2025;(4):56–78 Арктика XXI век / Arctic XXI Century

УДК 004.8; 81’32
https://doi.org/10.25587/3034-7378-2025-4-56-78

Original article

Transformer-Based Neural Network Approaches for Speech Recognition and Synthesis in the Sakha Language

***Sergei P. Stepanov*[1] ✉, *Dong Zhang*[2], *Timur Z. Zakharov*[1], *Altana A. Alekseeva*[1], *Vladislav L. Aprosimov*[1], *Djuluur A. Fedorov*[1], *Vladimir S. Leveryev*[1] , *Tuygun A. Novgorodov*[1], *Ekaterina S. Podorozhnaya*[1]**

[1]M.K. Ammosov North-Eastern Federal University,
Yakutsk, Russian Federation
[2]Qufu Normal University, Qufu, Shandong, P.R. China
✉ sp.stepanov@s-vfu.ru

Abstract

Recent breakthroughs in artificial intelligence and deep learning have fundamentally transformed the landscape of spoken language processing technologies. Automatic speech recognition (ASR) and text-to-speech (TTS) synthesis have emerged as essential components driving digital accessibility across diverse linguistic communities. The Sakha language, representing the northeastern branch of the Turkic language family, continues to face substantial technological barriers stemming from insufficient digital resources, limited annotated corpora, and the absence of production-ready speech processing systems. This comprehensive investigation examines the feasibility and effectiveness of adapting contemporary transformer-based neural architectures for bidirectional speech conversion tasks in Sakha. Our research encompasses detailed analysis of encoder-decoder frameworks, specifically OpenAI’s Whisper large-v3 and Meta’s Wav2Vec2-BERT for voice-to-text transformation, alongside Coqui’s XTTS-v2 system for text-to-voice generation. Particular emphasis is placed on addressing linguistic and technical obstacles inherent to Sakha, including its complex agglutinative morphological structure, systematic vowel harmony patterns, and distinctive phonemic inventory featuring sounds absent from most Indo-European languages. Experimental evaluation demonstrates that comprehensive fine-tuning of Whisper-large-v3 achieves exceptional recognition accuracy with word error rate (WER) of 8%, while the self-supervised Wav2Vec2-BERT architecture attains 13% WER when augmented with statistical n-gram language modeling. The neural synthesis system exhibits robust performance despite minimal training data availability, achieving average loss of 2.49 following extended training optimization and practical deployment via Telegram messaging bot. Additionally, ensemble meta-stacking combining both recognition architectures achieves 27% WER, demonstrating

effective complementarity through learned hypothesis arbitration. These findings validate transfer learning methodologies as viable pathways for developing speech technologies serving digitally underrepresented linguistic communities.
Keywords: Sakha language, automatic speech recognition, text-to-speech synthesis, neural networks, Whisper, Wav2Vec2-BERT, Coqui XTTS-v2, transformer architecture, low-resource languages, transfer learning, agglutinative morphology
Funding. This research was conducted with financial support from the Artificial Intelligence Laboratory of the Republic of Sakha (Yakutia) and with the financial support of the Russian Science Foundation "Languages and Cultures of the Peoples of the North and the Arctic of the Russian Federation: Comprehensive socio-humanitarian research (on the basis of big data)" No 25-78-30006 (22.05.2025)
Acknowledgments. The authors express their gratitude to everyone who participated in data collection
For citation: Stepanov S.P., Zhang D., Zakharov T.Z., Alekseeva A.A., Aprosimov V.L., Fedorov Dj.A., Leveryev V.S., Novgorodov T.A., Podorozhnaya E.S. Transformer-Based neural network approaches for speech secognition and synthesis in the Sakha language. *Arctic XXI Century*. 2025;(4):56–78. DOI: https://doi.org/10.25587/3034-7378-2025-4-56-78

Оригинальная научная статья

Подходы к распознаванию и синтезу речи для якутского языка на основе нейросетей архитектуры Transformer

***С. П. Степанов*[1] ✉, *Д. Чжан*[2], *А. А. Алексеева*[1], *В. Л. Апросимов*[1], *Дь. А. Федоров*[1], *В. С. Леверьев*[1], *Т. А. Новгородов*[1], *Е. С. Подорожная*[1], *Т. З. Захаров*[1]**
[1]Северо-Восточный федеральный университет им. М. К. Аммосова, Якутск, Российская Федерация
[2]Цюйфуский педагогический университет, Цюйфу, Шаньдун, КНР
✉ sp.stepanov@s-vfu.ru

Аннотация
Новейшие достижения в области искусственного интеллекта и глубокого обучения кардинально преобразовали ландшафт технологий обработки устной речи. Автоматическое распознавание речи (ASR) и синтез речи (TTS) стали ключевыми компонентами, обеспечивающими цифровую доступность для различных языковых сообществ. Якутский язык, представляющий северо-восточную ветвь тюркской языковой семьи, продолжает сталкиваться со значительными технологическими барьерами, вызванными недостаточностью цифровых ресурсов, ограниченностью размеченных корпусов и отсутствием

