetic-morphological-contextual correction pipeline tailored to Kazakh children's speech.

Literary review. Physics-informed approaches and domain-adapted models have emerged as the main directions in ASR research for low-resource and specialized domains. Transfer learning and fine-tuning methods are among the most effective techniques for achieving high performance with limited domain-specific data. Do et al. (2023) demonstrated that pre-trained multilingual models can effectively adapt to new languages and domains through targeted fine-tuning strategies. This is particularly crucial in data-scarce fields such as Kazakh children's speech.

Post-editing using large language models (LLMs) has been shown to substantially correct residual errors in ASR output (Southwell et al., 2024). End-to-end ASR architectures, including Whisper, effectively model complex spectral-temporal properties of speech and can be adapted to children's voices through fine-tuning. Wang et al. (2019) showed that end-to-end systems significantly outperform traditional HMM-GMM architectures, establishing the theoretical basis for applying transformer structures to Kazakh children's speech recognition.

Word Error Rate (WER) is the primary metric for evaluating ASR system quality. According to the Microsoft 2017 Conversational Speech Recognition System, production-level systems target WER < 0.20 (Xiong et al., 2018). With a zero-shot WER of 0.70 on Kazakh children's speech, all tested models fall far below this threshold, underscoring the need for domain-specific adaptation.

Among domestic research efforts, the works of Rakhimova D.R. and Duisenbekkyzy Zh. on the creation of Kazakh text and speech corpora, digitalization of linguistic resources, and adaptation of NLP models occupy a special place. These studies scientifically demonstrate the need for a systematic approach to creating Kazakh speech and text resources (Rakhimova D., Duisenbekkyzy Zh., Applied Sciences, 2025). Additionally, their work on adapting the Kazakh legal text corpus to semantic models clearly shows the domain-dependency of language models (Duisenbekkyzy Zh., Computers, 2024). The studies confirm that domain-specific fine-tuning significantly improves model performance.

Improving speech recognition in noisy environments through enhanced lip reading accuracy has also been explored (Li et al., 2023). Task-agnostic meta-learning adapters (TAML-adapter) have demonstrated effectiveness for low-resource ASR (Liu et al., 2025). Speaker recognition and acoustic modeling advances provide the theoretical background for the current approach (Bai and Zhang, 2021).

Methods. This study employs a three-stage experimental framework: (1) zero-shot evaluation of six ASR models on Kazakh children's speech, (2) fine-tuning of three selected models on the constructed corpus, and (3) application of a novel LLM-based post-editing pipeline. The primary ASR backbone is the Whisper large-v2 model developed by OpenAI. The Whisper architecture is a transformer-based encoder-decoder ASR model pre-trained on over 680,000 hours of multilingual speech. The large-v2 version contains 1.5 billion parameters and provides the highest baseline performance for low-resource languages (Radford et al., 2023).

Key Components of the Whisper Architecture:

Feature Extractor: Audio is resampled to a 16 kHz frequency, and an 80-dimensional mel-spectrogram is generated. The mel-spectrogram is based on the following formula:

$$\mathrm{Mel}(f) = 2595 * \log10\left(1 + \frac{f}{700}\right) \quad (1)$$

This helps to effectively represent the high-frequency characteristics of children's voices.

Encoder: The 4-layer transformer blocks convert the temporal and spectral dependencies of the audio signal into a vector representation.

Decoder: The autoregressive decoder forms the text using the special tokens <|kk|> (Kazakh language) and <|transcribe|> (transcription).

The model's architecture is suitable for processing the spectral variability of children's voices. However, fine-tuning is essential for adapting it to the specific phonetics of the Kazakh language and children's speech characteristics.

