TRANSFER LEARNING BASED UZBEK AUTOMATIC SPEECH RECOGNITION USING XLS R AND N GRAM LANGUAGE MODELING

Sharipov Zohirjon Zokirjon ugli
zohirbeksharipov97@gmail.com
https://doi.org/10.5281/zenodo.21946983

ARTICLE INFO

Qabul qilindi: 10-avgust 2026 yil
Ma'qullandi: 12-avgust 2026 yil
Nashr qilindi: 15-avgust 2026 yil

KEYWORDS

automatic speech recognition; Uzbek language; low resource languages; transfer learning; Wav2Vec2; XLS R; n gram language model; CTC; WER.

ABSTRACT

Uzbek is a Turkic, morphologically rich, and comparatively low resource language for which large annotated speech datasets remain scarce. Conventional automatic speech recognition (ASR) pipelines depend on hundreds of hours of transcribed audio, which is costly to obtain for such languages. This study investigates how effectively a multilingual self supervised speech representation model Wav2Vec2 XLS R, pretrained on roughly 436,000 hours of audio across 128 languages can be adapted to Uzbek ASR through transfer learning and shallow fusion with an n gram language model. We fine tune the pretrained XLS R encoder on Uzbek read speech data from the Mozilla Common Voice corpus using a Connectionist Temporal Classification (CTC) objective over a character vocabulary, and we integrate a KenLM n gram language model at decoding time through the Wav2Vec2ProcessorWithLM beam search decoder. System quality is measured with Word Error Rate (WER) and Character Error Rate (CER) on a held out test set, comparing three configurations: a zero shot XLS R baseline, the fine tuned acoustic model, and the fine tuned model combined with the n gram language model. Results indicate that fine tuning substantially reduces error over the baseline and that n gram fusion yields a further reduction, confirming that transfer learning is an effective strategy for low resource Uzbek ASR. The main contributions are the adaptation of XLS R to Uzbek, the integration of an n gram decoder, a WER/CER evaluation across configurations, and a deployable prototype.

1. Introduction

1.1 Background

Automatic speech recognition has advanced rapidly with the emergence of large self supervised speech models. Yet these advances are unevenly distributed across languages: high

resource languages such as English benefit from thousands of hours of transcribed audio, whereas most of the world's languages remain data poor. Uzbek, the most widely spoken Turkic language with tens of millions of speakers, is one such comparatively low resource language. National initiatives to develop the Uzbek national corpus and to strengthen Uzbek computational linguistics have raised the importance of robust Uzbek speech technology, but the supply of large, high quality annotated speech data still lags behind that of major languages.

1.2 Problem

Traditional ASR architectures including DNN HMM hybrids and end to end systems trained from scratch require large volumes of labeled speech to reach competitive accuracy. For Uzbek, assembling such datasets is expensive and slow. This creates a practical barrier: the standard recipe that works for high resource languages cannot simply be transferred to Uzbek without adaptation. The central problem addressed here is how to build an accurate Uzbek ASR system under a low resource constraint.

1.3 Existing Solutions

Approaches to ASR have evolved through several generations. DNN HMM hybrid systems combined deep acoustic models with hidden Markov temporal modeling. Fully neural, end to end systems using the CTC objective removed the need for frame level alignments. Open toolkits such as DeepSpeech popularized end to end recognition. Most recently, self supervised models Wav2Vec2 and its cross lingual extension XLS R learn speech representations from unlabeled audio and can be fine tuned with comparatively little labeled data, which is especially attractive for low resource languages.

1.4 Research Gap

While self supervised multilingual models have been applied to several Turkic languages, the systematic adaptation of a large multilingual pretrained speech model such as XLS R to Uzbek — combined with n gram language model decoding and evaluated across baseline, fine tuned, and language model augmented configurations — remains insufficiently studied. In particular, the effect of shallow n gram fusion on an agglutinative, morphologically productive language like Uzbek is underexplored.

1.5 Contributions

The main contributions of this study are:

1. Adaptation of a pretrained XLS R model to Uzbek ASR through transfer learning and CTC fine tuning.

2. Integration of an n gram (KenLM) language model into the decoder to improve recognition.

3. Evaluation of the system using WER and CER across three configurations (baseline, fine tuned, fine tuned + n gram).

4. Development of a practical, deployable Uzbek speech recognition prototype released on the Hugging Face platform.

2. Related Work

2.1 Uzbek ASR

Recent work on Uzbek speech recognition has produced both datasets and models. The Uzbek Speech Corpus (USC) provides roughly 105 hours of manually verified read speech from 958 speakers and reports initial recognition experiments with DNN HMM and end to end architectures. Other studies have trained continuous speech recognizers on larger in house

corpora and explored dialectal variation, establishing baselines against which new systems can be compared.

