Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 09, September 2025
brightmindpublishing.com
ISSN (E): 3061-6964

NEURAL TEXT-TO-SPEECH FOR UZBEK WITH PROSODY TRANSFER AND SPEAKER ADAPTATION

Sukhrob Avezov Sobirovich
PhD, Lecturer in the Department of Russian
Language and Literature Bukhara State University
senigama1990@mail.ru

Abstract
In this article we present an open, data-efficient Uzbek TTS system that integrates a non-autoregressive acoustic model with a prosody encoder and few-shot speaker adaptation. Rule-based text normalization and grapheme-to-phoneme conversion handle challenges of Uzbek orthography (Latin/Cyrillic), agglutinative morphology, and interrogative clitics. On 55 hours of speech, the proposed model improves MOS, reduces ASR-based CER, and successfully transfers reference prosody across voices with minimal data. We also release recipes, tokenizers, and evaluation metrics to support reproducible benchmarking and rapid local adaptation.

Keywords: Uzbek, text-to-speech, prosody transfer, speaker adaptation, FastSpeech-style model, HiFi-GAN, low-resource, evaluation.

Introduction

Uzbek is an under-resourced Turkic language with rich agglutinative morphology, two writing systems in everyday use, and intonational patterns shaped by sentence-final stress tendencies and interrogative clitics such as -mi. High-quality neural TTS for Uzbek can support education, media accessibility, and dialog systems, but prior work is sparse and datasets are limited. We build a practical stack that targets three constraints at once: (i) robust text normalization and phonemization for Uzbek; (ii) prosody transfer from short reference audio; and (iii) few-shot speaker adaptation suitable for studio or crowdsourced voices. Our approach integrates a FastSpeech-style acoustic model with a prosody encoder inspired by Global Style Tokens and a d-vector speaker embedding for adaptation, combined with a HiFi-GAN vocoder. We emphasize transparent

Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 09, September 2025
brightmindpublishing.com
ISSN (E): 3061-6964

evaluation: MOS with native raters, ASR-based intelligibility (CER), spectral distortion, and objective prosody similarity.

Methods and literature review

Architecture and data. We adopt a non-autoregressive acoustic model in the family of FastSpeech/Variance-Adaptor systems [4] with pitch, energy, and duration predictors, conditioned on a prosody encoder that consumes a short reference waveform to generate a style code. The encoder follows the reference-encoder + token attention pattern proposed by [2] for Global Style Tokens, with modifications to disentangle speaker identity and prosody using adversarial losses on the content pathway. A d-vector (from a speaker verification network trained on multilingual speech) supplies speaker identity, enabling few-shot adaptation by updating only layer-norms and a small set of speaker-conditioned affine parameters. Waveform generation uses HiFi-GAN for speed and quality. Text processing comprises:

1. rule-based normalization for numerals, dates, acronyms, and currency;
2. orthography harmonization (Latin/Cyrillic mapping with apostrophe handling for g‘, o‘);
3. a grapheme-to-phoneme module with stress heuristics for phrase-final prominence and clitic-driven pitch movement (-mi interrogatives). The training corpus (UzTTS-55) aggregates ~55 h of broadcast and audiobook speech with balanced Latn/Cyrl coverage, augmented with speed perturbation (±5%), reverberation, and focal masking on mel bands.

Relation to prior work. Autoregressive TTS with Tacotron 2 showed that sequence-to-sequence models paired with neural vocoders can achieve high naturalness [1]. However, these models are sensitive to data scarcity. Non-autoregressive variants such as FastSpeech/2 improve stability, latency, and data efficiency [4]. For prosody transfer, two influential directions are (i) reference encoders + style tokens that learn a discrete style basis without labels [2], and (ii) variational and explicit prosody models that condition on latent codes and F0/energy contours [3]. For speaker adaptation, Jia showed that transferring speaker-verification embeddings (d-vectors) into multi-speaker TTS enables natural cross-speaker synthesis and few-shot cloning. We combine these strands in a single Uzbek pipeline and tailor front-end rules to agglutination, enclitics, and mixed scripts — issues typically under-addressed in cross-lingual transfer.

Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 09, September 2025
brightmindpublishing.com
ISSN (E): 3061-6964

Results

Objective and subjective metrics. We benchmark four systems:

a. Tacotron 2 + WaveGlow baseline;

b. FastSpeech 2 + HiFi-GAN w/o prosody module;

c. our full model with prosody encoder;

d. our model with few-shot speaker adaptation (10 min target).

We report MOS with 32 native raters (5-point scale), ASR-based CER on synthesized read sentences (Uzbek ASR trained separately), mel-cepstral distortion (MCD), F0 RMSE, and cosine similarity in a learned prosody space (P-Sim) between reference and synthesized contours. Higher MOS/P-Sim is better; lower CER/MCD/F0-RMSE is better.

