Volume 3, Issue 9, September – 2025 **ISSN (E):** 2938-379X

END-TO-END UZBEK-RUSSIAN SPEECH TRANSLATION WITH SELF-SUPERVISED PRETRAINING

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

Abstract
In this article we study end-to-end Uzbek→Russian speech translation under realistic low-resource and code-switching conditions. We couple a wav2vec-style encoder pre-trained on unlabeled audio with a Transformer decoder, add multi-task ASR/CTC objectives, and distill from a strong cascade teacher. Script-aware tokenization and data augmentation reduce sparsity. On conversational and broadcast tests the model improves BLEU/chrF at fixed latency and yields fewer morphology and NE errors.

Keywords: End-to-end speech translation, Uzbek-Russian, self-supervised pretraining, wav2vec 2.0, XLS-R, knowledge distillation, code-switching, low-resource.

Introduction

Building speech-to-text translation (ST) for Uzbek→Russian is challenging: labeled paired speech–translation data are scarce; Uzbek is agglutinative with productive suffixation; communication frequently code-switches across Uzbek (Latn/Cyrl) and Russian (Cyrl); and practical systems must operate with limited compute. While cascaded ASR→MT pipelines are robust, their latency and error compounding motivate direct, end-to-end ST. We investigate whether self-supervised pretraining on unlabeled Uzbek/Russian audio [1], [2] combined with multi-task learning and teacher-student distillation can close the data gap. We design an encoder-decoder architecture with wav2vec-style representations, add CTC and ASR decoders during training, and use script-aware subwords to handle mixed alphabets. Evaluated on a new mixed-domain benchmark, our approach outperforms a tuned cascade in BLEU and chrF, with larger gains on conversational code-switching and named entities.

Methods and related work

Self-supervised speech pretraining has transformed low-resource ASR by learning general acoustic features from raw audio without transcripts [1]. wav2vec 2.0 [1] learns contextualized units through contrastive masking; XLS-R extends this to multilingual, cross-domain settings [2]. In ST, early end-to-end work [3] showed feasibility but required large paired corpora or strong regularization. Subsequent research explored multi-task learning with CTC/ASR auxiliaries, SpecAugment, speech encoders initialized from self-supervised checkpoints, and

knowledge distillation from cascades to stabilize training and reduce exposure bias. Whisper [4] popularized large-scale weakly supervised pretraining, providing strong ASR teachers but not direct low-resource ST.

Our system follows this line. *Architecture.* A 24-block Conformer-wav2vec encoder (initialized from XLS-R-0.3B) feeds a 6-layer Transformer decoder (512-d, 8 heads) that predicts Russian subwords. A CTC head and a lightweight ASR decoder share the encoder. During training we sample tasks: ST (70%), ASR (20%), CTC-only (10%). *Text side.* We build a 32k SentencePiece vocabulary over Russian (Cyrl) and Uzbek (Latn/Cyrl) after applying a reversible normalization that preserves script boundaries via inline tags («UZL», «UZC», «RUC»). *Data.* We compile ~480 h of unlabeled Uzbek/Russian speech for pretraining and ~210 h of paired Uzbek speech with Russian translations for ST fine-tuning (conversational calls, lectures, news, voice notes). We expand coverage with (i) speed perturbation and SpecAugment; (ii) back-translation from 1.1 M Russian sentences using a tagged TTS pipeline (20 h synthetic speech); (iii) code-switch injection by mixing Russian loanwords and Uzbek enclitics. *Distillation.* The teacher is a cascade: a fine-tuned Whisper-small ASR (Uzbek+RU) plus a Transformer MT trained on 6.4 M Uzbek↔Russian text pairs. We minimize a mixture of label-smoothed cross-entropy to references and KLD to the teacher's sequence-level distributions. *Decoding.* Beam = 5 with length-penalty 0.6; nucleus-guided re-ranking by ASR confidence when teacher scores are available. *Evaluation.* We report BLEU (sacreBLEU), chrF, TER, and entity F1 on two held-out sets: CONV (code-switching, 8.5 h) and BCAST (news/lectures, 9.2 h). Latency is estimated via Average Proportion under chunked streaming input (320 ms frames) for a deploy-oriented analysis.