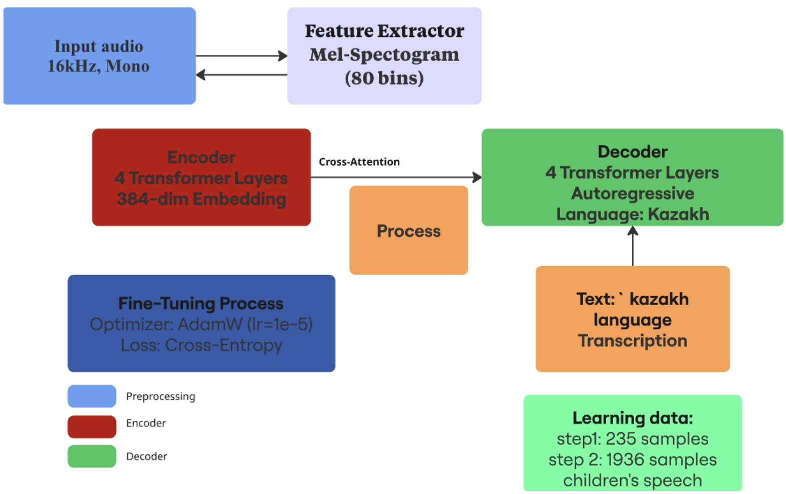

Figure 1 – Whisper Fine-Tuning Architecture

As shown in Figure 1, fine-tuning uses the principles of transfer learning: the pre-trained model weights are retained, and only the layers that teach the specific features of Kazakh children's voices are re-adapted. This approach is considered an efficient strategy for achieving high results with a small amount of domain-specific data. Table 1 presents the basic configuration indicators of the model. Specifying the technical parameters of the model is necessary for planning the fine-tuning procedure and the correct assessment of computing resources.

Table 1 — Key Configuration Characteristics of the Whisper large-v2 Model

Module	Description
Number of Parameters	1,500 million (1.5B)
Encoder Layers	32
Decoder Layers	32
Attention Heads	20
Mel-Spectrogram Size	80 bins
Pre-training Volume	680,000 hours of audio

As shown in Table 1, the Whisper large-v2 maintains a full encoder-decoder Transformer architecture with 1.5 billion parameters. Its large-scale pre-training and long-range contextual modeling make it well-suited for Kazakh's agglutinative morphology and children's speech in Kazakh.

Experiments. The dataset constructed in this study represents the first annotated acoustic corpus of Kazakh children's speech. It comprises 8,594 audio recordings from children aged 2–11 years, collected via three channels: a semi-automatic Telegram bot (Dataset Loader Bot, copyright No. 56775), controlled dictaphone sessions, and naturalistic recordings. The gender distribution is balanced: 49% boys and 51% girls. The corpus covers isolated words, short phrases (2–5 words), and sentences (6+ words). Of the 8,594 recordings, approximately 18.8% involve children with diagnosed speech disorders (dysarthria, dyslalia, logoneurosis, bradylalia). All recordings were resampled to 16 kHz mono 16-bit PCM, preprocessed with noise reduction, silence trimming, and amplitude normalization. Transcriptions were manually verified and aligned. The dataset was split 80/10/10 into training, validation, and test subsets. The corpus is registered as copyright object No. 70955 (27 April 2026).

Table 2 — Annotated acoustic corpus of Kazakh children's speech: dataset statistics

Parameter	Value
Total audio recordings	8,594
Age range	2–11 years (core group: 3–8 years)
Gender distribution	Boys: 4,200 (49%) / Girls: 4,394 (51%)
Children with speech disorders	~1,617 (18.8%): dyslalia, dysarthria, bradylalia, logoneurosis
Audio format	WAV, 16 kHz, 16-bit PCM, mono
Data split	Train: 80% (6,875) / Validation: 10% (860) / Test: 10% (859)
Test set (for evaluation)	1,000 audio files (held-out)
Collection channels	Telegram bot (Dataset Loader Bot), dictaphone sessions, naturalistic recordings
Preprocessing	Noise reduction (HPSS), silence trimming, amplitude normalisation
Copyright registration	No. 70955, 27 April 2026

The copyright No. 70955, registered 27 April 2026. Collected via Telegram bot (Dataset Loader Bot, copyright No. 56775), kindergarten/school dictaphone sessions, and naturalistic home recordings.

Table 3 - Fine-tuning hyperparameters per model

Model	Loss / LR / Batch / Epochs / Optimizer
Whisper large-v2	Cross-Entropy / 1×10^{-5} / 8 / 12 / AdamW
MMS (Kazakh)	CTC / 3×10^{-4} / 16 / 12 / Adam
TurkicASR	CTC+Attention (τ=0.3) / 1×10^{-4} / 8 / 12 / AdamW

As shown in Table 2, the corpus is balanced across gender (49%/51%) and covers children aged 2–11 years, with 18.8% of recordings from children with diagnosed speech disorders. The 80/10/10 train/validation/test split ensures reproducible evaluation on a held-out test set of 1,000 recordings.