готовых к промышленному использованию систем обработки речи. В данном комплексном исследовании изучается целесообразность и эффективность адаптации современных нейросетевых архитектур на основе трансформеров для задач двунаправленного речевого преобразования в якутском языке. Наша работа включает детальный анализ encoder-decoder моделей, а именно: Whisper large-v3 от OpenAI и Wav2Vec2-BERT от Meta для преобразования голоса в текст, а также системы XTTS-v2 от Coqui для генерации речи из текста. Особое внимание уделяется решению лингвистических и технических проблем, присущих якутскому языку, включая его сложную агглютинативную морфологическую структуру, системные законы сингармонизма и уникальный фонемный состав, содержащий звуки, отсутствующие в большинстве индоевропейских языков. Экспериментальная оценка показывает, что полное дообучение модели Whisper-large-v3 обеспечивает исключительно высокую точность распознавания с коэффициентом ошибок по словам (WER) 8%, в то время как самообучаемая архитектура Wav2Vec2-BERT достигает WER 13% при использовании статистического n-граммного языкового моделирования. Нейросетевая система синтеза демонстрирует устойчивую производительность даже при ограниченном объеме обучающих данных, достигая среднего значения функции потерь 2,49 после длительной оптимизации обучения и практического развертывания через бот в мессенджере Telegram. Кроме того, ансамблевый мета-стэкинг, объединяющий обе архитектуры распознавания, позволяет достичь WER 27%, что доказывает их эффективную взаимодополняемость через арбитраж гипотез. Полученные результаты подтверждают, что методы трансферного обучения представляют собой жизнеспособный путь для создания речевых технологий, обслуживающих цифрово недостаточно представленные языковые сообщества.

Ключевые слова: якутский язык, автоматическое распознавание речи, синтез речи из текста, нейронные сети, Whisper, Wav2Vec2-BERT, Coqui XTTS-v2, архитектура Transformer, малоресурсные языки, трансферное обучение, агглютинативная морфология

Финансирование. Данное исследование было выполнено при финансовой поддержке Лаборатории искусственного интеллекта Республики Саха (Якутия), а также при финансовой поддержке РНФ в рамках реализации проекта «Языки и культуры народов Севера и Арктики РФ: комплексные социогуманитарные исследования (на основе анализа больших данных)» № 25-78-30006 от 22.05.2025 г.

Благодарности. Авторы выражают благодарность всем, кто участвовал в сборе данных

Для цитирования: Степанов С.П., Чжан Д., Захаров Т.З., Алексеева А.А., Апросимов В.Л., Федоров Дь.А., Леверьев В.С., Новгородов Т.А., Подорожная Е.С. Подходы к распознаванию и синтезу речи для якутского языка на основе нейросетей архитектуры Transformer. *Арктика XXI век.* 2025;(4):56–78 (на англ.). DOI: https://doi.org/10.25587/3034-7378-2025-4-56-78

Introduction

The proliferation of voice-enabled computing interfaces has fundamentally reshaped human-machine interaction paradigms across virtually every domain of modern life. Intelligent virtual assistants, automated transcription services, real-time translation systems, and accessibility tools for individuals with disabilities increasingly depend on sophisticated speech processing algorithms capable of accurately interpreting and generating natural language in spoken form. These technologies have achieved remarkable performance levels for widely-spoken languages supported by extensive digital resources, yet remain largely inaccessible to speakers of linguistically marginalized communities worldwide [1].

The development of automatic speech recognition systems for underrepresented languages presents multifaceted challenges encompassing both technical and linguistic dimensions. From a technical perspective, the scarcity of annotated speech corpora, limited availability of text resources for language modeling, and absence of standardized evaluation benchmarks substantially constrain the applicability of data-intensive deep learning approaches that have proven successful for high-resource languages. Linguistically, many underrepresented languages exhibit structural properties that diverge significantly from the Indo-European languages dominating current speech technology research, necessitating specialized modeling strategies [2].

The Sakha language presents a particularly compelling case study for investigating low-resource speech technology development. As the northernmost representative of the Turkic language family, spoken by approximately 450,000 people primarily in the Republic of Sakha (Yakutia) within the Russian Federation, Sakha occupies a unique position both geographically and linguistically. The language exhibits distinctive phonological characteristics that differentiate it substantially from its Turkic relatives, having undergone extensive phonetic evolution during centuries of relative isolation in subarctic conditions [3].

Phonologically, Sakha maintains an extensive vowel inventory encompassing both short and long variants with systematic phonemic length distinctions affecting lexical meaning. The language preserves productive vowel harmony patterns governing the distribution of front and back vowels across morpheme boundaries, creating long-distance phonotactic dependencies that pose challenges for acoustic modeling approaches optimized for languages lacking such phenomena. Additionally, Sakha has developed several consonantal innovations absent from other Turkic languages, including distinctive realizations of historical uvular and velar segments [4].

Orthographically, the Sakha writing system employs several graphemes absent from standard Russian Cyrillic, representing sounds requiring specialized acoustic modeling: '*ө*' denoting the front rounded mid vowel [œ], '*ү*' representing the front rounded high vowel [y], '*ҥ*' marking the velar nasal consonant [ŋ], '*ҕ*' indicating the voiced uvular fricative [ʁ], and '*һ*' signifying the glottal fricative [h]. Automatic recognition systems frequently confuse these characters with visually or phonetically similar alternatives from Russian, substantially degrading transcription accuracy and producing outputs that violate Sakha phonotactic constraints. Phonologically, Sakha maintains an extensive vowel inventory encompassing four diphthongs, short and long variants with systematic phonemic length distinctions affecting lexical meaning [5; 6].

