Building ASR Resources for the Hutsul Dialect of Ukrainian

Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, Zakhar Guzii

Kyiv School of Economics
{rkyslyi,aorlovskyi,pkhomenko, bhnyshchenko,zguzii}@kse.org.ua

Abstract

Dialectal speech remains largely underexplored in Automatic Speech Recognition (ASR) research, particularly for Slavic languages. While Ukrainian ASR systems have rapidly improved in recent years with the adoption of Whisper, XLS-R, and Wav2Vec-based models, performance on dialectal variants remains unknown and often significantly degraded. In this work, we present the first dedicated effort to build ASR resources for the Hutsul dialect of Ukrainian. We develop a data preparation and segmentation pipeline, evaluate multiple forced alignment strategies, and benchmark state-of-the-art ASR models under zero-shot and fine-tuned conditions. We evaluate results using WER and CER demonstrating that large multilingual ASR models struggle with dialectal speech, while lightweight fine-tuning produces substantial improvements. All scripts, alignment tools, and training recipes are made publicly available to support future research on Ukrainian dialect speech.

1 Introduction

Ukrainian NLP has made significant progress in recent years, with new language models, ASR systems, and datasets becoming publicly available (Sereda, 2024). However, the majority of existing work targets standard Ukrainian as used in broadcast media or academic corpora, leaving dialectal variation largely unaddressed (Zhong et al., 2024). Yet dialects remain a central component of Ukrainian linguistic identity, especially in mountainous western regions such as Carpathian mountains, where phonetic and lexical differences from standard Ukrainian are quite huge[1].

Dialectal speech in general introduces challenges for ASR: vowel reduction, consonant softening, archaic lexical forms, code-switching with other languages (Romanian, Polish, German, etc.) and highly variable pronunciation shaped by geography and speaker generation (Michelsanti and et al., 2019). Zero-shot multilingual models can transcribe such speech, but with significantly reduced accuracy (Adams and et al., 2020). At the same time, collecting labeled dialect data is difficult due to speaker scarcity, limited dialectal written tradition, and available recordings being long-form oral narratives requiring segmentation and alignment (Klejch and et al., 2025).

To address these gaps, we present the first systematic pipeline for one of the dialects (Hutsul[2]) ASR development, centered around a curated speech corpus derived from the recordings of "Dido Yvanchyk"[3] - a unique novel written completely using dialect.

Our choice based first of all on availability of both recorded reading of the novel as well as a textual representation of it.

This paper introduces the dataset, describes alignment methodology, reports initial baselines, and outlines future work including dataset expansion, cross-dialect generalization, and LLM-powered transcript enhancement.

Our contributions in this work are:

1. Building a publicly available pipeline for segmentation, forced alignment, and dataset preparation for Hutsul dialect speech.

2. Evaluation of state of the art ASR models in zero-shot and fine-tuned settings using WER/CER metrics.

3. Release scripts and data, enabling reproducible training and evaluation for future research.

[1] https://en.wikipedia.org/wiki/Ukrainian_dialects

[2] https://en.wikipedia.org/wiki/Hutsuls

[3] https://shron1.chtyvo.org.ua/Shekeryk-Donykiv_Petro/Dido_Yvanchik.pdf

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 186–195

2 Related Work

Research on ASR for low-resource languages and dialects has grown in Arabic, Hindi, Italian, and Turkic languages (Ali and et al., 2014; Kumar and et al., 2025), where phonetic variation and lexical differences significantly impact recognition. Recent regional work also introduced RoDia, a Romanian dialect speech dataset for dialect identification (Rotaru et al., 2024), highlighting increasing interest in dialect-focused speech resources.

In Ukrainian, existing work focuses primarily on standard language, including XLS-R and Wav2Vec2-UA models for broadcast and conversational speech (Paniv, 2023). Community projects have fine-tuned Whisper for Ukrainian (Shas, 2024), but dialect performance remains undocumented. No prior work has addressed Hutsul ASR, making this study the first benchmark of its kind.

Forced alignment for speech corpora is typically performed using Montreal Forced Aligner (McAuliffe et al., 2017) or Aeneas (ReadBeyond, 2015), while WhisperX (Shi and et al., 2023) recently demonstrated strong alignment performance for multilingual speech. Our work compares alignment approaches in a dialect context where standard lexicons may be insufficient.

3 Hutsul Dialect Speech Corpus

The used dataset is primarily based on recordings of a native Hutsul speaker reading "Dido Yvanchyk" novel, the most extensive and culturally important written sources of Hutsul dialect. The novel represents the largest continuous literary record of Hutsul speech, featuring authentic lexical, morphological, and phonetic elements that are rarely present in standard Ukrainian corpora. We use publicly available audio recordings of oral readings of the text, sourced from a native speaker and accessible on YouTube.[4] The dataset is released on Hugging Face.[5]

Hutsul dialect is quite different comparing to standatd Ukrainian due to its geographical location. Here are some of the linguistic Characteristics of Hutsul dialect:

- *Phonetics*: vowel transformations, such as /je/ instead of /a/ or /ja/ (e.g., yak → yek, yahoda → yehoda).

