BUILDING A THEORETICALLY INFORMED UZBEK-ENGLISH SPEECH RECOGNITION SYSTEM: PRACTICAL IMPLEMENTATION AND EVALUATION

***Oltinbekova Nargiza*[1],**
***Abduvaliyeva Diyora*[2],**
Toshkent Kimyo Xalqaro universiteti Namangan filiali[1,2]

Abstract: *This paper presents the development and evaluation of a theoretically informed Uzbek-English speech recognition system, aimed at enhancing bilingual natural language processing capabilities for Central Asian linguistic environments. The study integrates linguistic theory, acoustic modeling, and neural network-based recognition techniques to improve accuracy and cross-lingual adaptability. Using hybrid deep learning architectures, the system was trained on a mixed corpus of Uzbek and English speech data. Experimental results indicate a significant reduction in word error rate (WER) compared to baseline monolingual models. The paper concludes by discussing the implications for multilingual AI development and future prospects for low-resource language technologies.*

Key words*: Uzbek-English bilingualism, speech recognition, deep learning, phonological theory, code-switching, low-resource languages, acoustic modeling, linguistic-informed neural networks, natural language processing, multilingual AI systems.*

Introduction. Speech recognition technology has become one of the most transformative innovations in artificial intelligence (AI), enabling natural and intuitive human–computer interaction. It allows machines to understand spoken language and respond intelligently, thereby facilitating automation, accessibility, and user-friendly digital systems. While major advancements have been made in speech recognition for high-resource languages such as English, Chinese, and Spanish, low-resource languages like Uzbek remain significantly underrepresented in global AI research. This lack of attention has created a technological gap that limits the development of inclusive and culturally relevant language technologies in Central Asia [1].

The Uzbek language, being agglutinative and characterized by vowel harmony, poses unique challenges for computational processing. These linguistic properties make it difficult for conventional speech recognition systems—originally designed for Indo-European languages—to achieve accurate results. Moreover, many Uzbek speakers frequently mix English words and phrases in their daily communication, a

phenomenon known as code-switching. This linguistic interaction increases the complexity of automatic speech recognition (ASR), as it requires models to handle two phonetic and syntactic systems simultaneously [2].

The purpose of this study is to design and evaluate a theoretically informed Uzbek-English speech recognition system that integrates linguistic theory with modern deep learning approaches. By embedding knowledge from phonology, morphology, and syntax into neural network architectures, the system seeks to improve recognition accuracy and adaptability to bilingual contexts. Unlike purely data-driven methods, this approach combines theoretical understanding of language structure with computational efficiency, ensuring that the model not only learns from data but also respects linguistic constraints.

Developing such a bilingual system also contributes to broader goals of digital inclusion and AI democratization, allowing speakers of underrepresented languages to benefit from modern technology. This research therefore addresses both scientific and social imperatives—bridging linguistic theory with practical AI engineering. The proposed system serves as a pioneering step toward establishing a robust multilingual speech recognition framework that can be extended to other Turkic and low-resource languages in the future [3].

Methods. The methodological framework of this research combines principles of linguistic theory and data-driven deep learning approaches to build a high-performance Uzbek-English bilingual speech recognition system. The design process includes several stages: corpus preparation, feature extraction, model architecture development, training, and evaluation. Each stage is grounded in both theoretical and practical considerations to ensure that the model is linguistically coherent and computationally efficient.

A bilingual speech corpus was developed to train and test the model. The dataset consists of 150 hours of recorded speech, including 60 hours of Uzbek, 80 hours of English, and 10 hours of mixed Uzbek-English code-switched material. The Uzbek data were collected from news broadcasts, radio interviews, podcasts, and conversational speech samples. The English data were adapted from publicly available corpora such as *LibriSpeech* and *TED-LIUM*. For code-switching data, native Uzbek speakers were asked to record spontaneous bilingual dialogues to reflect real-world language use. All audio samples were manually annotated and transcribed following the International Phonetic Alphabet (IPA) standard [4].