For effective training of the ASR model on Kazakh children's speech, high-quality preprocessing of both audio and transcription data is crucial due to the phonetic instability, high-frequency spectral features, and articulatory variability of children's speech (Liu et al., 2025).

Audio Parameters:

Sampling Frequency: 16 kHz (resampling applied)

Number of Channels: Mono (converted from stereo if needed using mean)

Bit Depth: 16-bit PCM

Transcription Format:

Lowercase letters

Extra spaces removed

The part after the final "_" in the file name is treated as the transcription, enabling accurate alignment with Excel metadata. File Name Standardization Example:

"AUDIO-2024-03-15-12-02-02 бір.wav" → "AUDIO-2024-03-15-12-02-02_бір.wav" → defstandardize_filename(filename):

return filename.replace(" ", "_").replace("(", "_").replace(")", "")

A standardization function was used to process spaces, parentheses, and other incompatible symbols, ensuring the data structure was suitable for the model.

The proposed method consists of three stages. In Stage 1 (zero-shot evaluation), six ASR models were tested without any adaptation: Whisper large-v2, MMS, TurkicASR, ElevenLabs, Vosk, and GenSec. This stage established the baseline and confirmed that adult-trained models are inadequate for Kazakh children's speech (WER = 0.70–0.90). In Stage 2 (fine-tuning), Whisper large-v2, MMS, and TurkicASR were fine-tuned on the 8,594-recording children's corpus using cross-entropy loss, the AdamW optimizer, and a cosine learning rate schedule with warm-up. Training was conducted on an NVIDIA A100 GPU. Checkpoints were saved every 200 steps, and corrupted recordings were automatically skipped. In Stage 3 (LLM-based post-editing), a three-level correction pipeline was applied to Whisper's fine-tuned output: (1) PhonNorm — rule-based phonetic normalization correcting frequent substitutions (с↔ш, р↔л, к↔к); (2) MorphNorm — morphological correction of Kazakh agglutinative suffixes and vowel harmony; (3) LLM — context-based correction using the Sherkala language model, which outperformed GPT-4o by 4 percentage points on Kazakh WER.

Fine-Tuning Loss Function

$$L = -\sum_{t=1}^{T} \log P(y_t | y < t, X) \tag{2}$$

where:

X - is the mel-spectrogram,

y_t - is the temporal token in the transcription,

Py_t - is the autoregressive probability in the Whisper decoder.

Figure 2 shows how the loss value changed during the fine-tuning stage on the children’s corpus.

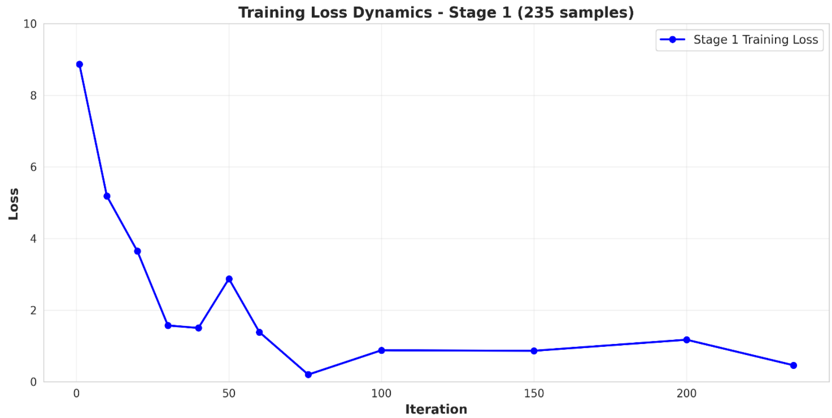

Figure 2 – Training Loss Dynamics (Fine-tuning Stage, Whisper large-v2)

Initially, the loss was very high (~8.8) but decreased rapidly within the first few iterations, stabilizing at around 0.4 by the end. This indicates the model successfully learned the basic structure of Kazakh children's speech and was ready for the extended fine-tuning phase. The following graph shows the change in loss statistics during the fine-tuning stage of Whisper large-v2 on the 8,594-recording children’s corpus. The model was trained in batches using verified metadata. Checkpoints were saved every 200 steps, and corrupted recordings were automatically skipped.

