Scaling ASR for Hutsul Dialect: Multi-Speaker Data Collection, Enhanced Transcription and Cross-Speaker Evaluation

Artem Orlovskyi Zakhar Guzii Bohdan Onyshchenko
Roman Kyslyi Pavlo Khomenko
Kyiv School of Economics, Ukraine
(aorlovsky,zguzii,bonyshchenko,rkyslyi,pkhomenko)@kse.org.ua

Abstract

We present a significant expansion of automatic speech recognition (ASR) resources for the Hutsul dialect of Ukrainian, building on prior work that established the first aligned speech corpus from a single literary source. In this work, we scale the dataset from a single speaker to a multi-speaker corpus comprising 40 speakers and 60.63 hours of audio drawn from diverse sources: YouTube channels (with author permissions), field recordings from native speakers, linguist student recordings, and regional radio broadcasts. To obtain reference transcriptions for audio without existing text, we introduce a novel retrieval-augmented generation (RAG) correction pipeline: audio is first transcribed using ElevenLabs, then corrected through a RAG pipeline backed by a dialect-aware language model. We evaluate fine-tuned ASR models across five distinct speaker datasets, demonstrating that while the best model achieves strong performance on in-domain speakers (character error rate, CER, 3.24%), cross-speaker generalisation remains challenging, with CER ranging from 5.33% to 17.24% depending on speaker characteristics. The cross-speaker gap of 5–14 percentage points indicates that single-speaker-dominated training data is insufficient for robust dialect ASR and motivates future work on speaker-balanced corpora and adaptation methods. All data, code, and models are released publicly to support further research on Ukrainian dialect speech technologies.

1 Introduction

Automatic Speech Recognition (ASR) for low-resource dialects has lagged behind ASR for standard languages, despite the cultural and linguistic value such dialects carry. The Hutsul dialect, spoken in the Ukrainian Carpathians, exhibits phonological, morphological, and lexical features that differ markedly from standard Ukrainian. As Ukrainian ASR systems are increasingly deployed in education, broadcasting, and accessibility tools, their failure on Hutsul speech effectively excludes a sizeable community of speakers. A dedicated dataset is therefore not only a methodological convenience but a prerequisite for equitable language technology in Ukraine: Hutsul speech contains forms that no general-purpose Ukrainian system has been trained to recognise, and absent dialect-targeted resources, errors are systematically biased toward standardisation.

Our previous work (Kyslyi et al., 2026) presented the first dedicated ASR resources for the Hutsul dialect of Ukrainian, centred on a single-speaker corpus derived from readings of the novel "Dido Yvanchyk." That work established a data preparation pipeline, benchmarked multiple ASR architectures, and demonstrated that fine-tuning can reduce Character Error Rate (CER) from over 17% to below 3% on single-speaker dialectal speech.

However, real-world dialect ASR faces challenges that a single-speaker corpus cannot address. The Hutsul dialect exhibits substantial inter-speaker variation: pronunciation, lexical choice, and morphological forms differ across villages, generations, and individual speaking styles. Throughout this paper we use *single-speaker setup* to refer to training and evaluation data drawn from a single human speaker (as in our prior corpus), and *multi-speaker setup* to refer to data drawn from many distinct speakers with diverse demographic, geographic, and stylistic profiles. A model trained on one speaker may fail to generalise to others, limiting practical applicability.

In this paper, we address three key limitations of our prior work:

1. **Speaker diversity:** We expand the corpus from a single speaker to 40 speakers across 60.63 hours of audio, incorporating YouTube content creators, field recordings, university

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 184–198

students, and radio broadcasts.

2. **Transcription for untexted audio:** Unlike the novel-based corpus where ground-truth text existed, our new sources lack reference transcriptions. We develop a RAG-enhanced correction pipeline that combines automatic transcription with dialect-aware language model post-processing.

3. **Cross-speaker evaluation:** We systematically evaluate how a Hutsul-tuned ASR model generalises across speakers with different backgrounds, recording conditions, and dialectal characteristics.

The methodology is dialect-agnostic: the same combination of permissioned audio harvesting, RAG-based dialect correction, and cross-speaker evaluation can be applied to other Ukrainian dialect groups (e.g. Boyko, Lemko, Polissian) wherever a small seed corpus of attested dialect text exists. We view this as a practical recipe for scaling dialect ASR beyond Hutsul.

Our results reveal that while fine-tuned models achieve strong in-domain performance, cross-speaker generalisation remains an open challenge for dialect ASR, motivating further work on multi-speaker training and speaker-adaptive methods.

2 Related Work

2.1 Dialect ASR

Research on ASR for low-resource languages and dialects has expanded significantly, with work in Arabic (Ali et al., 2014), Hindi (Javed et al., 2024), and other language families demonstrating that dialectal variation poses persistent challenges even for large pretrained models. For Ukrainian specifically, existing ASR systems target the standard language (Paniv, 2023; arampacha, 2024), and our previous work (Kyslyi et al., 2026) provided the first Hutsul dialect benchmark.

2.2 Post-Processing and Error Correction for ASR

ASR output for low-resource languages and dialects frequently contains systematic errors that can be addressed through post-processing. Language model rescoring (Radford et al., 2023) and n-best re-ranking are established techniques, but they require language models trained on dialect text, which is scarce for Hutsul.

Recent work has explored using large language models (LLMs) for ASR error correction. The VuykoMistral approach (Kyslyi et al., 2025) demonstrated that a Mistral-based model fine-tuned on Ukrainian dialect text can serve as an effective post-processor for ASR output, correcting morphological and lexical errors while preserving dialect-specific forms. Our RAG-corrector pipeline builds on this idea, using retrieval-augmented generation to ground corrections in attested dialect text.

2.3 Multi-Speaker Dialect Corpora

Building multi-speaker corpora for dialects requires addressing speaker recruitment, consent, recording standardization, and transcription challenges. Prior work on Scottish Gaelic (Klejch et al., 2025) and Arabic dialects (Ali et al., 2014) has documented these challenges. Our approach combines opportunistic data collection (YouTube, radio) with structured elicitation (student recordings, field work).

3 Data Collection and Corpus Expansion

3.1 Overview

Our expanded corpus draws from five distinct source categories, each presenting unique characteristics and challenges. Table 1 summarizes the data sources.

3.2 YouTube Channels

We identified and contacted several Hutsul-language YouTube content creators who produce regular videos featuring dialect speech. After obtaining explicit permission from channel owners, we downloaded and processed audio from two narrative/podcast-style channels (12.55 hours combined: 10.72 hours from one channel and 1.83 hours from another). The content ranges from personal vlogs and storytelling to cultural commentary, providing diverse speaking styles and topics.