Before model training, all recordings were normalized for volume and background noise using *SoX* and *Audacity* tools. Audio signals were converted into Mel-frequency cepstral coefficients (MFCCs), which are widely used to represent speech acoustics. In addition, *spectrogram augmentation* and *phoneme alignment*

were applied to improve the robustness of the model under various acoustic conditions. The linguistic structure of Uzbek, especially vowel harmony and morphological suffixes, was encoded using special phoneme embeddings [5].

The architecture of the proposed system integrates a **Hybrid CTC-Attention Model** with a **Bi-directional Long Short-Term Memory (BiLSTM)** network and a **Convolutional Neural Network (CNN)** frontend. The CNN layers extract local acoustic features, while BiLSTM layers capture long-range temporal dependencies in speech. The attention mechanism aligns input and output sequences, allowing the model to focus dynamically on relevant phonetic segments. Furthermore, linguistic rules were embedded as constraints within the decoding stage, ensuring that predicted words follow the phonological and morphological norms of Uzbek and English.

Model training was performed using the *PyTorch* deep learning framework on NVIDIA GPUs. The Adam optimizer was used with an adaptive learning rate scheduler. To prevent overfitting, *SpecAugment* and *dropout* regularization techniques were employed. The model was trained for 50 epochs, and early stopping was applied when validation accuracy plateaued. Evaluation was based on *Word Error Rate (WER)* and *Character Error Rate (CER)* metrics across three datasets: Uzbek-only, English-only, and mixed bilingual data [6].

Performance evaluation followed a cross-validation approach, where 80% of the corpus was used for training and 20% for testing. Statistical analysis was applied to compare the results of the proposed theoretically informed model with conventional monolingual and bilingual baselines. Qualitative evaluation was also conducted through error analysis, focusing on how the system handled Uzbek vowel harmony, English stress patterns, and code-switched expressions.

Results and Discussion. The results of the proposed theoretically informed Uzbek-English speech recognition system demonstrate significant improvements over baseline models in terms of accuracy, linguistic adaptability, and cross-lingual performance. The system was evaluated using three datasets: monolingual Uzbek, monolingual English, and mixed Uzbek-English code-switched speech. Quantitative and qualitative analyses were conducted to assess the efficiency of the model in recognizing diverse phonetic structures and linguistic patterns [7].

The overall performance was measured using *Word Error Rate (WER)* and *Character Error Rate (CER)*. The results indicate that the proposed model achieved substantial accuracy improvements across all datasets.

Model Type	WER (Uzbek)	WER (English)	WER (Code-switched)	CER (Average)
Monolingual (Uzbek-only)	18.6%	–	–	11.4%

Model Type	WER (Uzbek)	WER (English)	WER (Code-switched)	CER (Average)
Monolingual (English-only)	–	9.4%	–	6.3%
Baseline Bilingual	15.2%	10.8%	17.9%	8.9%
Proposed Theoretically Informed Model	**11.3%**	**8.9%**	**12.1%**	**5.7%**

These results show an average **27% relative reduction in WER** compared to traditional bilingual systems. The improvement was most pronounced in Uzbek recognition, demonstrating the benefit of incorporating phonological and morphological knowledge into the model. The lower CER also indicates that the system handles phoneme-level decoding more effectively, especially in acoustically complex or code-switched sentences [8].

Beyond numerical performance, qualitative evaluation revealed that the system could accurately recognize complex Uzbek word forms and English insertions within the same utterance. For example, it successfully processed sentences such as *"Men online meetingda qatnashdim"* ("I participated in an online meeting") without misinterpreting code-switched words. The model also captured vowel harmony and suffix variations, which are critical features of Uzbek morphology.

Moreover, attention-based decoding enhanced recognition of context-dependent pronunciation, reducing errors caused by coarticulation or regional accents. Compared to the baseline model, the theoretically informed system demonstrated better stability when speakers used informal or spontaneous speech [9].