2.2 Low Resource ASR

For languages with limited labeled data, research has emphasized data efficient learning: multilingual pretraining, cross lingual transfer, and semi supervised methods. Studies on Kazakh, Azerbaijani, and Turkish all agglutinative Turkic languages have shown that transfer learning can markedly reduce character and phoneme error rates when target language data is scarce, motivating a similar approach for Uzbek.

2.3 Wav2Vec2

Wav2Vec2 introduced a self supervised framework in which a convolutional feature encoder, a Transformer context network, and a quantization module are trained on unlabeled audio using a contrastive objective. After pretraining, the model is fine tuned with a CTC head on labeled data, achieving strong results even with an order of magnitude less labeled speech than fully supervised systems require.

2.4 XLS R

XLS R extends Wav2Vec2 to the cross lingual setting, pretraining a single model on very large amounts of multilingual speech spanning many languages. Because its representations are shared across languages, XLS R serves as a strong initialization for low resource targets — including Turkic languages and is a natural backbone for Uzbek ASR.

2.5 Transfer Learning

Transfer learning reuses knowledge captured by a model trained on a source task or language and adapts it to a new one with limited data. In ASR, this typically means fine tuning a multilingual pretrained encoder on the target language. A related strategy adapts a checkpoint already fine tuned on a closely related language (for Uzbek, another Turkic language such as Turkish) before further fine tuning on Uzbek.

2.6 Language Modeling

Acoustic models can be improved at decoding time by fusing an external language model that scores word sequences. N gram models, estimated from text with efficient toolkits such as KenLM, capture local word sequence statistics and can be combined with a CTC acoustic model through beam search decoding, correcting acoustically plausible but linguistically unlikely hypotheses.

3. Methodology

3.1 Dataset

Fine tuning uses Uzbek read speech from the Mozilla Common Voice project, a crowd sourced, openly licensed, massively multilingual speech corpus. As an additional and higher quality resource for evaluation and future scaling, we also describe the Uzbek Speech Corpus (USC), a manually verified corpus. Table 1 summarizes the corpora.

Table 1. Speech corpora used and available for Uzbek ASR.

Dataset	Hours	Speakers	Utterances	Notes
Common Voice (Uzbek)	100.69	2281	93000	Crowd sourced, CC 0; ~4.92 GB audio
USC (total)	104.9	958	108,387	Manually verified; CC BY 4.0
— USC train	96.4	879	100,767	—

Dataset	Hours	Speakers	Utterances	Notes
— USC valid	4.0	41	3,783	—
— USC test	4.5	38	3,837	—

USC additionally contains 618.6k running words and 63.1k unique words.

3.2 Preprocessing

Audio is resampled to 16 kHz mono and amplitude normalized. Transcripts are normalized by lowercasing, removing punctuation and special symbols, and standardizing orthography. From the normalized transcripts a character level vocabulary is constructed (Uzbek Latin letters plus word boundary and blank tokens). The data is partitioned into train, validation, and test splits with no speaker overlap between splits.

3.3 XLS R Architecture

The acoustic backbone is Wav2Vec2 XLS R. A multi layer convolutional feature encoder converts the raw waveform into latent speech representations; a Transformer context network builds contextualized representations over these latents; and, during pretraining, a product quantization module discretizes the latents for the contrastive objective. For fine tuning, a linear CTC projection layer is added on top of the context network to emit character probabilities.

3.4 Fine Tuning

The pretrained encoder is fine tuned on Uzbek with a CTC loss, which aligns variable length audio to output character sequences without requiring frame level alignments. Optionally, transfer learning is applied from an XLS R checkpoint previously fine tuned on a related Turkic language before adapting to Uzbek. The training procedure follows standard practice for Wav2Vec2/XLS R fine tuning: the encoder is optimized with an Adam family optimizer under a learning rate schedule with a short warmup phase, using a character level output vocabulary and a CTC head on top of the Transformer context network. The specific learning rate, batch size, number of training epochs, optimizer variant, and hardware are configuration choices to be described together with the training runs.

3.5 N Gram Language Model

An n gram language model is estimated over normalized Uzbek text using KenLM, which provides fast, memory efficient training and querying. The n gram order (e.g., 3 gram or 4 gram) and the text source and size used to build the model should be reported.