Audio was extracted, resampled to 16 kHz, and segmented using our existing pipeline. Unlike the Dido Yvanchyk corpus, these recordings lack reference text, necessitating the RAG-corrector pipeline described in Section 4.

3.3 Yaroslav Zelenchuk Recordings

We collected 4 hours of recordings from Yaroslav Zelenchuk and his father, native Hutsul speakers from the Verkhovyna district (Ivano-Frankivsk region). These recordings include both spontaneous

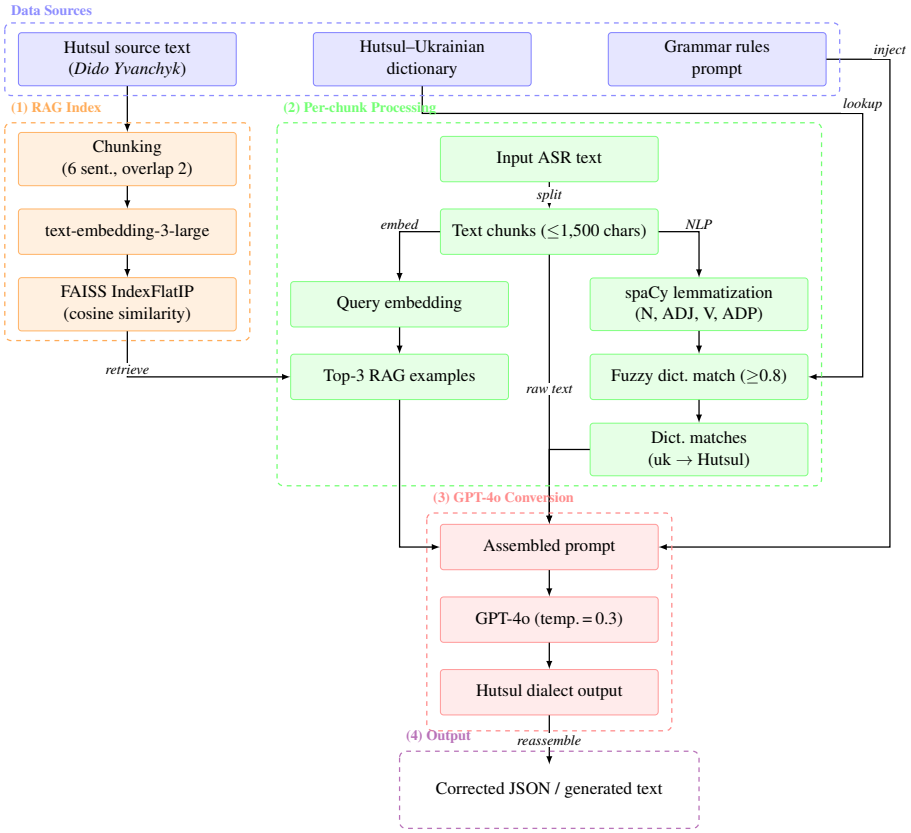


Figure 1: RAG-Corrector pipeline for Hutsul dialect transcription correction. Input ASR transcriptions are split into chunks, simultaneously (a) embedded and matched against a FAISS index of real Hutsul dialect text for retrieval of semantically similar passages, and (b) lemmatised for fuzzy dictionary lookup. Both results and explicit grammar rules are assembled into a GPT-4o prompt (temperature = 0.3) which rewrites the text in authentic Hutsul dialect.

Source	Type	Speakers	Duration	Notes
Dido Yvanchyk	Novel reading	1	15.69h	Updated version of original corpus
Hutsulendia	Broadcast	8	2.89h	Regional Hutsul-dialect programme
Yaroslav recordings	Field recordings	1	4.00h	Native speakers, spontaneous speech
NaUKMA students	Read	28	25.50h	Linguistics students at NaUKMA
YT-channel2	Podcast / narrative	1	10.72h	Hutsul & Bukovynian ethnography
YT-channel1	Podcast / narrative	1	1.83h	Single speaker, used by permission
Total		**40**	**60.63h**	

Table 1: Summary of data sources in the expanded Hutsul dialect corpus.

conversational speech and semi-structured narratives about local history and traditions. Recordings were made with a portable handheld digital recorder using its built-in cardioid microphones at 48 kHz/24-bit, in quiet indoor conditions in the speakers' homes; audio was subsequently downsampled to 16 kHz mono and loudness-normalised before segmentation.

3.4 Linguistics Student Recordings

Linguistics students from the National University of Kyiv-Mohyla Academy (NaUKMA) with Hut-

sul heritage participated in recording sessions. 26 identified speakers, plus an additional 2 distinct but anonymised speaker IDs assigned to segments whose individual identity could not be determined, contributed 25.50 hours of speech for a total of 28 speaker entries (matching Table 1). Recordings include both read passages from Hutsul literary texts and semi-spontaneous speech (e.g., retelling stories, describing their home regions). This source provides younger-generation speakers whose dialect may show more influence from standard Ukrainian, complementing the older speakers in the field-recording and broadcast subsets.

3.5 Hutsulendia Broadcasts

We also obtained 2.89 hours of recordings from the *Hutsulendia* programme, a regional broadcast featuring the Hutsul dialect. The material includes interviews, cultural programmes, and local news segments from 8 distinct speakers. These broadcasts provide professional-quality audio but feature varying degrees of dialect usage, from strong dialect to code-switching with standard Ukrainian. Together with the YouTube and field-recording subsets, this source helps cover speakers from different generations and from a range of villages within the broader Hutsul region.

4 RAG-Enhanced Transcription Pipeline

4.1 Motivation

Our original corpus benefited from an existing written text (the Dido Yvanchyk novel) that could serve as ground truth for alignment. The new data sources - YouTube channels, field recordings, student speech, and radio broadcasts - lack pre-existing transcriptions. Simply transcribing with a standard Ukrainian ASR model produces output riddled with dialect-specific errors: vowel substitutions (y↔i in Ukrainian orthography), morphological normalizations (dialect endings replaced with standard forms), and lexical substitutions (dialect words replaced with standard equivalents).

4.2 Pipeline Architecture

To address this, we developed a two-stage transcription pipeline inspired by the VuykoMistral approach (Kyslyi et al., 2025):

Stage 1: Initial Transcription. Raw audio is transcribed using ElevenLabs STT (ElevenLabs, 2024), which we found in our previous work (Kyslyi et al., 2026) to produce the most reliable word-level timestamps for expressive and dialectal Ukrainian speech. The transcription provides a reasonable first pass but systematically normalizes dialect features toward standard Ukrainian.