Results

Overall accuracy. Table 1 summarizes main results. The end-to-end model with XLS-R initialization and distillation (E2E-XLSR+KD) surpasses the tuned cascade on BLEU (+2.5 CONV, +1.2 BCAST) and chrF, with lower TER. The gains are larger where code-switching, hesitations, and colloquialisms dominate (CONV). Removing either pretraining or distillation degrades performance, indicating complementary benefits: pretraining stabilizes acoustic modeling; distillation transfers teacher phrasing and punctuation.

Error profile. On manual audits (600 sentences), the cascade often preserves source word order verbatim, propagating ASR deletion errors into mistranslations of Uzbek case endings (-ga/-ni/-dan). The end-to-end system better resolves Uzbek agglutinative morphology into correct Russian case and preposition choices, especially for indirect objects and ablatives, and improves transliteration of toponyms. Named-entity F1 rises from 81.2 → 86.5 on CONV. Typical residual errors involve numerals (spoken Uzbek forms "o'n besh" / Russian inflection) and rare honorifics.

Latency analysis. In chunked streaming, E2E-XLSR+KD keeps Average Proportion similar to the cascade while producing fewer long pauses. Because decoding conditions directly on acoustics, the model commits earlier on local phrases and hedges less on filler words, which yields more stable partials in incremental UIs.

Ablations. Removing script-aware tags reduces chrF on CONV by 0.9 due to tokenization mismatches at switch points. Dropping the auxiliary ASR decoder (keeping only CTC) slightly harms punctuation and clause boundary placement (BLEU −0.4). Synthetic speech adds modest but consistent gains on BCAST, with diminishing returns beyond ~20 h.

Table 1 — Uzbek→Russian speech translation (test sets CONV/BCAST). Higher is better (↑), lower is better (↓).

System	Init	Distill	CONV BLEU ↑	CONV chrF ↑	CONV TER ↓	BCAST BLEU ↑	BCAST chrF ↑	BCAST TER ↓	Latency AP ↓
Cascade (Whisper-ASR → MT)	—	—	24.6±0.2	52.3	57.8	28.9±0.1	56.4	51.1	0.56
E2E-scratch	random	—	18.7±0.4	47.1	64.9	22.3±0.3	51.2	58.3	0.52
E2E-XLSR	XLS-R	—	26.1±0.3	54.1	55.2	29.6±0.2	57.3	50.2	0.53
E2E-XLSR+KD (ours)	XLS-R	✓	27.1±0.2	55.0	53.9	30.1±0.2	57.9	49.6	0.52
– no script tags	XLS-R	✓	26.2	54.1	55.0	29.8	57.5	50.1	0.52
– no ASR aux	XLS-R	✓	26.7	54.6	54.6	29.7	57.6	50.0	0.52

Discussion

Why pretraining and distillation help. Self-supervised encoders [1] internalize phonetic and prosodic structure, which is crucial for Uzbek vowels and long suffix chains. Multilingual XLS-R [2] offers cross-lingual robustness that transfers to mixed Uzbek/Russian acoustic conditions. Distillation from a tuned Whisper-based cascade [4] provides stable targets where references are sparse or stylistically variable, aligning punctuation and clause structure in the decoder without over-regularizing translation choices. The combination acts like curriculum: acoustics first, then phrasing.

Conclusion

We demonstrate that end-to-end Uzbek→Russian ST with self-supervised pretraining, multi-task learning, and knowledge distillation is competitive with, and often better than, a strong cascade, particularly on code-switched speech. Script-aware subwording and modest synthetic augmentation further reduce sparsity. These results suggest an actionable recipe for low-resource Turkic→Russian ST: leverage unlabeled audio at scale, keep multi-task auxiliaries during training, distill from a production-ready cascade, and encode script boundaries explicitly.

References

1. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – T. 33. – C. 12449-12460.
2. Babu A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale //arXiv preprint arXiv:2111.09296. – 2021.
3. Bérard A. et al. End-to-end automatic speech translation of audiobooks //2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2018. – C. 6224-6228.
4. Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – C. 28492-28518.
5. Khamidovna N. L. et al. An Online Platform for Uzbek-Russian and Russian-Uzbek Parallel Corpora:: Linguistic Challenges and Prospects Exemplified by A. Kadyri's Novel “Bygone Days” //2024 9th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2024. – C. 11-16.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1