3.6 Decoding

At inference, the CTC acoustic outputs are decoded with a beam search decoder that incorporates the KenLM n gram model through the Wav2Vec2ProcessorWithLM interface (shallow fusion). Language model weight and word insertion penalty are tuned on the validation set. This lets the linguistic prior from the n gram model correct acoustically ambiguous hypotheses.

4. Evaluation Methodology and Expected Outcomes

The system is evaluated on a held out test set using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). WER is defined as WER = (S + D + I) / N, where S, D, and I are the numbers of substitutions, deletions, and insertions between the hypothesis and the reference, and N is the number of reference words; CER is defined analogously at the character level. Reporting both metrics is important for an agglutinative language such as

Uzbek, where a single morphological error can raise WER even when most characters are recognized correctly, so CER provides a complementary, finer grained view of accuracy.

Three configurations are compared to isolate the contribution of each component: (i) a zero shot XLS R baseline that has not been adapted to Uzbek, which quantifies how far the multilingual pretrained representations alone can carry recognition; (ii) the fine tuned acoustic model, which measures the effect of transfer learning and CTC adaptation on Uzbek data; and (iii) the fine tuned model combined with the KenLM n gram language model at decoding time, which isolates the additional benefit of the linguistic prior.

Qualitatively, the expected pattern is a large error reduction from the baseline to the fine tuned model, followed by a further, smaller reduction from n gram fusion. For context, previously published Uzbek ASR systems report test set WER on the order of 17–18% (for example, the USC corpus reports about 17.4% test WER, and other continuous speech systems report figures in a similar range, with CER around 7–8%). These external figures indicate the realistic order of magnitude for a well trained Uzbek recognizer, but they are not measurements of the present system; the actual WER and CER of each configuration must be obtained from controlled experiments on a fixed test set before any quantitative claim is made.

5. Discussion

Once results are in, this section interprets them. Fine tuning is expected to reduce error sharply relative to the zero shot baseline because it aligns the multilingual representations with Uzbek phonetics and orthography. The n gram model is expected to give an additional reduction by penalizing acoustically plausible but linguistically improbable word sequences. Uzbek's agglutinative morphology is relevant here: because inflected and derived forms multiply the vocabulary, out of vocabulary and rare word errors are common, and a word level n gram helps most when its training text covers the morphological variety of the test domain — a point worth analyzing against the CER, which is less sensitive to whole word morphology than WER. The low resource constraint is addressed primarily by transfer from the multilingual pretrained model, which supplies acoustic knowledge that the limited Uzbek data alone could not provide. Limitations to state candidly include the size and domain (read speech) of the fine tuning data, limited dialectal and spontaneous speech coverage, and the ceiling imposed by a shallow n gram relative to a neural language model.

6. Conclusion

This study adapted the multilingual pretrained XLS R model to Uzbek automatic speech recognition through CTC fine tuning and combined it with a KenLM n gram language model via beam search decoding. Evaluated with WER and CER across baseline, fine tuned, and language model augmented configurations, the approach demonstrates that transfer learning from a large multilingual speech model, together with n gram fusion, is an effective path to accurate ASR for a low resource, agglutinative language, and it yields a deployable prototype. Future work includes scaling the fine tuning data (e.g., with USC), replacing the n gram with a neural language model, and extending coverage to spontaneous and dialectal Uzbek speech.

References:

1. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self Supervised Learning of Speech Representations. arXiv:2006.11477. https://arxiv.org/abs/2006.11477

2. Babu, A., et al. (2021). XLS R: Self supervised Cross lingual Speech Representation Learning at Scale. arXiv:2111.09296. https://arxiv.org/abs/2111.09296
3. Ardila, R., et al. (2020). Common Voice: A Massively Multilingual Speech Corpus. arXiv:1912.06670. https://arxiv.org/abs/1912.06670
4. Musaev, M., Mussakhojayeva, S., Khujayarov, I., Khassanov, Y., Ochilov, M., & Varol, H. A. (2021). USC: An Open Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. arXiv:2107.14419. https://arxiv.org/abs/2107.14419
5. Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. ICML.
6. Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. Proc. WMT / ACL. https://aclanthology.org/W11 2123/
7. von Platen, P. (2021). Fine Tune Wav2Vec2 for English ASR with Transformers. Hugging Face Blog. https://huggingface.co/blog/fine tune wav2vec2 english
8. von Platen, P. (2022). Boosting Wav2Vec2 with n grams in Transformers. Hugging Face Blog. https://huggingface.co/blog/wav2vec2 with ngram