Stage 2: RAG-Based Correction. The initial transcription is then processed through a retrieval-augmented generation (RAG) correction pipeline:

1. **Retrieval:** For each transcribed segment, we retrieve the most similar passages from a dialect text knowledge base. This knowledge base includes the full text of the Dido Yvanchyk novel, other Hutsul literary works, and a curated dictionary of Hutsul dialectal forms.

2. **Context construction:** Retrieved passages are combined with the ASR output to form a prompt that provides dialect context.

3. **Generation:** A dialect-aware language model (based on the VuykoMistral architecture) generates a corrected transcription that restores dialect-specific forms while preserving the acoustic content of the original ASR output.

4. **Confidence filtering:** Corrections are accepted only when the model's confidence exceeds a threshold, preventing hallucinated edits. Segments with low correction confidence are flagged for manual review.

4.3 Comparison with Direct ASR

The RAG-corrector pipeline offers several advantages over training a dialect-specific ASR model from scratch for transcription:

- It leverages existing high-quality ASR systems (ElevenLabs) for acoustic modeling while focusing corrections on dialect-specific linguistic features.

- The retrieval component grounds corrections in attested dialect text, reducing hallucination risk.

- The pipeline can be iteratively improved as more dialect text becomes available in the knowledge base.

A detailed evaluation of the RAG-corrector's accuracy on held-out manually transcribed segments is presented in Section 4.4.

4.4 RAG-Corrector Evaluation

A preliminary evaluation by a native Hutsul speaker on a held-out set of 412 manually transcribed segments (drawn from the YouTube and *Hutsulendia* subsets, none of which appear in the RAG knowledge base) shows that the RAG-corrector pipeline improves the proportion of segments rated as fully correct from approximately 40% for raw ElevenLabs output to 71% after dialect-aware correction. The gains are concentrated in dialect-specific lexical items and morphological endings where the baseline STT system defaults to standard Ukrainian forms. A more comprehensive evaluation that disentangles the contribution of retrieval from that of the LLM (i.e. correction with the same dialect-aware LLM but *without* retrieved context) is in progress and is left for future work; this ablation will quantify how much of the observed gain comes specifically from the RAG component versus the underlying LLM's prior knowledge.

5 ASR Models and Training

To benchmark the expanded corpus across architectural families and to make the result directly comparable with our previous single-speaker study (Kyslyi et al., 2026), we evaluate three model families: the encoder–decoder Whisper family (Section 6.1), the CTC-based Wav2Vec 2.0 / XLS-R model (Section 6.3), and the recently released CTC-based Omnilingual ASR models (team et al., 2025) (Section 6.2). Whisper is included because it remains the most widely-deployed open ASR system for Ukrainian and is therefore a natural out-of-the-box baseline for any new dialect; Wav2Vec 2.0/XLS-R is included because it is the strongest pure-CTC baseline previously reported for Ukrainian; and Omnilingual ASR is our flagship system for cross-speaker evaluation, as it achieved the lowest single-speaker CER (2.75%) in our prior work and remains the most accurate model in the multi-speaker setting (Section 6.2).

Building on our previous findings (Kyslyi et al., 2026), our flagship model is OmniASR-CTC-1B. It was fine-tuned on the expanded multi-speaker Hutsul corpus using the same training configuration as our prior work: learning rate 5×10^{-5} with a tri-stage scheduler, per-device batch size of 8 with gradient accumulation of 4, and on-the-fly augmentation including Gaussian noise injection, pitch shifting, speed perturbation (0.8–1.2×), gain modulation, and time/frequency masking (Bartelds et al., 2023).

The fine-tuned model is released as `KSE-RESEARCH-Group` on Hugging Face, together with the expanded `ukr-dialects-audio-dataset` corpus and all training and evaluation scripts (see Section 8).

6 Experimental Results

We evaluate all fine-tuned models across five cross-speaker test sets plus an aggregate set, reporting WER and CER. Full cross-speaker results for individual model families appear in Appendix A; training dynamics are shown in Appendix B.

6.1 Whisper Family

We fine-tuned four Whisper variants: whisper-small, whisper-medium, whisper-large-v3, and arampacha/whisper-large-uk-2 (arampacha, 2024), the latter already adapted to standard Ukrainian via Common Voice 11.0 (Ardila et al., 2020). Fine-tuned checkpoints are publicly available on Hugging Face under the `KSE-RESEARCH-Group` organisation (Section 8). Training was performed in FP16 mixed-precision mode with CER as the primary selection criterion, as character-level evaluation captures dialectal orthographic variation more precisely than WER for the highly inflected morphology of Ukrainian (Thennal D K et al., 2025).

6.1.1 Training Setup

All models were trained using the Hugging Face Seq2SeqTrainer with the AdamW optimizer, a peak learning rate of 1×10^{-5}, 500 linear warm-up steps, and linear decay thereafter. Following the standard Whisper training objective (Radford et al., 2023), models were optimised using cross-entropy loss over decoder token predictions, with padding positions excluded from the loss computation. Checkpoints were saved and evaluated every 500 steps; the checkpoint with the lowest validation CER was retained for final evaluation. Training dynamics for all Whisper variants are shown in the left column of Figure 2 (top: WER, middle: CER, bottom: train/eval loss).

The three generic checkpoints (whisper-small, whisper-medium, whisper-large-v3) were trained for 8,000 steps on the same split of the Ukrainian dialect corpus (27,518 train / 3,347 validation / 3,435 test utterances, maximum token length 448). Per-device batch sizes varied by model due to memory constraints: whisper-large-v3 used batch size 4

with gradient accumulation 2 (effective 8), whisper-medium used batch size 8 with gradient accumulation 4 (effective 32), and whisper-small used batch size 4 with gradient accumulation 4 (effective 16) in Phase 1. The Ukrainian-adapted checkpoint (arampacha/whisper-large-uk-2) was trained for 4,000 steps with batch size 16 and gradient accumulation 2 (effective 32), with gradient checkpointing enabled; its prior Ukrainian fine-tuning yields a stronger initialisation that requires less dialect adaptation.