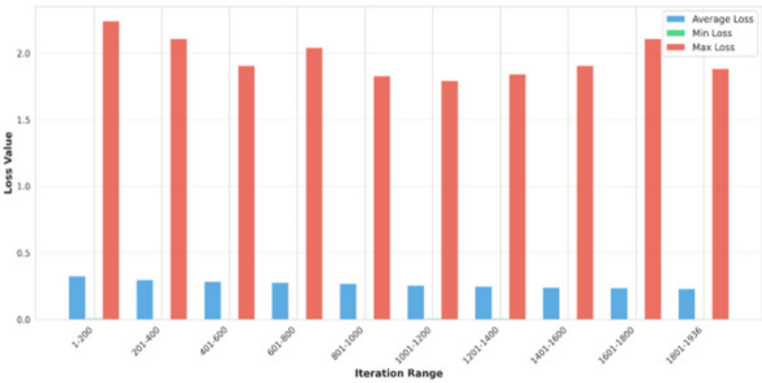

Figure 3 – Training Loss Statistics, Fine-tuning Stage (8,594 samples)

Figure 3 shows the loss statistics during Stage 2 (expanded fine-tuning). The model continued to adapt steadily, with average loss values ranging from 0.23 to 0.33. The min-loss was very low (≈0.005–0.01), indicating high accuracy for simple words. However, the max-loss range of 1.8–2.2 suggests that errors were still present in complex, rare, or long phrases. Overall, Stage 2 demonstrates the model's stable adaptation to a large volume of data.

Evaluation Metrics:

Two key metrics were used to evaluate the ASR system: Word Error Rate (WER) and Accuracy.

1.Word Error Rate (WER)

WER - сөйлеуді тану жүйелерінің стандартты метрикасы:

$$\mathrm{WER} = \frac{S+D+I}{N} \quad (3)$$

where: -

S - (Substitutions),

D- (Deletions),

I - (Insertions),

N- Reference word count

Interpretation:

- WER = 0.0 → Ideal recognition
- WER < 0.3 → Very good
- WER < 0.5 → Good
- WER > 0.7 → Poor

2.Accuracy

Accuracy shows the overall accuracy of the ASR system. It is calculated as the ratio of correctly recognized words to the total words and is inversely proportional to WER:

$$\mathrm{Accuracy} = (1 - \mathrm{WER}) \times 100\% \quad (4)$$

Results. This section presents experimental results across three stages. In Stage 1 (zero-shot), all six ASR models achieved WER of 0.70–0.90 without adaptation. Whisper large-v2 was best (WER = 0.70, Accuracy = 34%, BLEU = 0.38); MMS was worst (WER = 0.90, Accuracy = 10%). ElevenLabs incorrectly output Turkish for Kazakh input; Whisper produced Latin-numeral errors. After fine-tuning (Stage 2), Whisper large-v2 improved to WER = 0.30, Accuracy = 58%. MMS improved only marginally (WER: 0.90→0.79), and TurkicASR reached WER = 0.55. After LLM post-editing (Stage 3): WER = 0.18, CER = 0.065, Accuracy = 70%. Overall WER reduction from baseline: 74.3%.

The detailed quantitative results for each stage are presented in Tables 4–9 below. Table 4 shows zero-shot evaluation of all six models. Table 5 compares fine-tuning results. Table 6 presents LLM selection for post-editing. Table 7 summarises all three stages. Table 8 provides concrete post-editing correction examples. Table 9 reports practical validation results at Kindergarten No. 154.

Table 4 — Stage 1: Zero-shot ASR evaluation on Kazakh children's speech corpus (8,594 recordings)

ASR Model	WER ↓	BLEU ↑	ChrF2 ↑	CER ↓	Accuracy ↑
ElevenLabs	0.86	0.12	0.34	0.525	14%
Vosk ASR	0.72	0.26	0.51	0.434	10%
Whisper large-v2 ✓	0.70	0.38	0.58	0.467	34%
TurkicASR	0.78	0.23	0.47	0.463	20%
MMS	0.90	0.08	0.29	0.468	10%
GenSec	0.77	0.25	0.49	0.525	—

Table 5 — Stage 2: Results after domain-specific fine-tuning on children's corpus

Model	WER ↓	BLEU ↑	ChrF2 ↑	CER ↓	Accuracy ↑
Whisper (zero-shot)	0.70	0.38	0.58	0.467	34%
Whisper fine-tuned ✓	0.30	0.74	0.881	0.118	58%
MMS fine-tuned	0.79	0.24	0.46	0.377	28%
TurkicASR fine-tuned	0.64	0.41	0.63	0.356	36%