Furthermore, Sakha demonstrates pronounced agglutinative morphological structure characteristic of Turkic languages, wherein lexical stems concatenate with extensive chains of suffixes encoding diverse grammatical information including case, number, possession, tense, aspect, mood, and evidentiality. This productive word-formation mechanism generates substantial vocabulary expansion, with theoretically unlimited numbers of distinct word forms derivable from individual roots. Such morphological complexity challenges conventional recognition approaches relying on fixed lexicons, as out-of-vocabulary rates become prohibitively high without appropriate subword modeling strategies [7].

This investigation pursues dual complementary objectives: systematically evaluating state-of-the-art neural architectures for Sakha speech processing and establishing reproducible benchmarks facilitating future research advancement. We comprehensively assess recognition performance across multiple training configurations examining the contributions of individual architectural components, while simultaneously developing synthesis capabilities demonstrating viable quality despite severely limited acoustic training data. The methodological framework and empirical findings presented herein provide foundational resources supporting continued development of speech technologies serving the Sakha-speaking community.

Related Works

Substantial research investment has driven remarkable progress in speech technology capabilities over recent decades, though benefits remain unevenly distributed across the world's linguistic diversity. This section surveys relevant prior work addressing speech recognition and synthesis for resource-constrained languages, with particular attention to approaches applicable to Turkic language family members sharing structural similarities with Sakha.

Neural Approaches to Automatic Speech Recognition

Deep learning methodologies have fundamentally restructured automatic speech recognition system design over the past decade. Traditional hybrid architectures combining hidden Markov [8] models with Gaussian mixture acoustic models have progressively yielded to unified neural frameworks capable of direct acoustic-to-linguistic mapping without intermediate phonetic representations. Contemporary end-to-end systems leverage attention mechanisms enabling dynamic alignment between variable-length acoustic observation sequences and corresponding textual outputs [9].

Self-supervised representation learning has emerged as particularly promising paradigm for resource-constrained recognition scenarios. These approaches exploit large quantities of unlabeled audio through contrastive or predictive pretraining objectives, learning general-purpose acoustic representations subsequently fine-tuned on limited transcribed data for specific target languages. Research demonstrates that representations acquired through such procedures transfer effectively across typologically diverse languages, enabling competitive recognition performance with dramatically reduced annotation requirements [10].

The Whisper system developed by OpenAI represents current state-of-the-art in multilingual speech recognition, trained on approximately 680,000 hours of weakly-supervised audio-text pairs harvested from internet sources spanning nearly 100 languages. The architecture implements sequence-to-sequence transduction via transformer encoder-decoder networks [11], demonstrating remarkable robustness to acoustic variation including background noise, reverberation, and speaker diversity. However, performance varies substantially across languages depending on representation within training data, with many underrepresented languages receiving minimal coverage [12].

Speech Technology for Turkic Languages

Languages belonging to the Turkic family share numerous structural properties presenting systematic challenges for speech technology adaptation. Vowel harmony systems require acoustic models capable of capturing long-distance phonotactic dependencies spanning multiple syllables, as vowel quality in suffixes depends on preceding vowels within the same phonological word. Agglutinative morphology produces extensive lexical inventories with highly skewed frequency distributions, wherein most word types occur rarely while substantial probability mass concentrates on common function words [13].

Prior investigations have explored multilingual training strategies leveraging cross-linguistic similarities among related Turkic languages. Shared phonetic inventories, similar phonotactic constraints, and parallel morphological processes enable positive transfer during model adaptation, though language-specific fine-tuning remains essential for achieving

optimal performance. Research on Kazakh, Kyrgyz, Uzbek and Turkish has demonstrated that models pretrained on related languages substantially outperform those initialized from unrelated linguistic material [14].

Text-to-speech synthesis for Turkic languages has received comparatively limited research attention, though recent work demonstrates promising results using neural vocoder approaches [15–19]. Zero-shot synthesis strategies utilizing phonetic transcription intermediaries have shown particular promise for extending coverage to new languages without requiring parallel audio-text training data, leveraging shared phonetic spaces across related languages [20].

Sakha Language Resources and Prior Work

Despite growing interest in documenting and digitalizing the Sakha language, available speech resources remain substantially limited compared to high-resource languages and even relative to some other Turkic family members (Table 1). The Mozilla Common Voice project [21] includes a Sakha contribution comprising approximately 20 hours of crowdsourced recordings with corresponding transcriptions, representing the largest publicly available annotated corpus. Additional resources exist within academic institutions but remain restricted in accessibility.

Table 1

Overview of available Sakha language speech resources

Таблица 1

Обзор доступных речевых ресурсов якутского языка

Resource Name	Content Description	Volume	Access Status
Mozilla Common Voice (Sakha)	Crowdsourced recordings with verified transcriptions	~10 hours	Open access
NEFU Internal Corpus	Studio-quality readings of literary texts	~5 hours	Institutional
Augmented Training Set	Original recordings with noise augmentation	29,000 samples	Institutional

Prior speech technology research specifically targeting Sakha remains limited, with most relevant work addressing broader Turkic language coverage without detailed analysis of Sakha-specific challenges. This investigation aims to address this gap by providing systematic evaluation of contemporary neural approaches adapted specifically for Sakha linguistic characteristics.

Materials and Methods

Corpus Development and Data Preparation

Effective training of neural speech processing systems requires carefully curated datasets meeting stringent quality standards. Our corpus development efforts combined multiple data sources while implementing rigorous preprocessing protocols ensuring consistency across experimental conditions.