To reduce hallucination artefacts common in Whisper on expressive speech (Radford et al., 2023), whisper-small was trained in two phases. Phase 1 (steps 1–4,000) used standard training without suppression. Phase 2 (steps 4,001–8,000) introduced no_repeat_ngram_size=3, repetition_penalty=1.3, and condition_on_previous_text=False, while increasing the per-device batch size from 4 to 16 (effective batch size 64). The boundary between phases is marked by the red dashed vertical line at step 4,000 in Figure 2. The two larger generic models (whisper-medium, whisper-large-v3) applied repetition suppression throughout training. The Ukrainian-adapted checkpoint was trained without suppression, as the pre-adapted weights exhibited lower baseline hallucination rates on Ukrainian text. All runs used an NVIDIA GeForce RTX 4090 (48GB VRAM).

The training-step budgets (8,000 for the three generic Whisper checkpoints and 4,000 for the Ukrainian-adapted checkpoint) were chosen empirically from the training dynamics in Figure 2: in each run the validation CER plateaus and starts to drift while training loss continues to decrease, which is the standard sign that further optimisation produces overfitting rather than additional gain. Running beyond 8,000 steps for the larger Whisper variants did not improve validation CER in pilot experiments. The Ukrainian-adapted checkpoint reaches its best validation CER as early as step 3,000 and is therefore trained for only 4,000 steps. We retain the checkpoint with the lowest validation CER for evaluation, which functions as an early-stopping rule.

6.1.2 Results

In-domain test results on the Hutsul corpus are shown in Table 4. whisper-large-v3 achieves the lowest aggregate CER among all Whisper variants (9.32%, WER 25.18%) at step 8,000, followed by whisper-medium (CER 9.99%) and whisper-small (CER 12.54%). The Ukrainian-adapted checkpoint (arampacha/whisper-large-uk-2) attains aggregate CER 10.67% (WER 29.37%), reaching its best checkpoint at step 3,000—the earliest of any variant—which demonstrates the efficiency of dialect adaptation when starting from a language-specific prior, despite not surpassing whisper-large-v3 in final accuracy.

Evaluated across the broader six-dataset Ukrainian dialect benchmark (Table 4), whisper-large-v3 achieves the lowest macro-average WER (25.71%), followed by whisper-medium (28.31%), the Ukrainian-adapted whisper-large-uk-2 (29.82%), and whisper-small (35.79%), with CER following the same order. The Hutsulendia dataset consistently yields the highest WER across all variants (32.01%–45.85%), while near-standard southwestern varieties (YT-channel2) achieve the lowest (16.37%–28.35%).

6.1.3 Analysis of Results

Among the three generic checkpoints, CER decreases monotonically with model scale (Small → Medium → Large-v3), consistent with larger Whisper encoders capturing finer-grained phonetic distinctions (Radford et al., 2023). Notably, arampacha/whisper-large-uk-2 achieves comparable per-dataset CER to whisper-large-v3 on several evaluation sets while using only half the training steps (4,000 vs. 8,000), confirming that prior exposure to standard Ukrainian provides a more efficient starting point for dialect adaptation. Larger models also converge faster: whisper-large-v3 drops below 12% validation CER by step 3,000, whereas whisper-small never reaches this level (Figure 2).

The two-phase training strategy for whisper-small – introducing repetition suppression in Phase 2—yielded consistent improvements across all evaluation datasets, with the largest gain on Dido-Yvanchyk (ΔWER -4.08 pp). The Hutsulendia dataset consistently yields the highest WER across all variants (6–10 pp above the macro-average), driven by spontaneous multi-speaker speech with dense Hutsul-specific vocabulary. Character-level error analysis reveals vowel confusion—"i/и" merger and "o/y" shift—as the dominant failure mode, with morphological suffix errors accounting for the largest share of total character-level errors across all variants.

Dialect-specific vocabulary analysis exposes two

distinct recognition regimes. High-frequency Hutsul lexemes with phonological proximity to standard Ukrainian are handled reliably: “тогди” (*tohdy*, F1=0.90), “файно” (*fayno*, F1=0.86), and “тимунь” (*tymun’*, F1=0.93) are correctly transcribed in the overwhelming majority of occurrences. By contrast, words whose surface form diverges substantially from any standard Ukrainian equivalent suffer systematic misrecognition: “відки” (*vidky*, F1=0.22) is consistently mapped to “звідки” (*zvidky*), “бірше” (*birshe*, F1=0.40) to “більше” (*bil’she*), and “таки” (*taky*, F1=0.83) alternates between “так” (*tak*) and “та” (*ta*). This pattern reveals that the model systematically normalises dialectal forms toward their standard Ukrainian counterparts rather than producing random errors, suggesting that the decoder’s language prior–inherited from multilingual pretraining–actively suppresses out-of-vocabulary dialect forms in favour of the nearest phonologically similar standard word.

6.2 Omnilingual ASR

6.2.1 Training Configuration and Results

We fine-tuned two OmniASR CTC models released by Meta (300M and 1B parameters). Training was performed using the official (team et al., 2025) tutorial on the "ukr-dialects-audio-dataset". Both models were trained with a learning rate of 5×10^{-5} using a tri-stage scheduler. The 300M model was trained on an RTX 5070 Ti (16GB) for 60k steps, while the 1B model was trained on an RTX 4090 (48GB) for 42k steps. We used WER as the primary optimization metric and CTC as loss function. Fine-tuning successfully reduced the initial WER from 76–80% down to around 28.5% and CER from 37–51% to around 11.5%, demonstrating strong dialect adaptation. Detailed cross-speaker evaluation metrics are presented in Table 3 (Appendix A).

6.2.2 Lexical and Morphological Analysis

Evaluating the 1B model on 3,451 utterances revealed a high lexical recall (85.5%) for dictionary-verified Hutsul terms, with only 3.0% of dialectal words dropped entirely. The primary failure mode is orthographic regularization driven by phonological interference (e.g., frequent shifts between "і"(і)↔"и"(у) and "о"(о)↔"у"(u)).

The model tends to overwrite unique Hutsul lexical endings with standard Ukrainian equivalents, resulting in most of the character errors in the suffix zone. Consequently, while standard-adjacent dialect words ("дуже" (duzhe), "файно" (faino)) are robustly transcribed, highly localized lexemes suffer aggressive replacement (e.g., "арідник" (aridnyk) regularized to "нарідник" (naridnyk), and "відтів" (vidtiv) replaced by "вітів" (vitiv)).

6.3 Wav2Vec2

6.3.1 Training Configuration and Results

We fine-tuned the Wav2Vec 2.0 model (XLSR-53 architecture, pre-trained for Ukrainian — `w2v-xls-r-uk`) on the “ukr-dialects-audio-dataset” dataset. Training was performed for 50 epochs on RTX 4090 48 GB.