[4] https://www.youtube.com/@didoyvanchik7322

[5] https://huggingface.co/datasets/KSE-RESEARCH-Group/Dido-Yvanchyk-Audio-Dataset-v2

Property	Value
Total duration	19h 27m
Total samples	8,412
Average segment duration	8.35 s
Median segment duration	6.68 s
Duration range	0.11-78.84 s
Total words	162,204
Unique vocabulary	32,247 words
Average words per segment	19.4
Average characters per segment	108.8
Audio sample rate	16 kHz
Dialect marker coverage	72.4%
Total dialect markers	16,874

Table 1: Dido-Yvanchyk Hutsul Dialect Dataset Summary

- *Morphology*: unique case endings `-yed`, `-si` and preserved dual forms of a word `apples`, e.g., `yablutsi` instead of plural `yabluka`.
- *Lexicon*: Romanian, Polish and German borrowings such as `brynza` (cheese) and `spazieruvaty` (go for a walk, from German `spazieren gehen`).[6]

The audio is rich in dialect vocabulary, archaic expressions, and non-standard pronunciation, making it a valuable material for building dialectal ASR systems. Segmentation and alignment were performed on long-form recordings to create paired audio–text samples suitable for model training. All recordings are read by a single native Hutsul speaker, however, careful train–dev–test splitting and augmentation were applied to mitigate speaker memorization effects.

The current release serves as a baseline version of the corpus. In future work, we plan to expand the dataset with additional speakers, spontaneous speech, and regional variation, enabling multi-speaker modeling, dialect classification, and deeper linguistic analysis.

General dataset statistics are summarized in Table 1, with a detailed description provided in Appendix A.

4 Forced Alignment Pipeline

Recordings were normalized and resampled to 16kHz. Before the text–transcript alignment, the

[6] https://en.wikipedia.org/wiki/Eastern_Romance_influence_on_Slavic_languages

audio recordings are automatically transcribed with word-level time stamps. We evaluate two speech-to-text pipelines: WhisperX (Shi and et al., 2023) and the ElevenLabs (ElevenLabs, 2024) speech recognition system.

WhisperX extends Whisper (Radford et al., 2022) with forced alignment, enabling word-level timing extraction from neural acoustic models, while ElevenLabs provides native word-level timestamps as part of its transcription output. In practice, ElevenLabs yields more accurate and stable word boundaries (ElevenLabs, 2024), particularly in conversational and expressive speech, and is therefore used as the primary source of word-level timing information for subsequent alignment and segmentation steps.

We perform sentence–level alignment between ground-truth text and automatic speech transcription using a fuzzy string-matching strategy (Source, 2011). Both the transcript generated by the ASR system and the reference text are first normalized through lowercasing, punctuation removal, and whitespace standardization to reduce superficial mismatches.

For each ground-truth sentence, candidate hypothesis segments are searched within a sliding temporal window, and similarity is computed using token-based fuzzy matching (RapidFuzz (Bachmann, 2020)). The highest-scoring segment above a predefined threshold is selected as the alignment, enabling robust matching even in the presence of transcription errors, paraphrasing, or minor omissions (Gao and et al., 2025; Chen and et al., 2025; Abdjul and et al., 2025). The process is executed in batches while maintaining sentence and segment indices, ensuring sequential consistency and scalability to long recordings.

The selected ASR system produces a continuous transcript with word-level timestamps, which serves as the temporal backbone for subsequent sentence matching. The resulting aligned segments preserve the original temporal ordering of the recording and enable consistent downstream processing, including sentence-level segmentation and corpus construction for dialectal speech analysis.

Pipeline overview:

1. Transcribe raw audio using WhisperX or ElevenLabs STT.
2. Normalize reference text (lowercase, remove punctuation, unify spacing).
3. Align reference sentences to ASR output using RapidFuzz similarity search.
4. Export word- or sentence-level timestamps in WebVTT/JSON format.
5. Filter low-confidence matches to construct clean training segments.

Initial observations indicate WhisperX yields the most stable results, while Montreal Forced Aligner (McAuliffe et al., 2017) requires dialect lexicon adaptation.

5 ASR Models and Training

We evaluate four ASR model families representing different architectural paradigms and multilingual capabilities:

- **OmniASR** – Meta's new omnilingual speech model designed for cross-language generalization across hundreds of languages (team et al., 2025).
- **Wav2Vec2-XLSR-300M-UA** – a self-supervised multilingual wav2vec2 model with a Ukrainian CTC head, commonly used in Ukrainian ASR applications (Baevski et al., 2020; Babu and et al., 2021).
- **Whisper-large-v3** – a transformer-based encoder–decoder ASR model trained on 680K+ hours of weakly supervised multilingual data (Radford et al., 2022), widely adopted as a strong zero-shot baseline.
- **Wav2Vec2-BERT-UK-v2.1** – a hybrid SSL+MLM architecture integrating wav2vec2 acoustic features with BERT-style masked language modeling (Smoliakow and et al., 2024).