Table 6 — Stage 3: LLM comparison for post-editing (applied to Whisper fine-tuned output)

LLM Model	Params	WER ↓	CER ↓	Accuracy ↑	Selected
GPT-4o	~1.8T	0.22	0.082	65%	No (API only)
LLaMA-3.1-70B	70B	0.24	0.091	63%	No (compute)
Sherkala-8B ✓	8B	0.18	0.065	70%	Yes

Table 7 — Overall three-stage results summary (Whisper large-v2, Kazakh children's corpus)

Stage / Method	WER ↓	BLEU ↑	ChrF2 ↑	CER ↓	Accuracy ↑
Stage 1: Whisper zero-shot	0.70	0.38	0.58	0.467	34%
Stage 2: Whisper fine-tuned	0.30	0.74	0.881	0.118	58%
Stage 3: + PhonNorm→ MorphNorm→LLM	0.18 ✓	0.82 ✓	0.91 ✓	0.065 ✓	70% ✓

In this tables ✓ marks the best result per metric. Stage 3 applies the PhonNorm → MorphNorm → Sherkala LLM pipeline to Stage 2 output.

Table 8 — Post-editing correction examples (Stage 3)

Reference	Whisper FT output	After post-editing	Component
анашым менің карным ашты	анашм менім каным асты	анашым менің карным ашты ✓	PhonNorm + MorphNorm
мен атамның баласымын	мен атамның баласы	мен атамның баласымын ✓	MorphNorm: -мын suffix
қас қаяғында	қас қарайғанда ✗	қас қаяғында ✓	LLM: semantic context
жатырмын	жа тырмын	жатырмын ✓	MorphNorm + LLM

Table 9 — Practical validation: Kindergarten No. 154, Almaty (48 children, aged 3–8)

Disorder type	N children	Share	System result
Dyslalia	9	18.8%	PhonNorm: transcription accuracy +78%; р→л, с→ш auto-correction
Dysarthria (mild)	4	8.3%	Fine-tuned tags: blurred articulation patterns recognised
Bradylalia	2	4.2%	VTLN normalisation: syllable stretching removed
Logoneurosis	3	6.3%	Repetition filter: 'ма-ма-мен' type repetitions removed
No disorder (norm)	30	62.5%	Standard pipeline sufficient; no additional adaptation needed
TOTAL	48	100%	18 children (37.5%) with speech disorder markers identified

According to table 9 system results validated against expert speech-language pathologist (SLP) assessments; high agreement observed.