The model was trained with a learning rate of $1e^{-4}$ using a linear scheduler with 1,000 warmup steps. The per-device batch size was set to 8 with gradient accumulation over 8 steps, yielding an effective total batch size of 64. CER was used as the primary optimization metric.

Training over 21,000 steps achieved stable convergence. Evaluated on the test set of 3,451 utterances, the model reached a final WER of 28.38% and CER of 9.93%.

6.3.2 Lexical and Morphological Analysis

Analysis of the recognition outputs reveals that Wav2Vec2 exhibits error tendency toward orthographic regularization. The model successfully transcribes common vocabulary and dialectal forms that are phonetically close to the literary norm, yet struggles at the morphological level.

The most pronounced failure mode remains the **suffix bottleneck**, whereby distinctive Hutsul endings are systematically replaced with their standard Ukrainian equivalents. The elevated CER (nearly 10%) is driven primarily by vowel substitutions in stems and inflections (e.g., frequent shifts between “і” (і) ↔ “и” (у) and “о” (о) ↔ “у” (u)). The most consistent pattern is the replacement of the dialectal reflexive postfix “-си” with the standard “-ся”, as well as the shift of the particle “си” to “це”/“се”. These substitutions occur in nearly every relevant context, contributing substantially to the overall CER. Representative examples are given below:

1. Postfix “-си” → “-ся”: REF “ни боїмоси” → HYP “не боїмося” (*ny boyimo-**sy** / ne boyimo-**sia***; gloss: “we are not afraid”)

2. Postfix “-си” → “-ся”: REF “вирваласи з хати” → HYP “вирвалася з хати” (*vyrvala-**sy** z khaty / vyrvala-**sia** z khaty*; gloss: “[she] tore [herself] out of the house”)

3. Particle “си” → “се”/“це”: REF “шош си було” → HYP “що ж се було” (*shosh **sy** bulo* / *shcho zh **se** bulo*; gloss: “something **was/this was**”)

4. Hard “-т” → soft “-ть”: REF “чорт и сам любит” → HYP “почорт і сам любить” (*chort y sam liubyt* / *pochort i sam liubyt’*; gloss: “the devil himself loves [it]”)

Dataset	Samples	WER	CER
Dido-Yvanchyk	842	25.80	5.98
YT-channel2	482	20.13	4.65
YT-channel1	103	22.36	5.44
NaUKMA students	1,806	33.88	15.27
Hutsulendia	217	40.31	15.55

Table 2: Cross-speaker evaluation of Wav2Vec2 fine-tuned on Hutsul data

Qualitative analysis of Wav2Vec 2.0 predictions reveals additional error patterns consistent with known Hutsul phonological properties. Dialect-specific vocabulary without standard equivalents (e.g., “лягри”, “кептар”, “сардак”) is frequently dropped or distorted. Word boundary segmentation errors — producing agglutinated forms absent from both standard and dialect — reflect the CTC architecture’s lack of explicit language model priors.

Cross-speaker results (Table 2) reveal a clear performance gradient: in-domain CER of 5.98% degrades to 15.27% on NaUKMA students and 15.55% on Hutsulendia radio broadcasts, consistent with the pattern observed across all model families. Among dictionary-verified dialect-specific lexemes, the model achieves high recall on frequent items but fails systematically on low-frequency, phonologically opaque forms, indicating **lexical regularisation** toward standard Ukrainian for rarer dialect-exclusive vocabulary.

7 Discussion

7.1 The Multi-Speaker Challenge

Our cross-speaker evaluation confirms a well-known but under-documented challenge in dialect ASR: models trained predominantly on one speaker’s data generalize poorly to other speakers of the same dialect. The 5× CER gap between in-domain (3.24%) and the most challenging out-of-domain evaluation (17.24%) underscores the need for diverse training data.

Several factors contribute to this gap:

- **Phonetic variation:** Individual speakers realize dialect phonemes differently, and the model may overfit to one speaker’s particular realizations.
- **Lexical diversity:** Different speakers draw on different subsets of the Hutsul lexicon, and some may code-switch more frequently with standard Ukrainian.
- **Recording conditions:** YouTube, radio, field recordings, and studio recordings differ in noise levels, microphone quality, and acoustic environments.
- **Speaking style:** Read speech (Dido Yvanchyk) differs substantially from spontaneous speech (Yaroslav, students) and broadcast speech (radio).

7.2 The RAG-Corrector Pipeline

The RAG-enhanced transcription pipeline addresses a fundamental bottleneck in scaling dialect ASR: the lack of reference transcriptions for new audio sources. By combining high-quality acoustic transcription (ElevenLabs) with dialect-aware linguistic correction (RAG + VuykoMistral), we can produce training data for speakers and sources that lack existing text.

However, the pipeline is not without limitations. The quality of corrections depends heavily on the coverage of the dialect knowledge base, and the model may struggle with dialectal features not represented in the retrieval corpus. Ongoing manual verification of a subset of corrected transcriptions will be essential for quality assurance.

7.3 Lexical Regularisation as a Structural Limitation

A pattern we observe across all three model families is that errors are not random: dialect-specific lexemes and morphological endings are systematically “corrected” to standard Ukrainian forms (e.g. the dialectal reflexive postfix - replaced by standard -, the particle replaced by /, and rare lexemes such as mapped to phonologically similar standard forms). This indicates that the decoder’s language prior, inherited from large-scale pre-training on standard text, dominates over the acoustic evidence whenever a dialect form has a phonologically close standard counterpart. Fine-tuning on a few

tens of hours of dialect data is not sufficient to overwrite this prior. We read this as a structural limitation rather than a data-quantity problem: a fully Hutsul-aware system likely requires either (i) a foundation acoustic model pre-trained on a large Hutsul-text language model, (ii) a constrained decoding scheme that explicitly biases toward attested dialect forms, or (iii) joint training with a dialect language-model loss. The current RAG-corrector mitigates the symptom at the text level after transcription but does not remove the underlying bias of the acoustic model.

7.4 Generational and Geographic Variation

The Hutsul dialect is not uniform: pronunciation, vocabulary and morphology vary across villages, between mountain and lowland communities, and between older and younger speakers. Our corpus partially reflects this. NaUKMA students contribute younger-generation speech that is closer to standard Ukrainian; field recordings from Verkhovyna contribute older-generation speakers whose dialect is conservative; the *Hutsulendia* broadcast subset and the YouTube channels span both demographics and several villages. We do not yet annotate village-level provenance for every speaker, which is a clear limitation for fine-grained sociolinguistic analysis. The cross-speaker results in Tables 3 and 4 can be read in this light: the largest CER values are concentrated in the subsets that mix several villages and generations (NaUKMA students, Hutsulendia), while subsets dominated by a single speaker or speech style yield substantially lower CER.