Raw audio recordings underwent systematic normalization procedures prior to model training. All files were resampled to 16 kHz mono format using high-quality interpolation algorithms implemented in the librosa library. Silence segments exceeding 500 milliseconds duration were trimmed from recording boundaries, while amplitude normalization ensured consistent volume levels across heterogeneous source materials. Audio quality filtering removed samples exhibiting excessive background noise, clipping artifacts, or unintelligible speech segments.

Textual transcriptions received comprehensive orthographic standardization addressing inconsistencies in punctuation conventions, numeral representation formats, and abbreviation usage. Character vocabulary construction enumerated all unique graphemes occurring within the training partition, enabling appropriate tokenizer configuration for Sakha-specific characters. The base dataset comprised approximately 6,710 utterance-transcription pairs drawn from Common Voice contributions and institutional recordings.

Data augmentation procedures [22] expanded training set diversity through systematic acoustic transformations. Augmentation strategies included additive noise injection at varying signal-to-noise ratios, tempo perturbation within ±10% range, and pitch shifting across semitone intervals. The augmented collection totaled approximately 29,000 training samples, substantially increasing effective dataset size while maintaining linguistic content validity.

For synthesis system development, specialized data collection targeted single-speaker consistency essential for voice cloning applications. Recordings captured readings of Sakha literary texts performed by university students under controlled acoustic conditions. Quality requirements mandated minimal background noise, consistent microphone positioning, and precise utterance boundary segmentation. The final synthesis training set comprised approximately 90 minutes of validated single-speaker audio.

Speech Recognition System Architectures

Whisper Architecture and Adaptation

The Whisper system implements sequence-to-sequence speech recognition via transformer encoder-decoder networks trained on massive

multilingual corpora. Acoustic inputs undergo mel-spectrogram feature extraction computing 80 frequency bins over 25-millisecond windows with 10-millisecond frame shifts at 16 kHz sampling rate. The encoder module processes spectrogram representations through 32 stacked transformer layers employing multi-head self-attention mechanisms, producing contextualized frame-level embeddings capturing both local acoustic detail and broader temporal context [12].

The decoder component generates output token sequences autoregressively, attending to encoder representations through cross-attention layers while maintaining causal masking preventing information leakage from future positions. Tokenization employs byte-pair encoding [23] with vocabulary size of 50,000 subword units supporting multilingual text generation. The complete Whisper-large-v3 configuration comprises approximately 1.55 billion trainable parameters distributed across encoder and decoder modules.

Our adaptation experiments examined multiple fine-tuning configurations to identify optimal strategies for Sakha recognition. The primary experimental condition performed comprehensive parameter updates across all model components, while comparison conditions explored partial adaptation strategies including encoder freezing and selective layer training. Training utilized AdamW optimization with learning rate warmup and cosine decay scheduling.

Wav2Vec2-BERT Architecture

The Wav2Vec2-BERT framework combines self-supervised acoustic representation learning with transformer-based sequence modeling (Table 2). Unlike Whisper's spectrogram-based approach, Wav2Vec2 processes raw audio waveforms through convolutional feature extraction layers producing latent representations at 50 Hz frame rate. These representations undergo quantization via learned codebook vectors enabling contrastive pretraining objectives [24].

The transformer encoder processes quantized representations through multiple self-attention layers, learning contextualized embeddings capturing phonetic and prosodic information. During fine-tuning, a linear projection layer maps encoder outputs to character-level probability distributions, with connectionist temporal classification (CTC) loss [25] enabling alignment-free training without requiring frame-level phonetic annotations.

Recognition accuracy benefits substantially from integration with external language models providing linguistic context during decoding. Our experiments employed KenLM toolkit for training n-gram language models on Sakha text corpora, with shallow fusion combining acoustic model posteriors and language model probabilities during beam search decoding.

Table 2

Comparative overview of recognition system architectures

Таблица 2

Сравнительный обзор архитектур систем распознавания

Architecture	Primary Advantages	Notable Limitations
Whisper-large-v3	Extensive multilingual pretraining; robust noise handling; integrated language identification; strong cross-lingual transfer	High computational requirements; substantial fine-tuning needed for rare languages; large model size
Wav2Vec2-BERT	Effective self-supervised learning; efficient acoustic feature extraction; flexible LM integration; moderate resource requirements	Requires substantial unlabeled audio; CTC limitations for long sequences; alignment challenges

Speech Synthesis System Configuration

Text-to-speech synthesis capabilities utilized the XTTS-v2 framework developed by Coqui, designed specifically for cross-lingual voice cloning scenarios with minimal target language training data. The architecture implements neural text-to-spectrogram conversion followed by vocoder-based waveform reconstruction [17], enabling high-fidelity speech generation from textual inputs [26].

Text encoding employs byte-pair tokenization supporting multilingual character inventories. For Sakha adaptation, tokenizer vocabulary required extension incorporating language-specific graphemes absent from default character sets. Custom preprocessing rules handled numeral verbalization, abbreviation expansion, and punctuation normalization according to Sakha orthographic conventions.

The acoustic modeling component predicts mel-spectrogram frames from encoded text representations using attention-based sequence-to-sequence architecture. Speaker conditioning enables voice characteristic control through reference embeddings extracted from short audio samples, supporting zero-shot voice cloning for new speakers not present in training data. The neural vocoder reconstructs time-domain waveforms from generated spectrograms, producing natural-sounding speech output.

