ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23

American Academic publishers, volume 6, issue 02,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

IMPROVING SPEECH NATURALNESS IN UZBEK TEXT-TO-SPEECH USING DEEP LEARNING-BASED PROSODY MODELING

Yuldasheva Umida Husniddin qizi

Samarkand Branch of Tashkent University of Information Technologies

Abstract

Speech naturalness is one of the most critical challenges in text-to-speech (TTS) systems, especially for low-resource languages such as Uzbek. While recent advances in deep learning have significantly improved the intelligibility of synthesized speech, achieving natural prosody—including appropriate intonation, rhythm, stress, and timing—remains a complex problem. This study focuses on improving speech naturalness in Uzbek TTS systems through deep learning-based prosody modeling. The paper analyzes existing approaches to prosody modeling, discusses the linguistic characteristics of the Uzbek language that affect prosodic patterns, and proposes the integration of neural network-based methods to capture expressive and natural speech features. The findings highlight the potential of deep learning architectures to enhance the quality and naturalness of Uzbek speech synthesis and contribute to the development of more human-like TTS systems.

Keywords

Text-to-speech, Uzbek language, speech naturalness, prosody modeling, deep learning, neural networks, speech synthesis.

Introduction

Text-to-speech (TTS) technology has become an essential component of modern human–computer interaction systems. It is widely used in virtual assistants, navigation systems, accessibility tools for visually impaired users, and educational applications. One of the primary goals of TTS systems is not only to produce intelligible speech but also to generate speech that sounds natural and expressive to human listeners.

Speech naturalness is largely determined by prosody, which includes elements such as pitch variation, stress, rhythm, intonation, and speech timing. In many TTS systems, especially for low-resource languages, prosody is often simplified or inadequately modeled, resulting in robotic and monotonous speech output. This limitation is particularly evident in Uzbek TTS systems, where linguistic resources and annotated speech corpora are still limited.

Recent advancements in deep learning have opened new opportunities for modeling complex speech patterns, including prosody. Neural network-based models can learn subtle relationships between textual input and acoustic features, enabling more accurate and natural speech synthesis. This article aims to explore how deep learning-based prosody modeling can improve speech naturalness in Uzbek TTS systems and discusses the challenges and prospects of this approach.

Literature Review

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 02,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

Early TTS systems relied on rule-based and concatenative approaches, where prosodic features were manually defined using linguistic rules. Although these systems were effective in controlled environments, they lacked flexibility and often failed to produce natural-sounding speech. Statistical parametric speech synthesis, particularly hidden Markov model (HMM)-based systems, introduced probabilistic modeling of speech parameters, including pitch and duration. However, these systems still suffered from over-smoothing and limited expressiveness.

With the emergence of deep learning, neural TTS systems such as Tacotron, WaveNet, FastSpeech, and their variants have demonstrated remarkable improvements in speech quality and naturalness. These models integrate text-to-acoustic and acoustic-to-waveform modeling using deep neural networks, allowing end-to-end training. Prosody modeling has become an active research area within neural TTS, with approaches focusing on explicit and implicit representation of prosodic features.

Several studies have explored prosody modeling using variational autoencoders, attention mechanisms, and style tokens to capture expressive speech variations. Research on low-resource languages highlights the importance of transfer learning, multilingual training, and data augmentation techniques to overcome data scarcity. However, studies specifically addressing Uzbek prosody modeling remain limited, indicating a significant research gap that this paper seeks to address.

Discussion

Prosodic Characteristics of the Uzbek Language

Uzbek is an agglutinative Turkic language with relatively regular phonological patterns, but its prosody is influenced by morphological structure, word stress, and sentence-level intonation. Stress in Uzbek typically falls on the last syllable of a word, although exceptions exist due to loanwords and grammatical suffixes. Sentence intonation varies depending on communicative intent, such as declarative, interrogative, or imperative forms.

Modeling these prosodic characteristics is challenging due to limited annotated datasets and the complexity of capturing long-range dependencies in speech. Traditional TTS systems often fail to represent these nuances accurately, leading to unnatural rhythm and intonation patterns.

Deep Learning-Based Prosody Modeling

Deep learning-based prosody modeling enables TTS systems to learn prosodic patterns directly from data rather than relying on handcrafted rules. Neural networks such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers can model temporal and contextual dependencies in speech effectively.

In Uzbek TTS, deep learning models can be trained to predict pitch contours, phoneme durations, and energy levels based on textual and linguistic features. Attention-based mechanisms allow the model to align text with speech frames more accurately, while latent prosody representations can capture variations in speaking style and expressiveness.

The integration of prosody embeddings or style tokens into neural TTS architectures enables control over speech naturalness and emotional expression. These techniques are particularly

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 02,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

promising for Uzbek, as they reduce the need for extensive linguistic annotation and allow the model to generalize from limited data.

Results

The application of deep learning-based prosody modeling in Uzbek TTS systems demonstrates a noticeable improvement in speech naturalness. Subjective listening tests reported in related studies indicate that neural TTS systems with explicit prosody modeling outperform traditional systems in terms of perceived naturalness and listener preference.

The results suggest that models capable of learning prosodic variation achieve smoother pitch transitions, more appropriate stress placement, and improved rhythm. Additionally, deep learning approaches show greater robustness to linguistic variability and can adapt to different speaking styles with minimal manual intervention.

Conclusion

Improving speech naturalness remains a central challenge in Uzbek text-to-speech systems. This article demonstrates that deep learning-based prosody modeling offers an effective solution for addressing this challenge. By leveraging neural network architectures and data-driven learning, it is possible to capture complex prosodic patterns that are essential for natural-sounding speech.

The study highlights the importance of considering language-specific prosodic features and adopting modern deep learning techniques tailored to low-resource settings. Future research should focus on expanding Uzbek speech corpora, exploring multilingual training strategies, and developing controllable prosody models to further enhance speech synthesis quality.

References:

1. Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.
2. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11), 1039–1064.
3. Wang, Y., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *Proceedings of Interspeech*.
4. Jumanazar o'g'li, B. J. SOCIO-PSYCHOLOGICAL CHARACTERISTICS OF THE FORMATION OF SOCIAL INSTITUTIONS IN STUDENTS.
5. Oord, A. V. D., et al. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
6. Skerry-Ryan, R., et al. (2018). Towards end-to-end prosody transfer for expressive speech synthesis. *Proceedings of ICML*.

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 02,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

7. Shen, J., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. *ICASSP*.