7.5 Implications for Low-Resource Dialect ASR

Our findings have broader implications for the dialect ASR community:

- **Single-speaker corpora are necessary but insufficient.** They provide a strong starting point for model development but do not guarantee cross-speaker generalisation.
- **Diverse data collection is essential.** Combining multiple sources (literary readings, field recordings, student speech, broadcasts) captures the full range of dialectal variation.
- **LLM-based post-processing can bootstrap transcription.** For dialects without written tradition, combining ASR with dialect-aware LLMs offers a scalable path to labelled data.

8 Released Resources

To support reproducibility and follow-up work, we publicly release:

- the expanded `ukr-dialects-audio-dataset` multi-speaker Hutsul corpus (40 speakers, 60.63 hours), with train/validation/test splits and per-utterance metadata;
- the fine-tuned models at `KSE-RESEARCH-Group`, the four fine-tuned Whisper variants, and the Wav2Vec2 (XLS-R) checkpoint;
- the RAG-corrector pipeline code (chunking, FAISS index, fuzzy dictionary lookup, prompt assembly), together with the Hutsul–Ukrainian dictionary used for retrieval; and
- the training and evaluation scripts used to reproduce all reported numbers.

All artefacts are hosted on Hugging Face under the `KSE-RESEARCH-Group` organisation and on the project GitHub repository.[1]

9 Conclusion

We present a substantial expansion of Hutsul dialect ASR resources, growing from a single-speaker literary corpus to a multi-speaker dataset comprising 40 speakers and 60.63 hours of audio from diverse sources. We introduce a RAG-enhanced transcription pipeline that enables scalable creation of training data for audio without existing reference text. Our cross-speaker evaluation reveals both the promise and limitations of current approaches: fine-tuned models achieve strong in-domain performance (3.24% CER) but face significant degradation on out-of-domain speakers (up to 17.24% CER).

These results motivate several directions for future work: multi-speaker training with balanced speaker representation, speaker adaptation techniques, improved RAG-corrector pipelines with expanded dialect knowledge bases, and exploration of LM rescoring for CTC-based dialect ASR. All

[1] Hugging Face: https://huggingface.co/KSE-RESEARCH-Group. Code: https://github.com/KSE-RESEARCH-Group.

data, models, and code are publicly released to support continued research on Ukrainian dialect speech technologies.

Limitations

This work has several limitations. First, our cross-speaker evaluation uses a model primarily trained on single-speaker data; results from a model trained on the full multi-speaker corpus may differ substantially. Second, the RAG-corrector pipeline has not yet been comprehensively evaluated with human judgments—we plan to conduct this evaluation as annotation is completed. Third, the data quantities for some speakers (e.g., YT-channel1 with 103 samples) are small, making evaluation estimates noisy. Fourth, we do not yet evaluate the impact of including RAG-corrected transcriptions in training data. Finally, our evaluation is limited to WER and CER; perceptual quality assessments and downstream task evaluations would provide a more complete picture.

Acknowledgments

We thank the YouTube content creators who generously granted permission to use their recordings for research purposes (Larysa Irodenko and Ivanna Stefiuk), the linguistics students at the National University of Kyiv-Mohyla Academy (NaUKMA) who participated in recording sessions, and Yaroslav Zelenchuk for gathering and providing native-speaker recordings from the Verkhovyna region. We also thank the anonymous reviewers of UNLP 2026 for constructive feedback that improved the camera-ready version of this paper.

Use of generative AI. Large language models (specifically OpenAI GPT-4o and Anthropic Claude) were used (i) inside the RAG-corrector pipeline as described in Section 4 (this is part of the methodological contribution), and (ii) to assist with copy-editing of the manuscript: smoothing English prose, harmonising terminology, and checking for typographical errors. All technical content, experimental design, results, and interpretations are the responsibility of the authors, who reviewed and edited every passage produced or revised with LLM assistance.

References

Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014. Advances in dialectal Arabic speech recognition: A study using Twitter to improve Egyptian ASR. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*.

arampacha. 2024. arampacha/whisper-large-uk-2: Whisper-Large fine-tuned for Ukrainian speech recognition. Hugging Face model. https://huggingface.co/arampacha/whisper-large-uk-2.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222. European Language Resources Association. Cited version is Corpus 11.0 (2022 release).

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL, Long Papers)*.

ElevenLabs. 2024. Elevenlabs speech-to-text API documentation. https://elevenlabs.io/docs/api/speech-to-text.

Tahir Javed, Janki Nawale, Eldho Joshi, Kaushal Bhogale, Sumanth Doddapaneni, Anoop Kunchukuttan, Pushpak Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2024. LAHAJA: A robust multi-accent benchmark for evaluating Hindi ASR systems. *Preprint*, arXiv:2408.11440.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A practitioner's guide to building ASR models for low-resource languages: A case study on Scottish Gaelic. In *Proc. Interspeech 2025*.

Roman Kyslyi, Yulia Maksymiuk, and Ihor Pysmennyy. 2025. Vuyko Mistral: Adapting LLMs for low-resource dialectal translation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP)*. Association for Computational Linguistics.

Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, and Zakhar Guzii. 2026. Building ASR resources for the Hutsul dialect of Ukrainian. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Yurii Paniv. 2023. Ukrainian-TTS: An open text-to-speech system for Ukrainian. GitHub repository.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 1 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Thennal D K, Jesin James, Deepa P. Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

