E CONF SERIES

International Conference on Scientific Research in Natural and Social Sciences
Hosted online from New York, USA
Website: econfseries.com 2nd October, 2025

BUILDING AN OPEN UZBEK SPEECH-TO-TEXT STACK FOR LOW-RESOURCE ASR

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, end-to-end speech-to-text stack for Uzbek that lowers the cost of building practical ASR in low-resource settings. The stack combines reproducible data pipelines, self-supervised pretraining, lightweight fine-tuning, and lexicon-free decoding. On public Uzbek splits, fine-tuned wav2vec 2.0 and Whisper outperform strong baselines. We release recipes, tokenizers, and evaluation scripts to enable fair, repeatable benchmarking and rapid local adaptation.

Keywords: Uzbek ASR, low-resource speech, wav2vec 2.0, Whisper, Kaldi, data augmentation, subword modeling, lexicon-free decoding.

Introduction

Automatic speech recognition (ASR) for Uzbek remains constrained by limited labeled audio, code-switching with Russian, and orthographic variation (Latin/Cyrillic). Recent self-supervised learning (Baevski), multilingual pretraining (Radford), and robust decoding with subword units have changed the trade-offs for low-resource languages. This paper describes a fully open Uzbek ASR stack designed for constrained compute and data regimes. We articulate design choices, document a clean data pipeline, and report replicable results on public splits to provide a baseline that future work can extend.

Methods and literature review

We adopt a modular architecture: (i) data curation and normalization for Uzbek (text normalization, script harmonization, punctuation stripping); (ii) grapheme- and BPE-level tokenization; (iii) model families — CTC with wav2vec 2.0, encoder-

decoder with Whisper, and a compact Conformer trained from scratch; (iv) decoding via beam search with and without a small n-gram LM; and (v) evaluation scripts for WER/CER with careful Uzbek-specific preprocessing. Data augmentation uses speed perturbation, SpecAugment (Park), and noise mixing at controlled SNRs to simulate real acoustic conditions.

Our choices follow established evidence that supervised training alone is insufficient for low-resource ASR. Kaldi recipes (Povey) established strong GMM/TDNNF baselines but struggle without large lexicons. Self-supervised approaches (Baevski) learn useful acoustic units from unlabeled audio, while multilingual instruction-tuned models (Radford) transfer well to narrow domains with small fine-tuning sets. Surveys on under-resourced ASR (Besacier) emphasize the importance of reproducible pipelines, transparent evaluation, and language-specific normalization rules. We operationalize these insights in one cohesive stack with minimal external dependencies.

Results

We evaluate on public Uzbek speech data splits (read and spontaneous), reserving speaker-disjoint dev/test sets. Text normalization converts Cyrillic to Latin with deterministic rules; numerals and dates are verbalized; mixed Uzbek–Russian tokens are kept as is if acoustically expressed. Models are trained on a single mid-range GPU (24 GB) to reflect typical university lab constraints. We report case-insensitive WER after normalization.

Fine-tuning wav2vec 2.0-Base yields strong performance with limited labeled hours and remains stable across domains. Whisper-Small, further fine-tuned on Uzbek, achieves the lowest WER on read speech and remains competitive on spontaneous speech despite domain mismatch. A compact from-scratch Conformer benefits most from augmentation but trails pretrained models, confirming the value of self-supervision in low-resource settings.

Removing script harmonization increases WER through token inflation and inconsistent spacing. Turning off SpecAugment reduces robustness to microphone variability. Adding a small 4-gram LM trained on normalized Uzbek web text

modestly helps the CTC system but provides limited gains for Whisper, whose decoder already internalizes a language model.

All numbers are averaged over three seeds; ± indicates standard deviation. Hours denote labeled fine-tuning data only.

Model (params)	Labeled hrs	Dev WER (%)	Test WER (%)	Train time (h)
Conformer-S (15M) scratch	100	29.6 ±0.4	31.2 ±0.6	11.2
wav2vec 2.0-Base (95M) FT	50	22.1 ±0.3	23.8 ±0.5	6.9
wav2vec 2.0-Base (95M) FT + 4-gram LM	50	21.4 ±0.3	23.1 ±0.4	7.1
Whisper-Small (244M) FT	60	18.7 ±0.2	20.9 ±0.3	9.4

Notes. FT = fine-tuned. All systems trained with speed perturbation {0.9, 1.0, 1.1} and SpecAugment; CTC decoding beam=20; LM interpolated with α tuned on dev.

Discussion

Pretrained representations consistently reduce data requirements and training time while improving accuracy. In Uzbek, script harmonization and careful normalization are not superficial details: they are core to reliable scoring and to stable BPE vocabularies. The marginal utility of external LMs diminishes for encoder-decoder models such as Whisper, but remains useful for CTC when transcripts are short and numeric. Beyond accuracy, the stack's value lies in its reproducibility: one command to rebuild features, tokens, data manifests, and scores.

Conclusion

We provide a practical, open Uzbek ASR stack that runs on modest hardware and attains competitive WER with limited labeled data. By combining transparent data handling, self-supervised backbones, and lexicon-free decoding, the system offers a strong, reproducible baseline for research and deployment in Uzbek and related

Turkic languages. The recipes are intended to be extended — by more unlabeled audio, better normalization, and targeted domain adaptation.

References

1. Povey D. et al. The Kaldi speech recognition toolkit //IEEE 2011 workshop on automatic speech recognition and understanding. – 2011. – Т. 1. – С. 5.1.
2. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-12460.
3. Sobirovich S. A. A PRAGMATICALLY ORIENTED APPROACH TO GENERATIVE LINGUISTICS //CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES. – 2024. – Т. 5. – №. 04. – С. 69-75.
4. Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – С. 28492-28518.
5. Besacier L. et al. Automatic speech recognition for under-resourced languages: A survey //Speech communication. – 2014. – Т. 56. – С. 85-100.
6. Авезов С. КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ПОДХОДЫ К АНАЛИЗУ ЯЗЫКА И ИХ ПРИЛОЖЕНИЯ В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ //International Bulletin of Applied Science and Technology. – 2023. – Т. 3. – №. 7. – С. 177-181.