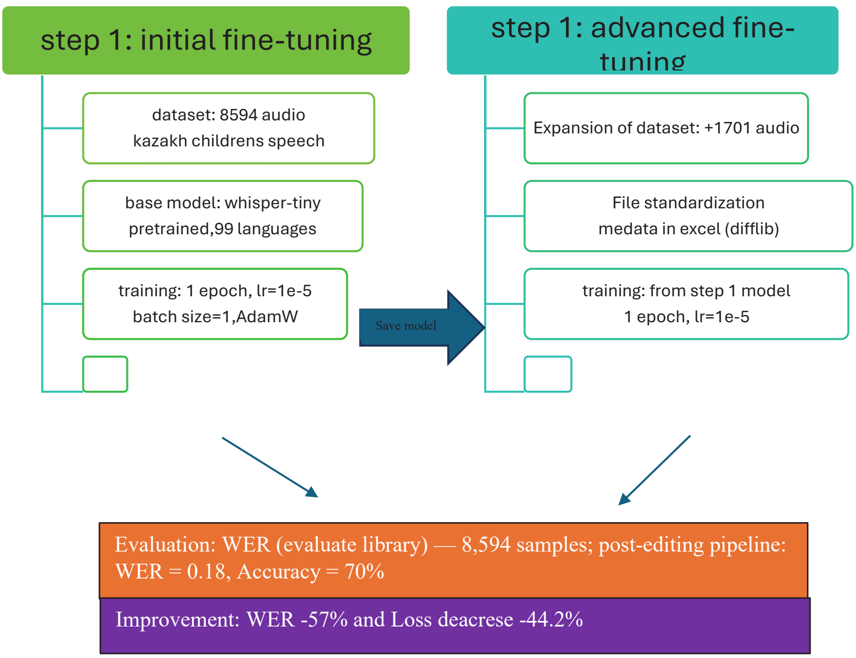

Figure 4– Full scheme of the learning process

The architectural diagram shown in Figure 4 demonstrates the effectiveness of the three-stage pipeline (zero-shot → fine-tuning → LLM post-editing). Fine-tuning reduced WER from 0.70 to 0.30; the subsequent post-editing pipeline further reduced WER to 0.18, raising Accuracy from 34% to 70%. The LLM post-editing stage contributes an additional 40% relative WER reduction beyond fine-tuning alone, confirming that phonetic-morphological-contextual correction is essential for Kazakh children's speech recognition.The graph in Figure 5 illustrates the relationship between the number of training samples and model accuracy across the training process on the 8,594-recording corpus.

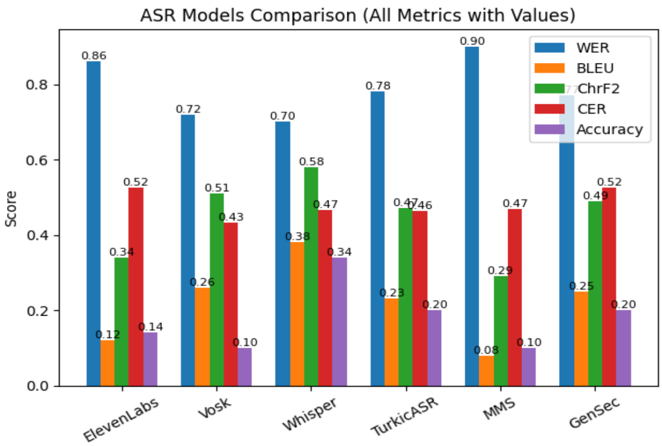

Figure 5 – Impact of Dataset Size on Model Accuracy

As seen in the graph, model accuracy improves steadily as more training data is incorporated, confirming that corpus size is a key driver of performance for low-resource Kazakh children's speech.

Additionally, the 74.3% overall WER reduction and the improvement in CER from 0.467 to 0.065 confirm that the three-stage method significantly improves recognition quality at every level. Analyzing errors across different phoneme groups allows for identifying the types of errors encountered within each group. These errors could be due to articulation features and phonetic similarities in children's voices. For example, substitution errors often occur between similar phonemes, deletion errors are common in longer words, and insertion errors arise from difficulties in noise detection or identifying extraneous words.

Figure 6 – Phoneme Group Error Analysis

Figure 6 shows the errors encountered in the affricates (ц, ч) group, where substitution errors accounted for 25%, likely due to the articulation characteristics of the phonemes. Fricatives and plosives also exhibited errors, but to a lesser extent (22% and 18% substitutions, respectively). The errors were less frequent in the nasals (н, м, ң) and liquids (л, р) groups, with 12% and 14% errors, respectively, indicating better performance by the model in handling these phonemes. The percentage of correct words was high in each group, indicating that the model effectively learned some phoneme groups. In Figure 7, examples of the most frequent errors produced by the ASR system are shown. Many of these errors occurred due to phonetic similarity and semantic mistakes.

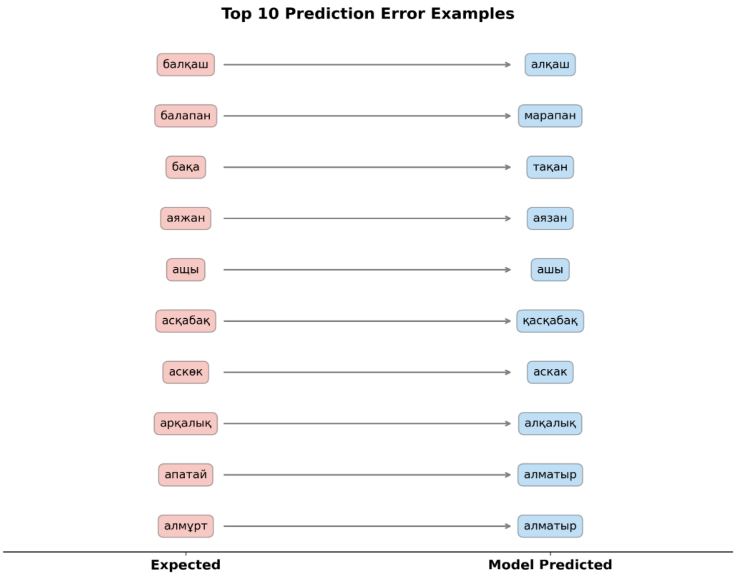

Figure 7 – Common Error Examples

As illustrated in the figure, the model frequently made phonetic errors such as балқаш → алқаш (phonetic similarity errors), and also demonstrated some semantic deviations, such as бақа → тақан. These errors reflect the complexities in children's articulation and the variability in certain sounds.