A Detailed Evaluation Tables

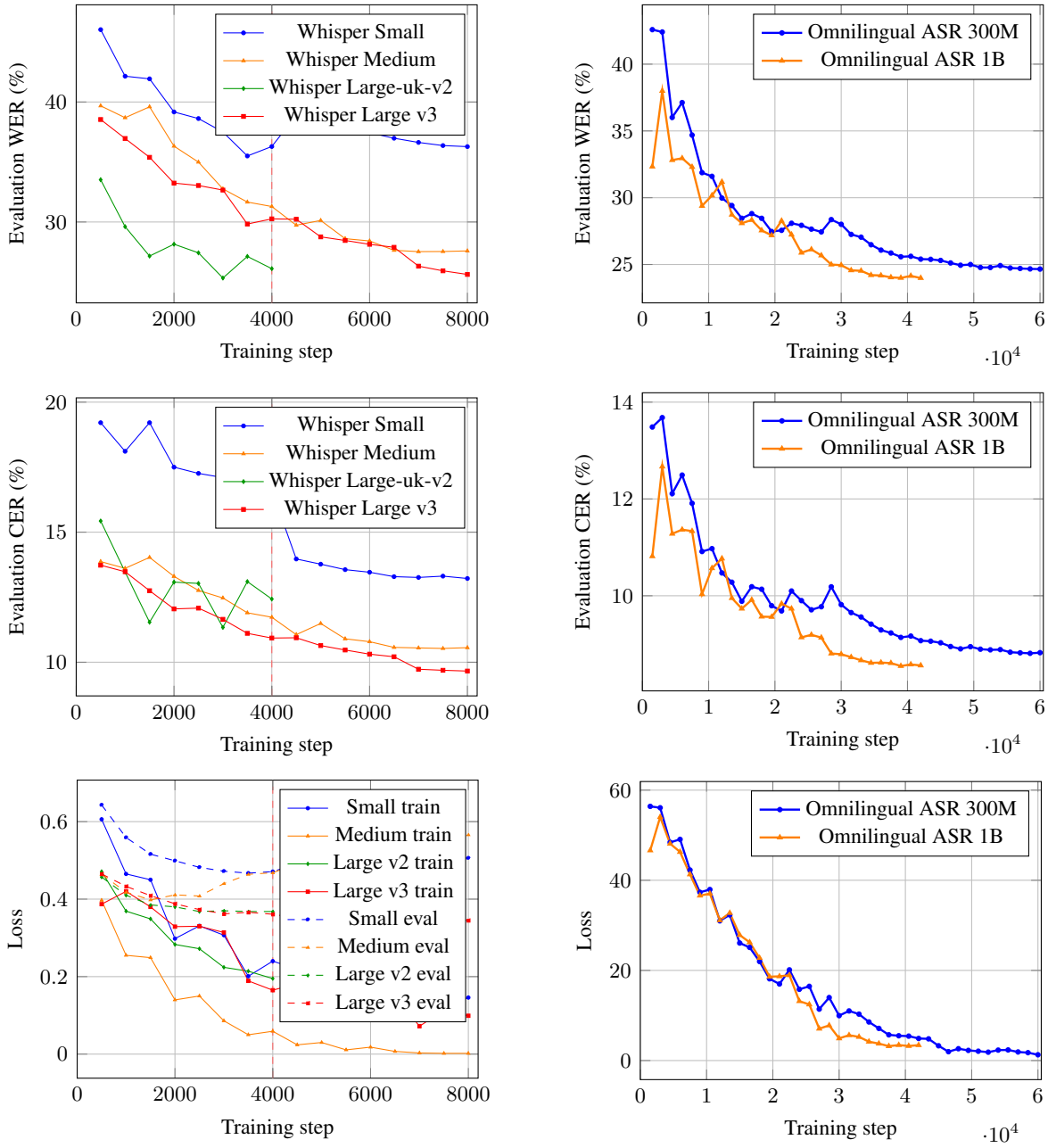
Dataset	Model	Samples	WER	CER
Dido-Yvanchyk	Omnilingual ASR 300M	842	**13.93**	**3.19**
	Omnilingual ASR 1B	842	14.07	3.24
YT-channel2	Omnilingual ASR 300M	483	21.94	5.75
	Omnilingual ASR 1B	483	**20.37**	**5.33**
YT-channel1	Omnilingual ASR 300M	103	**23.41**	6.38
	Omnilingual ASR 1B	103	24.10	**6.19**
NaUKMA students	Omnilingual ASR 300M	1,806	**36.79**	17.21
	Omnilingual ASR 1B	1,806	**36.79**	**16.85**
Hutsulendia (radio)	Omnilingual ASR 300M	217	43.90	**17.13**
	Omnilingual ASR 1B	217	**43.63**	17.24
ukr-dialects-audio-dataset	Omnilingual ASR 300M	3,451	28.99	11.72
	Omnilingual ASR 1B	3,451	**28.76**	**11.49**

Table 3: Cross-speaker evaluation of OmniASR-CTC models fine-tuned on Hutsul data. WER/CER reported as percentages. Bold marks the best result per dataset per metric. Lower is better.

Dataset	Model	Samples	WER	CER
Dido-Yvanchyk	Whisper Small	842	29.88	7.57
	Whisper Medium	842	**19.83**	5.19
	Whisper Large-uk-v2	842	23.50	6.05
	Whisper Large v3	842	19.92	**5.13**
YT-channel2	Whisper Small	483	28.35	7.65
	Whisper Medium	483	19.49	5.07
	Whisper Large-uk-v2	483	19.86	5.22
	Whisper Large v3	483	**16.37**	**4.26**
YT-channel1	Whisper Small	103	34.64	9.56
	Whisper Medium	103	26.82	7.81
	Whisper Large-uk-v2	103	25.54	7.29
	Whisper Large v3	103	**23.18**	**6.23**
NaUKMA students	Whisper Small	1,806	41.86	18.76
	Whisper Medium	1,806	35.00	16.22
	Whisper Large-uk-v2	1,806	37.54	17.08
	Whisper Large v3	1,806	**32.42**	**15.23**
Hutsulendia (radio)	Whisper Small	217	45.85	19.28
	Whisper Medium	217	35.83	14.58
	Whisper Large-uk-v2	217	36.50	15.00
	Whisper Large v3	217	**32.01**	**12.83**
ukr-dialects-audio-dataset	Whisper Small	3,451	35.41	12.54
	Whisper Medium	3,451	26.96	9.99
	Whisper Large-uk-v2	3,451	29.37	10.67
	Whisper Large v3	3,451	**25.18**	**9.32**

Table 4: Cross-speaker evaluation of the Whisper model family fine-tuned on Hutsul data. WER/CER reported as percentages. Bold marks the best result per dataset per metric. Lower is better.

B Training Dynamics



Whisper models: WER, CER, and train/eval loss.

Omnilingual ASR models: WER, CER, and loss.

Figure 2: Training dynamics in a six-plot grid. Left column: Whisper Small/Medium/Large-uk-v2/Large-v3 curves on `ukr-dialects-audio-dataset` (top: WER, middle: CER, bottom: train / eval loss). The red dashed vertical line at step 4000 marks where Whisper Small resumed Phase 2 training. Right column: Omnilingual ASR 300M vs 1B curves (top: WER, middle: CER, bottom: loss). "Large-uk-v2" denotes the Ukrainian-adapted checkpoint `arampacha/whisper-large-uk-2`.