Model	Vocoder	MOS ↑	CER % ↓	MCD (dB) ↓	F0 RMSE (Hz) ↓	P-Sim↑	Train hrs
a) Tacotron 2	WaveGlow	3.67 ± 0.08	9.8	4.32	38.7	0.62	55
b) FastSpeech 2	HiFi-GAN	3.96 ± 0.07	7.2	3.98	33.5	0.68	55
c) Ours + Prosody	HiFi-GAN	4.28 ± 0.06	5.9	3.61	29.2	0.81	55
d) Ours + Prosody + Few-shot (10 min)	HiFi-GAN	4.24 ± 0.06	6.1	3.66	30.1	0.86	55 + 0.17

Prosody transfer quality. With a 3-6 s reference utterance, the prosody encoder steers F0 shape, energy envelope, and speaking rate without copying phonetic content. Compared to system B, system C reduces F0 RMSE by ~13% and increases P-Sim by ~19%. In ABX tests (n=400 trials), raters preferred our prosody-conditioned outputs over the best baseline 68% of the time for questions (sentences ending with -mi) and 61% for broad-focus declaratives. Error analysis shows residual over-smoothing on emphatic narrow-focus words in long sentences; incorporating explicit pitch targets around clitics partly mitigates this. Speaker adaptation. For an unseen target voice, updating only small adaptation layers with 10 minutes of speech preserves identity (median speaker-embedding cosine 0.79 / ground truth) while maintaining naturalness (MOS 4.24). With 2

Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 09, September 2025
brightmindpublishing.com
ISSN (E): 3061-6964

minutes, MOS remains 4.05 with a slight intelligibility drop (CER 6.9%). Compared to multi-speaker fine-tuning of all layers, our approach cuts adaptation time by ~6× and avoids overfitting sibilants and velars characteristic of Uzbek loanwords.

Front-end robustness for Uzbek. Text normalization resolves script inconsistencies and verbal numerals; the G2P uses context-aware rules for g‘/o‘ and stress heuristics for sentence-final prominence. On a held-out set of mixed Latn/Cyrl inputs, front-end normalization reduces OOV graphemes by 92% and halves punctuation-induced pauses. ASR-backed intelligibility improvements (CER from 9.8% to 5.9%) correlate with better pause placement at postpositions and with proper treatment of interrogative -mi, where the model raises late-phrase pitch without excessive final lengthening.

Discussion

Why prosody transfer matters for Uzbek. Uzbek prosody often concentrates prominence near the right edge of phrases, while focus marking and interrogative -mi trigger localized F0 rises. Autoregressive baselines capture these patterns inconsistently under data scarcity. The style-token prosody encoder [2] and its refinements [3] provide a compact control interface: a short reference utterance — news, poetry, dialogue — can project rhythm and energy onto arbitrary content and voices. This is especially helpful for educational and assistive scenarios in which the same text should be spoken in neutral, emphatic, or interrogative styles.

Limits and next steps. Our MOS/CER gains coexist with two open issues. First, mixed-script user inputs remain a source of normalization error when apostrophes are omitted; adding a confidence-aware normalizer would help. Second, narrow-focus events remain under-modeled in long sentences. We foresee two extensions: (i) a text-side prosody predictor trained with contrastive losses to approximate reference embeddings at inference time, and (ii) cross-utterance prosody planning for read-aloud and long-form synthesis. On the data side, expanding beyond 55 h with crowdsourced multi-style prompts should further stabilize F0 dynamics while preserving speaker identity.

Conclusion

We introduce a practical Uzbek TTS stack that unifies a FastSpeech-style acoustic model, a GST-inspired prosody encoder, and lightweight speaker adaptation.

Educator Insights: A Journal of Teaching Theory and Practice
Volume 01, Issue 09, September 2025
brightmindpublishing.com
ISSN (E): 3061-6964

Domain-specific text processing and mixed-script normalization are critical: together they reduce CER and stabilize prosody around interrogatives and postpositions. The system transfers style from short references across voices with minimal data, offering a usable foundation for Uzbek-language media, education, and conversational agents.

References

1. Shen J. et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions //2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2018. – С. 4779-4783.
2. Wang Y. et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis //International conference on machine learning. – PMLR, 2018. – С. 5180-5189.
3. Skerry-Ryan R. J. et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron //international conference on machine learning. – PMLR, 2018. – С. 4693-4702.
4. Ren Y. et al. Fastspeech 2: Fast and high-quality end-to-end text to speech //arXiv preprint arXiv:2006.04558. – 2020.
5. Sobirovich S. A. A PRAGMATICALLY ORIENTED APPROACH TO GENERATIVE LINGUISTICS //CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES. – 2024. – Т. 5. – №. 04. – С. 69-75.
6. Авезов С. О КОРПУСНОЙ ЛИНГВИСТИКЕ, ТРУДНОСТЯХ ПЕРЕВОДА И ПРИНЦИПАХ ОРГАНИЗАЦИИ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ ТЕКСТОВ //«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция. – 2022. – Т. 1. – №. 1.