Experimental Results and Discussion

Speech Recognition Performance Evaluation

Recognition system evaluation employed word error rate (WER) as the primary accuracy metric, computing minimum edit distance between hypothesized and reference transcriptions normalized by reference word count. This standard metric captures substitution, insertion, and deletion errors providing comprehensive assessment of transcription quality. Experiments systematically varied training configurations to isolate contributions of individual factors.

Table 3

Whisper-large-v3 recognition results across training configurations

Таблица 3

Результаты распознавания модели Whisper-large-v3 при различных конфигурациях обучения

Training Configuration	WER	Loss	Analysis Notes
Complete model fine-tuning	**0.08**	**<0.1**	Optimal configuration
Encoder parameters frozen	**0.42**	**~0.2**	Substantial degradation
Training with noise augmentation	**0.35**	–	Counterproductive effect

Complete model adaptation yielded markedly superior results compared to partial fine-tuning strategies. The comprehensive parameter update condition achieved word error rate of 8%, representing exceptional accuracy for a low-resource language scenario. Training loss converged below 0.1, indicating successful optimization without overfitting concerns.

Complementary experiments with Whisper-small architecture (approximately 244 million parameters) explored computational efficiency trade-offs for resource-constrained deployment scenarios. Training utilized approximately 12.0 hours of combined audio data from Common Voice corpus [21, 27] including original recordings and augmented samples. The optimal configuration achieved 47% WER through comprehensive hyperparameter optimization including learning rate scheduling, gradient accumulation, and curriculum learning strategies. While exhibiting higher error rates than the large model variant, Whisper-small demonstrates viable accuracy

for practical applications requiring reduced computational overhead, with training completable on consumer-grade GPU hardware within reasonable timeframes.

Freezing encoder parameters during fine-tuning – a common strategy for reducing computational requirements and preventing catastrophic forgetting – produced approximately five-fold error rate increase to 42%. This substantial degradation demonstrates that pretrained acoustic representations, while providing valuable initialization, require significant modification for effective Sakha speech processing. The encoder evidently performs critical acoustic normalization and phonetic discrimination functions that cannot be adequately compensated through decoder adaptation alone.

Contrary to initial expectations, noise augmentation during training degraded rather than enhanced recognition performance, producing 35% error rate. Detailed error analysis suggests that artificially introduced acoustic distortions interfered with the model's capacity to learn Sakha-specific phonemic contrasts, particularly affecting discrimination between phonologically similar vowel pairs distinguished primarily by subtle formant frequency differences.

Table 4

Wav2Vec2-BERT recognition results with language model integration

Таблица 4

Результаты распознавания модели Wav2Vec2-BERT с интеграцией языковой модели

System Configuration	WER	Analysis Notes
Acoustic model with KenLM language model fusion	**0.13**	Best configuration
Acoustic model only (no LM integration)	**0.22**	Baseline reference

The self-supervised Wav2Vec2-BERT architecture demonstrated competitive recognition performance, particularly when augmented with statistical language modeling. Integration of n-gram language model probability estimates during beam search decoding provided crucial disambiguation capabilities for morphologically complex word forms, reducing error rate by approximately 40% relative to acoustic-only baseline from 22% to 13% (Table 6).

Systematic error analysis across all recognition systems revealed characteristic failure patterns reflecting Sakha linguistic complexity. Word-

level substitutions predominantly affected morphologically complex forms where suffix boundaries require precise acoustic discrimination. Deletion errors frequently omitted grammatical markers in agglutinative constructions, while insertion errors typically involved spurious word boundaries within long compound formations. These patterns underscore the importance of language model integration for morphologically rich languages and identify priorities for future corpus expansion targeting underrepresented grammatical constructions.

This finding underscores the importance of linguistic context modeling for agglutinative language recognition, where acoustic ambiguity between phonetically similar suffix variants requires resolution through higher-level language structure knowledge. The result suggests promising directions for future work incorporating more sophisticated language models trained on larger Sakha text corpora.

Further investigation explored ensemble strategies combining Wav2Vec2-BERT and Whisper outputs through meta-stacking methodology [28; 29; 30]. This approach employs a logistic regression meta-classifier trained to select the optimal transcription hypothesis for each utterance based on features extracted from both model outputs, including hypothesis length, inter-model Levenshtein distance, and word-level agreement metrics. The ensemble system achieved substantial performance improvements, reducing WER to 27% on the Sakha test set compared to individual model baselines, demonstrating effective complementarity between architectural approaches.

Table 5

Ensemble recognition results via meta-stacking

Таблица 5

Результаты распознавания ансамблевой модели с использованием мета-стэкинга

Model Configuration	WER	Improvement
Wav2Vec2-BERT standalone	0.34	Baseline
Whisper fine-tuned standalone	0.37	Baseline
Meta-stacking ensemble	**0.27**	21% relative reduction

The meta-stacking ensemble demonstrates that combining complementary model architectures through learned arbitration achieves superior recognition accuracy compared to either constituent system operating independently. The logistic regression classifier achieved approximately

85-90% accuracy in selecting the superior hypothesis, effectively leveraging the distinct error patterns exhibited by Wav2Vec2-BERT and Whisper across different acoustic conditions and morphological structures.