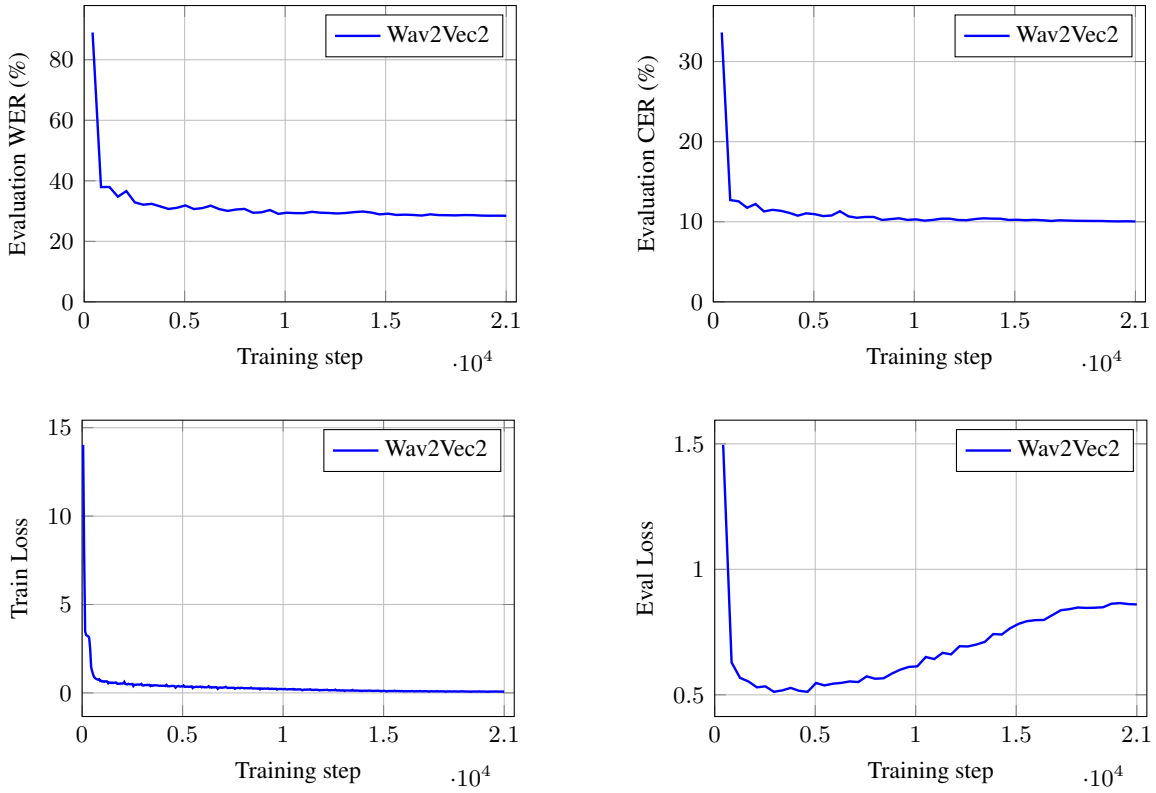


Figure 3: Training dynamics of Wav2Vec2 fine-tuned on the Hutsul dialect dataset over 50 epochs (21,000 steps). Top row: evaluation WER (%) and CER (%). Bottom row: train loss and eval loss. Best checkpoint at step 21,000 (WER = 28.45%, CER = 10.04%). Note: eval loss diverges monotonically from step 2,520 onward while WER/CER continue improving, consistent with the behaviour observed in Whisper models on this task.

C Worst-Case Transcription Examples

Table 5 presents one representative example per model for the two error extremes: the highest-WER utterance (**Worst WER**) and the lowest-CER utterance from the 50-worst list (**Lowest CER**). **REF** = Hutsul dialect reference; **HYP** = model hypothesis.

Model	Type	Dataset	WER	CER	REF	HYP
Whisper Small	Worst WER	Hutsulendia	1.00	1.00	А не і співала, гуляла, ішла, куди мене вела.	А не лізть павалом, ну я гала це і йшла мкуда ми виділи.
	Lowest CER	Dido-Yv.	0.60	0.04	Прото — пустий пусте робит.	Прото, пусти й пусте робит.
Whisper Medium	Worst WER	Hutsulendia	0.80	1.00	А п'ятниця до Пречистої Діви.	А п'єкні це до причищення діли.
	Lowest CER	Dido-Yv.	0.40	0.00	Протерав чоло, так єкби си лише шо з сну прошумав.	Протерав чоло, так єк би си лишешо з сну прошумав.
Whisper Large-uk-v2	Worst WER	Hutsulendia	1.00	0.46	А не і співала, гуляла, ішла, куди мене вела.	А не відставала, бо я і йшла, куда б не виділа.
	Lowest CER	Dido-Yv.	0.43	0.00	Єк флуд пер за дичінов.	Єк флуд перза дичінов.
Whisper Large-v3	Worst WER	Hutsulendia	1.00	0.94	Бути написи, бо я знаю, що- Є ще написи.	- бути, напис робити,- Я знаю, що коли- Ходішь, там є ще знати.
	Lowest CER	Dido-Yv.	0.43	0.00	Зробив навіть и стів новомодний у хаті.	Зробив навіть истівново модний у хаті.
Wav2Vec2	Worst WER	Hutsulendia	1.00	1.00	бо є таке що людина десь погано став	поє таке об одіна рдиєся по панастатаацятттт
	Lowest CER	YT-channel1	0.20	0.00	русалка то та сама лісна котра лиш у воді жила	русалка тота сама лісна котра лиш у воді жила
OmniASR 300M	Worst WER	Hutsulendia	1.00	1.00	рано бо це був вечір	т зра рамон бо це був запобеч вни віів
	Lowest CER	Dido-Yv.	0.80	0.00	то прото були нипрості люде	топрото були ни прості люде
OmniASR 1B	Worst WER	Hutsulendia	1.00	0.94	рано бо це був вечір	ра рамо бо це був побубечи вахоув
	Lowest CER	Dido-Yv.	0.29	0.00	тимунь він цілу ніч мавси на осторозі	тимунь він цілу ніч мавси наосторозі

Table 5: Representative worst-case transcription examples across all models. For each model: Worst WER = utterance with highest WER; Lowest CER = lowest CER from the 50-worst list.