

COMPARATIVE FINE TUNING STRATEGIES FOR UZBEK AUTOMATIC SPEECH RECOGNITION ACROSS MULTILINGUAL FOUNDATION MODELS

Avezov Sukhrob Sobirovich

PhD, Lecturer at the Department of Russian Language and Literature
Bukhara State University (Bukhara, Uzbekistan)

1990senigama@gmail.com

<https://orcid.org/0000-0003-1137-3403>

Abstract. *This article presents a comparative analytical study of six model families relevant to Uzbek automatic speech recognition, namely Whisper, wav2vec 2.0, XLS R, MMS, SeamlessM4T, and Uzbek end to end systems trained on the USC corpus. The analysis is based on published academic evidence rather than a newly run laboratory benchmark. The central question is practical and linguistic at once. Which architecture is most suitable for Uzbek fine tuning under limited transcribed data, script variation, and morphologically dense word forms. The comparison integrates corpus scale, pretraining logic, adaptation behavior in low resource settings, language model compatibility, and risk zones specific to Uzbek speech and orthography.*

Keywords: *Uzbek ASR, fine tuning, low resource speech recognition, multilingual foundation models, Whisper, XLS R, MMS, SeamlessM4T, speech corpora*

Introduction. Uzbek ASR remains a strategically important yet methodologically difficult area because the language combines agglutinative morphology, orthographic variation, and a still limited supply of openly reusable transcribed speech. The public resource situation has improved. USC

introduced 105 hours of transcribed Uzbek speech from 958 speakers, while FeruzaSpeech added a 60 hour read speech corpus with transcripts in both Cyrillic and Latin. At the broader multilingual level, Common Voice and FLEURS have made cross language evaluation and transfer learning far more realistic for low resource research.

Methods. Six model families were selected because each represents a distinct adaptation logic for low resource speech. Whisper stands for large scale weak supervision. wav2vec 2.0 provides the core self supervised paradigm. XLS R extends that paradigm cross linguistically. MMS maximizes language coverage and adapter based adaptation. SeamlessM4T adds unified multilingual multimodality. Uzbek end to end systems on USC represent the local empirical baseline against which imported architectures can be interpreted. The comparison used five criteria. First, pretraining scale and multilingual breadth. Second, suitability for Uzbek fine tuning under small and medium labeled data. Third, sensitivity to script variation between Latin and Cyrillic. Fourth, compatibility with language model rescoring and post correction. Fifth, downstream usefulness when the ASR output later feeds speech translation or dubbing pipelines. These criteria are directly motivated by the published properties of the models and by Uzbek specific corpus work.

The Uzbek side of the analysis was anchored in three corpus facts. USC contains 105 hours of transcribed audio from 958 speakers and produced a best reported test WER of 17.4 for the end to end system in the initial experiments. FeruzaSpeech contributes 60 hours of high quality read speech from a native speaker in Tashkent with dual script transcripts. FLEURS contributes a multilingual parallel benchmark with about 12 hours of supervision per language, which is useful for controlled cross language comparison even when it does not replace full domain evaluation.

Model family	Training logic	Published scale	Uzbek relevance	Main constraint
Whisper	weakly supervised encoder decoder	680,000 hours multilingual and multitask supervision	strongest off the shelf zero shot entry point for noisy and mixed web style audio	orthographic output is less controllable and full fine tuning is computationally heavy
wav2vec 2.0	self supervised pretraining plus supervised fine tuning	53k hours unlabeled speech in the original paper	conceptually strong when Uzbek labels are scarce and adaptation is targeted	original framework is not itself a multilingual Uzbek ready solution
XLS R	cross lingual self supervised speech representations	nearly half a million hours in 128 languages with models up to 2B parameters	excellent balance between transferability and controlled fine tuning for Uzbek	requires careful tokenizer and decoder choices in low script consistency settings

MMS	massively multilingual wav2vec based system with language specific adapters	1,406 languages for pretraining and 1,107 for multilingual ASR	most attractive when Uzbek labeled audio is extremely small	adapter centered setup still benefits from external language modeling and text cleanup
SeamlessM4 T	unified multimodal and multitask model	1 million hours of open speech audio for speech representations and ASR for up to 100 languages	best when Uzbek ASR is part of a larger speech translation workflow	heavier stack and less transparent pure ASR error control
USC end to end Uzbek baseline	locally trained E2E transformer conformer style systems	105 hour Uzbek corpus from 958 speakers	essential baseline for realistic Uzbek domain evaluation	weaker transfer compared with multilingual foundation models

Results. The first result is strategic rather than spectacular. No single model dominates every Uzbek scenario. Whisper is the most defensible baseline when one needs immediate transcription quality before any local adaptation. Yet the literature repeatedly shows that self supervised families



become increasingly competitive, and sometimes preferable, once fine tuning begins under controlled low resource conditions. That pattern is especially clear for XLS R and MMS.

For extremely small labeled sets, MMS deserves priority. Liang and Levow show a threshold effect. MMS is strongest at the smallest durations, while XLS R catches up after one hour of data. This matters for Uzbek projects that begin from fragmentary institutional recordings, student collections, or manually aligned pilot sets. In contrast, if a research group already has tens of hours of relatively clean transcribed audio, XLS R becomes a more balanced candidate because it combines large multilingual pretraining with a fine tuning regime that is easier to localize than Whisper.