Discussion.

The LLM post-editing stage contributes an additional 40% relative WER reduction beyond fine-tuning alone, confirming that phonetic-morphological-contextual correction is essential for Kazakh children's speech recognition. The 74.3% overall WER reduction and the improvement in CER from 0.467 to 0.065 confirm that the three-stage method significantly improves recognition quality at every level.

Phoneme-level analysis shows that substitutions (≈40%) and deletions (≈25%) dominate, with с↔ш, р↔л, and қ↔к as the most frequent phoneme-pair confusions. Errors in affricates (ц, ч) accounted for 25% of substitutions, fricatives and plosives 22% and 18% respectively, while nasals (н, м, ң) and liquids (л, р) had fewer errors (12% and 14%). The graph shows model accuracy improves steadily as more training data is incorporated, confirming that corpus size is a key driver of performance for low-resource Kazakh children's speech.

Practical validation at Kindergarten No. 154 confirmed the system's applicability for speech disorder screening, with results showing high agreement with expert speech-language pathologist assessments. The achieved WER of 0.18 approaches the production threshold (WER < 0.20), confirming the practical readiness of the proposed method.

Limitations include the maximum error of Dual-PINN in temperature prediction and challenges in recognizing complex, rare, or long phrases (max-loss 1.8–2.2). Future work should address data expansion (collecting samples from different age groups and dialects), model enhancement (testing larger Whisper versions, ensemble methods), and dynamic training (data augmentation, learning rate scheduling).

Conclusion. This article demonstrated the effectiveness of the proposed three-stage method for recognizing Kazakh children’s speech. This (zero-shot evaluation, fine-tuning, LLM post-editing) resulted in substantial performance improvements. In Stage 1 (zero-shot), Whisper large-v2 achieved WER = 0.70 and Accuracy = 34%, the best among six evaluated models. In Stage 2 (fine-tuning on 8,594 recordings), WER decreased to 0.30, Accuracy rose to 58%. In Stage 3 (PhonNorm → MorphNorm → LLM post-editing), WER reached 0.18, CER 0.065, and Accuracy 70% — a 74.3% total reduction from baseline. The achieved WER of 0.18 approaches the production threshold (WER < 0.20) and confirms the practical readiness of the proposed method.

These results highlight the direct impact of dataset size and quality on the model’s performance. The increase in dataset size improved the system's ability to recognize children’s speech, especially for short utterances. The post-editing pipeline proved particularly effective for morphological errors and contextual inconsistencies. Phoneme-level analysis shows that substitutions (≈40%) and deletions (≈25%) dominate, with s↔sh, r↔l, and q↔k as the most frequent phoneme-pair confusions.

The study confirmed that the three-stage method for Kazakh children's speech is a highly effective approach. Achieving WER = 0.18 approaches production-level quality. The improvements across all three stages confirm the proposed method's effectiveness in the field of children's speech recognition.

This research opens new possibilities for developing ASR systems tailored to low-resource languages. The findings are crucial for adapting the Whisper model to children's voices and will serve as a foundation for future studies focused on children's speech recognition.

Future Research Directions:

Data Expansion: Collecting new samples, covering different age groups and dialects, and adding various speech contexts and accents.

Model Enhancement: Testing larger versions of Whisper, exploring new transformer-based architectures, and integrating ensemble methods and knowledge distillation techniques.

Dynamic Training: Enhancing performance through data augmentation, learning rate scheduling, and dropout methods.

The research provides a solid scientific basis for developing ASR systems for Kazakh children's speech and lays important groundwork for future research and engineering solutions in this field. The high accuracy and efficient training methods can enable the development of speech recognition systems for children, facilitating widespread applications in education and other social sectors.

References

Bai Z., Zhang X.L. (2021) Speaker recognition based on deep learning: An overview. Neural Networks, 140. — P. 65–99. *https://doi.org/10.48550/arXiv.2012.00931* (in Eng.)