Speech Synthesis Quality Assessment

Synthesis system evaluation monitored training loss progression as quantitative indicator of acoustic modeling quality, with lower loss values reflecting improved alignment between generated and target spectral representations. Additionally, perceptual assessment examined naturalness, intelligibility, and phonetic accuracy of synthesized outputs.

Table 6

XTTS-v2 synthesis system training progression

Таблица 6

Процесс обучения системы синтеза XTTS-v2

Data Preparation Stage	Average Loss	Quality Impact
Initial training with unprocessed recordings	**2.67**	Baseline performance
After noise removal and amplitude normalization	**2.59**	Measurable improvement

Data preprocessing yielded measurable quality improvements reflected in approximately 3% reduction in training loss. Systematic noise elimination and amplitude standardization enabled the model to focus learning capacity on linguistically relevant acoustic patterns rather than fitting recording artifacts and volume inconsistencies present in heterogeneous source materials.

Extended training over 22 epochs with optimized hyperparameters (batch size 1 with gradient accumulation of 64, learning rate 5e-7, AdamW optimizer) [31; 32] achieved final average loss of 2.486, representing additional 4% improvement over initial preprocessing optimization. Training was conducted on NVIDIA GeForce RTX 4060 GPU over approximately 19 hours. The trained synthesis system was deployed as a Telegram messaging bot, providing accessible speech generation capabilities for Sakha language users and demonstrating practical applicability of the developed technology.

Perceptual evaluation through informal listening tests confirmed acceptable naturalness and intelligibility for most synthesized utterances. Generated speech exhibited natural prosodic contours with appropriate phrase-level intonation patterns and reasonable speaking rate. Phonetic accuracy

proved particularly strong for Sakha vowel qualities, successfully reproducing front rounded variants absent from the model's primary pretraining languages.

Consonantal accuracy demonstrated greater variability, with occasional substitution errors affecting uvular and glottal segments representing phonemes rare in the multilingual pretraining distribution. Emotional expressiveness remains limited, with generated speech exhibiting relatively neutral affect regardless of textual content. These limitations represent priorities for future development efforts.

Comparative Analysis and Discussion

Cross-architecture comparison reveals complementary strengths across evaluated recognition systems. Whisper achieves superior raw accuracy when computational resources permit comprehensive model adaptation, benefiting from extensive multilingual pretraining providing rich acoustic and linguistic knowledge transferable to Sakha. The 8% word error rate represents exceptional performance for a low-resource language, approaching accuracy levels typically observed only for high-resource languages with abundant training data.

Wav2Vec2-BERT offers more computationally efficient training path with competitive results through language model integration, achieving 13% error rate with substantially lower resource requirements. This architecture may prove preferable for deployment scenarios with constrained computational budgets or when rapid iteration cycles are desired during system development.

Both recognition architectures substantially outperform baseline expectations for low-resource scenarios, validating transfer learning efficacy for Sakha speech processing. The synthesis system demonstrates viable quality despite extremely limited training data, suggesting that neural vocoder approaches effectively leverage cross-lingual acoustic knowledge to produce intelligible speech from minimal language-specific resources.

Conclusions and Future Directions

This investigation systematically evaluated transformer-based neural network approaches for speech recognition and synthesis in the Sakha language, establishing benchmark results and identifying effective adaptation strategies for this low-resource Turkic language. The comprehensive experimental evaluation yields several principal findings with implications for future research and practical system development.

Key conclusions from our experimental work include:

1. Complete fine-tuning of Whisper-large-v3 achieves 8% word error rate, demonstrating that comprehensive encoder adaptation is essential for optimal low-resource recognition performance. Partial fine-tuning strategies substantially degrade accuracy.

2. Wav2Vec2-BERT with statistical language model integration reaches 13% error rate, offering computationally efficient alternative with strong accuracy suitable for resource-constrained deployment scenarios.

3. XTTS-v2 produces intelligible synthesis from minimal training data, validating cross-lingual transfer learning effectiveness for speech generation tasks in underrepresented languages.

− Ensemble meta-stacking combining Wav2Vec2-BERT and Whisper outputs achieves 27% WER, demonstrating 21% relative improvement over individual model baselines through learned hypothesis selection.

− Practical deployment through Telegram bot interface validates real-world applicability of developed synthesis capabilities, providing accessible voice generation for Sakha language community.

4. Data preprocessing quality significantly impacts both recognition and synthesis performance, emphasizing importance of careful corpus curation for low-resource language work.

These results confirm that contemporary neural architectures, when appropriately adapted through transfer learning methodologies, can deliver practical speech technology capabilities for linguistically underserved communities despite severe data limitations. The methodological framework and empirical benchmarks established herein provide foundation for continued Sakha language technology advancement.

Future research directions include corpus expansion through community crowdsourcing initiatives, integration of morphological analysis for improved language modeling capturing Sakha agglutinative structure, development of emotionally expressive synthesis capabilities, and exploration of end-to-end speech translation systems connecting Sakha with major world languages. Collaborative efforts across research institutions and community organizations will accelerate progress toward comprehensive speech technology infrastructure serving Sakha speakers.

References

1. Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*. 2014;(56): 85–100. DOI: https://doi.org/10.1016/j.specom.2013.07.008

2. Joshi P, Santy S, Buber A, Bali K, Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020: 6282–6293.

3. Pakendorf B. *Contact in the prehistory of the Sakha (Yakuts): Linguistic and genetic perspectives*. LOT Publications: Utrecht. 2007.