The integration of linguistic theory into neural modeling proved crucial for improving recognition accuracy in a low-resource bilingual setting. Traditional end-to-end models rely solely on data-driven learning, which often fails to generalize well for morphologically rich languages. By contrast, the inclusion of **phoneme embeddings, morphological rules, and linguistic constraints** enabled the system to develop a deeper structural understanding of both Uzbek and English.

The success of this approach supports the idea that artificial intelligence systems benefit from *hybrid intelligence*—the combination of human linguistic knowledge and machine learning. This synergy not only enhances recognition precision but also preserves the linguistic authenticity of underrepresented languages.

The findings have broader implications for multilingual AI development, particularly in regions where multiple languages coexist. The proposed system can be extended to other Turkic languages with similar phonological properties. Furthermore, its architecture could support real-time applications such as bilingual

virtual assistants, speech-to-text transcription tools, and educational software [10]. The study contributes to the advancement of *ethical and inclusive AI*, ensuring that speakers of lesser-studied languages can access cutting-edge technology.

Conclusion. This study set out to design, implement, and evaluate a theoretically informed Uzbek-English speech recognition system that bridges the gap between linguistic theory and modern deep learning technologies. The research demonstrates that incorporating phonological, morphological, and syntactic insights into neural architectures can significantly enhance recognition accuracy for low-resource and bilingual languages. Unlike traditional purely data-driven models, the proposed system leverages linguistic constraints to ensure that the model respects the structural properties of both Uzbek and English.

Experimental results clearly show that the theoretically informed model achieved higher performance compared to baseline systems, reducing the word error rate by more than 25% and the character error rate by over 30%. The model's ability to process code-switched speech effectively—particularly sentences mixing Uzbek morphology and English vocabulary—marks a substantial step toward practical bilingual speech recognition in Central Asian contexts. The inclusion of a bilingual lexicon and phoneme-based embeddings proved essential for capturing the unique acoustic-phonetic characteristics of Uzbek while maintaining the robustness of English recognition.

From a theoretical perspective, the study confirms that linguistic principles such as vowel harmony, agglutinative morphology, and stress patterns can be successfully integrated into neural network training through embedding design and decoding constraints. This hybrid approach not only improves model accuracy but also contributes to preserving linguistic identity in digital technologies. The findings thus advocate for an interdisciplinary approach in artificial intelligence—where language theory and computational modeling coexist and complement each other.

Practically, this research provides a foundation for developing multilingual AI systems tailored to the linguistic realities of Central Asia. The system can be extended to other Turkic or regional languages with similar morphological features, promoting greater inclusion of underrepresented languages in the global AI landscape. Future research should focus on expanding the code-switched corpus, exploring transformer-based architectures such as Whisper and Conformer, and integrating real-time speech-to-text functionalities for educational, governmental, and commercial applications.

In conclusion, the study illustrates that theory-guided design is not only compatible with modern AI but also essential for its advancement. By grounding neural systems in linguistic understanding, we can create speech recognition

technologies that are more accurate, culturally aware, and inclusive for all language communities.

References.

1. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), 602–610.
2. Chan, W., et al. (2016). Listen, Attend and Spell: A neural network for large vocabulary conversational speech recognition. *ICASSP 2016*.
3. Pratap, V., et al. (2020). Wav2letter++: The fastest open-source speech recognition system. *IEEE Transactions on Audio, Speech, and Language Processing*, 28, 1277–1289.
4. Bozkurt, F., et al. (2017). Speech synthesis for Turkic languages using linguistic and phonetic constraints. *Computer Speech & Language*, 46, 87–104.
5. Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
6. Zhou, J., & Zhao, W. (2021). Code-switching speech recognition: Challenges and solutions. *Frontiers in Artificial Intelligence*, 4(39), 1–12.
7. Hinton, G., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82–97.
8. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Pearson.
9. Muminov, S., & Rakhimov, A. (2021). Uzbek Speech Corpus Development for Automatic Speech Recognition. *Journal of Language Technologies*, 5(2), 45–59.
10. Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS 2017*, 5998–6008.