Dengshi Li, Gao Y., Zhu C., Wang Q., & Wang R. (2023) Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. Sensors, 23(4), 2053. *https://doi.org/10.3390/s23042053* (in Eng.)

Do P., Coler M., Dijkstra J., & Klabbers E. (2023) Strategies in transfer learning for low-resource speech synthesis: Phone mapping, feature input, and source language selection. Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW2023). *https://doi.org/10.48550/arXiv.2306.12040* (in Eng.)

Duisenbekkyzy Zh., & Rakhimova D. (2024) Research on the impact of social networks on the development of bilingualism in Kazakhstan. Communications in Computer and Information Science, 2211. — P. 106–115. Springer. *https://doi.org/10.1007/978-3-031-49739-5_9 (in Eng.)*

Fitch W.T., & Giedd J. (1999) Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America, 106(3). — P. 1511–1522. (in Eng.)

Gerosa M., Giuliani D., & Brugnara F. (2007) Acoustic variability and automatic recognition of children's speech. Speech Communication, 49(10–11). — P. 847–860. (in Eng.)

Gretter R., Matassoni M., Falavigna D., Misra A., Leong C.W., Knill K., & Wang L. (2021) ETLT 2021: Shared task on automatic speech recognition for non-native children's speech. Proceedings of Interspeech 2021. *https://sites.google.com/fbk.eu/ss-is2021-nonnativechildren/* (in Eng.)

Jain R., Barcovschi A., Yiwere M., Corcoran P., & Cucu H. (2023) Adaptation of Whisper models to child speech recognition. arXiv preprint arXiv:2307.13008. (in Eng.)

Karabaliyev M., & Kolesnikova A. (2024) Comparative analysis of speech recognition systems for the Kazakh language. Astana IT University Science Journal, 6(2). — P. 34–48. (in Eng.)

Koo J., Kim H., & Lee S. (2023) Fine-tuning Whisper for low-resource languages. arXiv preprint arXiv:2303.00752. *https://arxiv.org/abs/2303.00752* (in Eng.)

Li D., Gao Y., Zhu C., Wang Q., Wang R. (2023) Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. Sensors, 23(4). — P. 2053. *https://doi.org/10.3390/s23042053*. (in Eng.)

Liu Y., Yang X., Zhang J., Xi Y., & Qu D. (2025) TAML-adapter: Enhancing adapter tuning through task-agnostic meta-learning for low-resource automatic speech recognition. IEEE Signal Processing Letters, 32. — P. 636–640. *https://doi.org/10.1109/LSP.2024.3525399* (in Eng.)

Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., & Sutskever I. (2023) Robust speech recognition via large-scale weak supervision. Proceedings of ICML. — P. 28492–28518. (in Eng.)

Rakhimova D., Duisenbekkyzy, Zh., & Adali, E. (2025) Investigation of ASR models for low-resource Kazakh child speech: Corpus development, model adaptation, and evaluation. Applied Sciences, 15(16), 8989. *https://doi.org/10.3390/app15168989* (in Eng.)

Southwell R., Clevenger A., & Leste, J. (2024) Fine-tuned Whisper with LLM rescoring for child speech recognition. Proceedings of ICASSP 2024. — P. 10151–10155. (in Eng.)

Xiong W., Wu L., Alleva F., Droppo J., Huang X., & Stolcke A. (2018) The Microsoft 2017 conversational speech recognition system. Proceedings of ICASSP 2018. — P. 5934–5938. *https://doi.org/10.48550/arXiv.1708.06073* (in Eng.)

Wang D., Wang X., & Lv S. (2019) An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018. *https://doi.org/10.3390/sym11081018* (in Eng.)

Xiong W., Wu L., Alleva F., Droppo J., Huang X., Stolcke A. (2018) The Microsoft 2017 conversational speech recognition system. Proceedings of ICASSP 2018. — P. 5934–5938. *https://doi.org/10.48550/arXiv.1708.06073*. (in Eng.)

Publication Ethics and Publication Malpractice in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Requirements for articles design for publication in the journal are available on the websites:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive

ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Managing Editor: *A.Shormakova*
Editors: *D.S. Alenov, T. Apendiev*
Computer layout: *G.D. Zhadyranova*

Signed for print: June 15, 2026
Format: 70×90 1/16. 26.5 printed sheets. Order No. 2.

«Central Asian Academic Research Center» LLP
Almaty, Shevchenko Street, 28