4. Johanson L, Csató ÉA. *The Turkic Languages*. Routledge Language Family Series. Routledge: London. 2021. DOI: https://doi.org/10.4324/9781003243809

5. Dyachkovsky ND. *Sound structure of the Yakut language. Part 1: Vocalism* (Дьячковский Н.Д. Звуковой строй якутского языка. Вокализм). Yakutsk: Yakutsk publishing house. 1971 (in Russian).

6. Dyachkovsky ND. *Sound structure of the Yakut language. Part 2: Consonantism* (Дьячковский Н.Д. Звуковой строй якутского языка. Консонантизм). Yakutsk: Yakutsk publishing house. 1977 (in Russian).

7. Mussakhojayeva S, Dauletbek K, Yeshpanov R, Varol HA. Multilingual speech recognition for Turkic languages. *Information*. 2023;14(2): 74. DOI: https://doi.org/10.3390/info14020074

8. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989;77(2):257–286. DOI: http://dx.doi.org/10.1109/5.18626

9. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. 2013:6645–6649. DOI: https://doi.org/10.1109/ICASSP.2013.6638947

10. Conneau A, Baevski A, Collobert R, Mohamed A, Auli M. Unsupervised cross-lingual representation learning for speech recognition. *In Proceedings of Interspeech*. 2021:2426–2430. DOI: https://doi.org/10.48550/arXiv.2006.13979

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:5998–6008. DOI: https://doi.org/10.48550/arXiv.1706.03762

12. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. *In Proceedings of International Conference on Machine Learning*. 2023:28492–28518. DOI: https://doi.org/10.48550/arXiv.2212.04356

13. Du W, Maimaitiyiming Y, Nijat M, Li L, Hamdulla A, Wang D. Automatic speech recognition for Uyghur, Kazakh, and Kyrgyz: An overview. *Applied Sciences*. 2023;13(1):326. DOI: https://doi.org/10.3390/app13010326

14. Yeshpanov R, Mussakhojayeva S, Khassanov Y. Multilingual text-to-speech synthesis for Turkic languages using transliteration. *In Proceedings of Interspeech*. 2023:5521–5525. DOI: https://doi.org/10.48550/arXiv.2305.15749

15. Kim J, Kim S, Kong J, Yoon S. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *In Proceedings of the International Conference on Neural Information Processing Systems*. 2020:8067–8077. DOI: https://doi.org/10.48550/arXiv.2005.11129

16. Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *In Proceedings of the International*

Conference on Machine Learning. 2021:5530–5540. DOI: https://doi.org/10.48550/arXiv.2106.06103

17. Kong J, Kim J, Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *In Proceedings of the International Conference on Neural Information Processing Systems*. 2020:17022–17033. DOI: https://doi.org/10.48550/arXiv.2010.05646

18. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R., et al. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. *In Proceedings of IEEE ICASSP*. 2018:4779–4783. DOI: https://doi.org/10.48550/arXiv.1712.05884

19. Ren Y, Hu C, Tan X, Qin T, Zhao S, Zhao Z, Liu T-Y. FastSpeech 2: Fast and high-quality end-to-end text to speech. *In Proceedings of ICLR*. 2021. DOI: https://doi.org/10.48550/arXiv.2006.04558

20. Karibayeva A, Karyukin V, Abduali B, Amirova D. Speech recognition and synthesis models and platforms for the Kazakh language. *Information*. 2025;16(10):879. DOI: https://doi.org/10.3390/info16100879

21. Ardila R, Branson M, Davis K, Kohler M, Meyer J, Henretty M, Morais R, Saunders L, Tyers F, Weber G. Common Voice: A massively-multilingual speech corpus. *In Proceedings of LREC*. 2020:4218–4222. DOI: https://doi.org/10.48550/arXiv.1912.06670

22. Park DS, Chan W, Zhang Y, Chiu C-C, Zoph B, Cubuk ED, Le QV. SpecAugment: A simple data augmentation method for automatic speech recognition. *In Proceedings of Interspeech*. 2019:2613–2617. DOI: https://doi.org/10.21437/Interspeech.2019-2680

23. Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *In Proceedings of EMNLP System Demonstrations*. 2018:66–71. DOI: https://doi.org/10.48550/arXiv.1808.06226

24. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *In Proceedings of the 34th International Conference on Neural Information Processing System*. 2020:12449–12460. DOI: https://doi.org/10.48550/arXiv.2006.11477

25. Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *In Proceedings of ICML*. 2006:369–376. DOI: https://doi.org/10.1145/1143844.1143891

26. Casanova E, Weber J, Shulby C, Junior AC, Gölge E, Müller MA. XTTS: A massively multilingual zero-shot text-to-speech model. *In Proceedings of Interspeech*. 2024. DOI: https://doi.org/10.48550/arXiv.2406.04904

27. Panayotov V, Chen G, Povey D, Khudanpur S. LibriSpeech: An ASR corpus based on public domain audio books. *In Proceedings of IEEE ICASSP*. 2015:5206-5210. DOI: https://doi.org/10.1109/ICASSP.2015.7178964

28. Dyakonov AG. Ensembles in machine learning: Methods and applications. Data Science Course Materials (Дьяконов А.Г. Ансамбли в машинном обучении). 2019. Available at: https://alexanderdyakonov.wordpress.com (accessed 07.09.2025).

