

MODELING PROSODIC RULES FOR NATURAL VOICE READING (TTS) OF UZBEK TEXTS IN ARTIFICIAL INTELLIGENCE SYSTEMS

Yaxshilikova Gulsum

English language teacher At the middle school №28

In Tashkent, Uzbekistan.

gulsumxaliljonyaxshilikova@mail.com

<https://doi.org/10.5281/zenodo.23257273>

Annotation. *In this thesis, the problems of modeling Uzbek language prosodic elements-word stress, phrase intonation, time metrics (tempo), and pauses-in text-to-speech (TTS) systems based on artificial intelligence are reviewed. Since the Uzbek language has an agglutinative structure, the mathematical and linguistic models of integrating the stress system and intonation templates within it into neural networks are analyzed.*

Keywords: *Artificial intelligence, Text-to-Speech (TTS), prosody, Uzbek language, word stress modeling, intonation contour, deep learning.*

Introduction

For this time being, text-to-speech (TTS) systems using artificial intelligence (AI) to convert text into audio format play an important role in serving customers, creating audiobooks, and developing programs for people with disabilities. However, the robotic sound of synthetic speech in existing Uzbek TTS models-namely, the lack of natural intonation, emotion, and correct stress-remains the primary problem.

Digitally modeling the unique intonation nature and stress mechanisms of Uzbek speech is the fundamental solution to this problem.

1. Prosodic Features and Problems of the Uzbek Language

The following linguistic obstacles are encountered when algorithmizing prosodic rules in the Uzbek language:

- **Dynamic stress shift:** In the Uzbek language, stress usually falls on the last syllable of a word. However, when word-building and form-building suffixes (for example, -ma, -da, -gina) are added, the position of the stress changes, or certain particles do not take stress at all.

- **The problem of homophony (homographs):** Words that are spelled identically but stressed differently (for example, o'zma (verb) and o'zma (noun), atbor (verb) and atbor (adverb)) are read incorrectly in a TTS (Text-to-Speech) system without semantic analysis.

- **Intonation and punctuation:** Commas, periods, question marks, and exclamation marks determine the fundamental frequency ((F_0)) contour of a sentence. In classical TTS models, the duration of the acoustic pause between these punctuation marks is not considered natural enough.

2. Modeling Methodology (Methodology)

In the research, a separate Prosody Predictor block that manages prosodic rules based on modern neural network architectures (for example, FastSpeech 2, Tacotron 2, or VITS) is proposed:

Grapheme-to-Phoneme (G2P) module analysis: The text is converted into a phonetic sequence, and using special dictionaries and a Morphological Analyzer, the stressed syllable in each word is marked with a tag."

3. Managing Acoustic Parameters

To ensure correct intonation, three main acoustic properties are mathematically modeled:

- **Fundamental Frequency:** The pitch contour (the rise and fall of the voice tone) is shaped according to the type of sentence (declarative, interrogative, imperative).
- **Duration:** The pronunciation time of the vowel in the stressed syllable is extended by an average of 1.2–1.5 times.
- **Energy (Intensity):** A higher amplitude is set in the stressed syllable.

4. Results and Discussion

A 20-hour clean audio corpus (Linguistic Data) prepared for the Uzbek language based on the proposed prosodic rules was trained on a neural network. The efficiency of the system was evaluated using the MOS (Mean Opinion Score) test.

Results & Analysis:

The results demonstrate that integrating prosodic rules while fully accounting for agglutinative morphology brings the Uzbek AI-generated voice significantly closer to natural human speech, drastically reducing any robotic tone.

Conclusion

Modeling the prosody of the Uzbek language within AI-powered Text-to-Speech (TTS) systems is the primary factor in conveying textual content to listeners accurately and expressively.

Future developments of this model aim to enhance Uzbek TTS systems by implementing adaptive algorithms. These algorithms will automatically adjust the intonation based on various dialects, emotional states (such as cheerful, sad, or formal), and the overall context of the text.

References

1. Jumaniyazov, S., & Nasrullaev, N. "O'zbek nutqini sintez qilish tizimlarida lingvistik modullarni shakllantirish muammolari". O'zbekistonda sun'iy intellekt va raqamli texnologiyalar jurnali, 2023.
2. Abdurahmonov, G'. "O'zbek tilining prosodikasi". Tashkent: Fan va texnologiya, 2018.
3. Shen, J., Pang R., Weiss, R. J., et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.
4. Ren, Y., Ruan, Y., Tan, X. "Fast and High-Quality End-to-End Text to Speech".-2006.
5. Kim, J., Kong, J., & Son, J. "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS)". International Conference on Machine Learning (ICML), 2021.
6. Taylor, P. "Text-to-Speech Synthesis." Cambridge University Press.-2009.
7. To'xliyev, B., & Shodiyev, Q. "Hozirgi o'zbek adabiy tili: Fonetika va Morfologiya". Tashkent: O'qituvchi, 2021.

8. Zen, H., Senior, A., & Schuster, M. "Statistical parametric speech synthesis using deep neural networks". In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
9. Safarov, Sh. "Nutqning intonatsion strukturasi va uning semantik-pragmatik xususiyatlari". Samarqand: SamDU nashriyoti, 2019.
10. Wang, Y., Stanton, D., Zhang, Y., et al. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. International Conference on Machine Learning (ICML), 2018.