A second result concerns the role of local baselines. The USC paper remains crucial not because its architecture is globally dominant, but because it shows what can be learned from Uzbek specific data alone. Its reported 17.4 test WER for the best end to end system is not directly comparable to foundation model results from other papers, since corpora, splits, and decoding differ. Still, it proves that Uzbek ASR cannot be treated only as transfer learning imported from global benchmarks. Local corpus design, speaker diversity, and transcription discipline materially affect performance.

A third result is orthographic. FeruzaSpeech is unusually valuable because it includes both Cyrillic and Latin transcripts. For Uzbek, this is not a peripheral convenience. It changes token statistics, normalization rules, and the behavior of named entities, abbreviations, and loanwords. Models with strong multilingual priors can still drift when a training set mixes script standards or uses unstable apostrophe conventions. The practical implication is simple. Dual script normalization must precede serious fine tuning.

Uzbek challenge zone	Example	Why it matters	Most promising model family	Highest risk family
Agglutinative form density	kitoblaringizda n	one token compresses plural possessive and ablative information	XLS R and MMS after fine tuning	zero shot systems without Uzbek specific adaptation
Analytic predicate	qaror qabul qilmoq	lexical meaning is distributed across nominal and verbal parts	Whisper and XLS R with phrase aware fine tuning	purely local small baselines without phrase diversity
Named entity stability	Buxoro davlat universiteti	transliteration and casing inconsistency affects recognition and later translation	Whisper and SeamlessM4T	models trained on narrow local corpora only
Dual script variation	o'qituvchi and уқитувчи	tokenizer fragmentation	MMS and XLS R with	any model fine tuned on



		n and text normalization mismatch	script normalization	mixed raw text without harmonization
Scientific noun phrase	til birliklarining funktional semantik tavsifi	long dependencies amplify deletion and morphology loss	XLS R with LM rescoring	smaller end to end systems without strong text modeling
Evidential nuance	u kelgan ekan	short function material carries discourse meaning	Whisper plus Uzbek text correction or LM	models optimized only for rough phonetic matching

The model level comparison becomes clearer when the diagnostic probes are read against published architecture behavior. Whisper is likely to survive noise and accent shifts better at the initial stage because its weak supervision is vast. XLS R and MMS, however, are more attractive when the goal is disciplined Uzbek adaptation with explicit control over tokenization and CTC style decoding. SeamlessM4T becomes especially relevant once the ASR output must immediately feed speech translation, subtitling, or dubbing.

Discussion. Several implications follow from the comparison. First, Uzbek ASR research should stop framing the choice as Whisper versus everything else. The real question is staged deployment. Whisper is ideal for



the first pass and for rapidly estimating corpus quality. After that point, fine tuning decisions should depend on data volume and target workflow. If the objective is a robust monolingual recognizer, XLS R and MMS become more compelling. If the objective is an integrated speech translation system, SeamlessM4T has an architectural advantage that pure ASR papers cannot fully capture.

Second, Uzbek requires more than acoustic adaptation. The literature on Uzbek language modeling shows why rescoring remains important. N gram approaches help, yet neural language models help more, and Transformer based rescoring is supported by broader ASR research as well. For a language with long suffix chains and unstable orthographic practice across datasets, post decoding linguistic control is not optional. It is part of the model.

Third, corpus expansion must be philologically informed. More hours alone will not solve the problem if the added material reproduces only read speech, only one register, or only one script standard. USC improved the speaker dimension. FeruzaSpeech strengthened transcript quality and script coverage. The next Uzbek ASR corpora should explicitly incorporate dialectal diversity, academic speech, spontaneous dialogue, and controlled parallel transcription in Latin and Cyrillic. Otherwise the model will improve numerically while remaining narrow in real use.

Conclusion. The comparative analysis shows that Uzbek ASR should be organized as a staged research program rather than as a search for one universally best model. For immediate transcription, Whisper remains the strongest zero shot baseline. For controlled fine tuning on limited Uzbek data, XLS R is the most balanced option, while MMS is especially effective at the smallest labeled scale. SeamlessM4T is the best choice when recognition must directly support speech translation. Uzbek specific end to end systems retain

methodological value because they expose local error structure and ground evaluation in real language use.

REFERENCES

1. Dyomochkina V. V., Gruzdev D. Y., Lukyanova E. V. Machine translation in hindsight //Научный результат. Вопросы теоретической и прикладной лингвистики. – 2024. – Т. 10. – №. 2. – С. 21-45.
2. Камшилова О. Н., Беляева Л. Н. Машинный перевод в эпоху цифровизации: новые практики, процедуры и ресурсы //Terra Linguistica. – 2023. – Т. 14. – №. 1. – С. 41-56.
3. Абдурахманова М. Т., Абжалова М. А. Параллельные корпуса на примере узбекского, русского и английского языков //Конвергентные технологии XXI: вариативность, комбинаторика, коммуникация. – 2022. – С. 74-81.
4. Эргашева Г., Хаиткулов З., Кучимова Н. Параллельный корпус в автоматизированном переводе (на примере коллокаций) // O‘zbekistonda xorijiy tillar. – 2023. № 5 (52). – С. 162-173.
5. Митренина О. В., Мухамбеткалиева А. Г. Как и какой перевод (не) оценивают компьютеры //Journal of applied linguistics and lexicography. – 2021. – Т. 3. – №. 2. – С. 77-84.
6. Nigmatova L., Avezov S. ПРИМЕНЕНИЕ МЕТОДОВ NLP В КОРПУСНЫХ ИССЛЕДОВАНИЯХ: ОСОБЕННОСТИ И ОГРАНИЧЕНИЯ //«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция. – 2023. – Т. 2. – №. 2.

7. 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.