29. Wolpert DH. Stacked generalization. Neural Networks 1992;5(2): 241–259. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1

30. Breiman L. Stacked regressions. *Machine Learning*. 1996;24(1):49–64.

31. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al. Transformers: State-of-the-art natural language processing. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020:38–45. DOI: https://doi.org/10.18653/v1/2020.emnlp-demos.6

32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. PyTorch: An imperative style, high-performance deep learning library. *In Proceedings of the International Conference on Neural Information Processing Systems*. 2019:8026–8037. DOI: https://doi.org/10.48550/arXiv.1912.01703

About the authors

Sergei P. STEPANOV – Cand. Sci. (Physics and Mathematics), Head, the Laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, ORCID: https://orcid.org/0000-0001-9445-2726, WoS ResearcherID: F-7549-2017, Scopus Author ID: 56419440700, Elibrary AuthorID: 856700, SPIN 2943-2640, e-mail: sp.stepanov@s-vfu.ru

Dong ZHANG – Cand. Sci. (Physics and Mathematics), Associate Professor, Qufu Normal University, Qufu, Shandong, P.R. China, ORCID: https://orcid.org/0000-0003-4688-7762, WoS ResearcherID: ACW-5232-2022, Scopus Author ID: 57212194896, e-mail: dz_zhangdong@163.com

Timur Z. ZAKHAROV – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: timu.zaxarov40@gmail.com

Altana A. ALEKSEEVA – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: altana.alexeeva@gmail.com

Vladislav L. APROSIMOV – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: malysay88@gmail.com

Djuluur A. FEDOROV – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: fjuluur@mail.ru

Vladimir S. LEVERYEV – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: leverev.vs@svfu.ru

Tuygun A. NOVGORODOV – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: tuygun2000@gmail.com

Ekaterina S. PODOROZHNAYA – Research Assistant, laboratory "Computational Technologies and Artificial Intelligence", Institute of Mathematics and Information Science, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russian Federation, e-mail: ekpodor@gmail.com

Об авторах

СТЕПАНОВ Сергей Павлович – кандидат физико-математических наук, ведущий научный сотрудник, руководитель, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, ORCID: https://orcid.org/0000-0001-9445-2726, WoS ResearcherID: F-7549-2017, Scopus Author ID: 56419440700, Elibrary AuthorID: 856700, SPIN 2943-2640, e-mail: sp.stepanov@s-vfu.ru

ЧЖАН Дун – кандидат физико-математических наук, доцент, Цюйфуский педагогический университет, Цюйфу, Шаньдун, КНР, ORCID: https://orcid.org/0000-0003-4688-7762, WoS ResearcherID: ACW-5232-2022, Scopus Author ID: 57212194896, e-mail: dz_zhangdong@163.com

ЗАХАРОВ Тимур Захарович – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: timu.zaxarov40@gmail.com

АЛЕКСЕЕВА Алтана Александровна – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: altana.alexeeva@gmail.com

АПРОСИМОВ Владислав Леонидович – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: malysay88@gmail.com

ФЕДОРОВ Дьулуур Андрианович – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: fjuluur@mail.ru

ЛЕВЕРЬЕВ Владимир Семенович – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: leverev.vs@svfu.ru

НОВГОРОДОВ Туйгун Александрович – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: tuygun2000@gmail.com

ПОДОРОЖНАЯ Екатерина Сергеевна – лаборант, лаборатория «Вычислительные технологии и искусственный интеллект», Институт математики и информатики, Северо-Восточный федеральный университет им. М.К. Аммосова, Якутск, Российская Федерация, e-mail: ekpodor@gmail.com

Authors' contribution

Sergei P. Stepanov – conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – review & editing, supervision, project administration

Dong Zhang – methodology, validation, investigation, data curation, writing – original draft, writing – review & editing, visualization

Timur Z. Zakharov – formal analysis, investigation

Altana A. Alekseeva – formal analysis, investigation

Vladislav L. Aprosimov – formal analysis, investigation

Djuluur A. Fedorov – formal analysis, investigation

Vladimir S. Leveryev – formal analysis, investigation

Tuygun A. Novgorodov – formal analysis, investigation

Ekaterina S. Podorozhnaya – formal analysis, investigation

Вклад авторов

Степанов С.П. – разработка концепции, методология, программное обеспечение, верификация данных, проведение статистического анализа, проведение исследования, администрирование данных, редактирование рукописи, руководство исследованием, администрирование проекта

Чжан Д. – методология, верификация данных, проведение исследования, администрирование данных, создание черновика рукописи, редактирование рукописи, визуализация

Захаров Т.З. – проведение статистического анализа, проведение исследования

Алексеева А.А. – проведение статистического анализа, проведение исследования

Апросимов В.Л. – проведение статистического анализа, проведение исследования

Федоров Дь.А. – проведение статистического анализа, проведение исследования

Леверьев В.С. – проведение статистического анализа, проведение исследования

Новгородов Т.А. – проведение статистического анализа, проведение исследования

Подорожная Е.С. – проведение статистического анализа, проведение исследования

Conflict of interests

The authors declare no conflict of interests

2025;(4):56–78 Арктика XXI век / Arctic XXI Century

Конфликт интересов

Авторы заявляет об отсутствии конфликта интересов

Поступила в редакцию / Submitted 09.11.2025
Поступила после рецензирования / Revised 08.12.2025
Принята к публикации / Accepted 17.12.2025