We select these models to cover most popular models in current ASR research and score them for the dialectal recognitions task (Radford et al., 2022; Baevski et al., 2020; Babu and et al., 2021; Smoliakow and et al., 2024; team et al., 2025).

Whisper serves as a strong zero-shot baseline (Radford et al., 2022), Wav2Vec2-XLSR and Wav2Vec2-BERT represent Ukrainian-centric self-supervised CTC approaches (Baevski et al., 2020; Babu and et al., 2021; Smoliakow and et al., 2024), enabling us to assess how well standard-Ukrainian

systems adapt to dialects, and OmniASR allows us to test whether a unified omnilingual speech encoder can generalize to dialect without prior exposure (team et al., 2025).

This spectrum enables comparison across (i) multilingual vs. Ukrainian-only pre-training, (ii) encoder–decoder vs. CTC decoding, and (iii) zero-shot vs. fine-tuned dialect adaptation (Radford et al., 2022; Baevski et al., 2020).

All models are fine-tuned on the aligned Hutsul dataset using an 80/10/10 train–dev–test split. For Whisper, we apply parameter-efficient LoRA fine-tuning (Hu et al., 2021).

For Wav2Vec2-BERT, lower convolutional and transformer layers are frozen to prevent catastrophic forgetting, while adapter modules and the CTC head remain trainable (Smoliakow and et al., 2024).

OmniASR models are fine-tuned using Meta's official tri-stage learning rate schedule (team et al., 2025).

Also, given the limited amount of dialectal speech data, we use on-the-fly augmentation during training to improve robustness (Chen and et al., 2025). The pipeline includes Gaussian noise injection, pitch shifting, speed perturbation (0.8–1.2×), gain modulation, and time/frequency masking (Nguyen and et al., 2023). Such augmentation is essential for preventing overfitting and encouraged more stable convergence.

All training runs use the AdamW optimizer with FP16 mixed precision (Baevski et al., 2020). Whisper and Wav2Vec2 models employ a linear warm-up schedule followed by linear decay (Radford et al., 2022), while OmniASR follows a tri-stage scheduler (team et al., 2025). Checkpoints are evaluated every 500–1,000 steps, and the best checkpoint is selected based on development CER. Hyperparameters and training configurations for each model are summarized in next section.

6 Experimental Results

For the training we used mixed-precision training mode, with CER (character error rate) as the metric for selection of the best model. The issue between selecting WER and CER lies down to the nuances of the training ASR for the dialect task, since the changes are mostly in characters(Coll and et al., 2023; Thennal D K, 2024).

Ukrainian is a highly inflected Slavic language with:

- Seven grammatical cases
- Three genders with agreement across adjectives/participles
- Complex verb conjugation
- Extensive prefixation and suffixation

And other morphological complex structures and nuances(Pugh and Press, 2005; Sussex and Cubberley, 2006).

All models were evaluated on the same test split using WER and CER metrics.

Overall, fine-tuning consistently improved recognition quality on Hutsul speech across all architectures, with the best systems reaching sub-3% CER.

6.1 OmniASR

We fine-tuned two OmniASR CTC models released by Meta (300M and 1B parameters).[7] Training was performed using the official (team et al., 2025) tutorial on the "Dido Yvanchyk" dataset.

Training configuration. Both models were trained with a learning rate $5e^{-5}$ using a tri-stage scheduler. The 300M model was trained on an RTX 5070 Ti (16GB) for 48k steps, and 1B model was trained on an RTX 4090 (48GB) for 36k steps. We used a per-device batch size of 8 with gradient accumulation of 4, and WER served as the primary optimization metric.

Model	WER_b	CER_b	WER_a	CER_a
300M	76.78	37.08	**13.82**	**2.97**
1B	80.09	51.24	**13.09**	**2.75**

Table 2: OmniASR fine-tuning results on Hutsul dialect (in %). WER_b/CER_b: before fine-tuning; WER_a/CER_a: after fine-tuning.

After fine-tuning, both OmniASR reduced WER from 76–80% down to around 13%, indicating that OmniASR benefits strongly from dialect adaptation.

6.2 Wav2Vec2-XLSR-300M-UA

We fine-tuned the Wav2Vec2-XLSR-300M-UA model(Conneau and et al., 2020; Smoliakow and et al., 2023) on the aligned Hutsul corpus. All audio was resampled to 16 kHz, normalized, and cleaned

[7] https://huggingface.co/KSE-RESEARCH-Group/omniASR-CTC-300M-uk-dido-tuned-v2 https://huggingface.co/KSE-RESEARCH-Group/omniASR-CTC-1B-uk-dido-tuned-v2

using a regular-expression based text preprocessor. A Ukrainian CTC vocabulary was constructed to include all Cyrillic characters and dialectal orthographic forms.

Training setup. The model was trained on an RTX 4090 with an effective batch size of 64 (per-device BS=8, gradient accumulation=8), using FP16 mixed precision, AdamW optimizer, a peak learning rate of 1×10^{-4}, and a linear scheduler with 1,000 warmup steps. Training ran for 50 epochs (5,000 steps total).

Results. The best checkpoint was reached at step 4,700 with validation WER=12.99% / CER=2.53%, and the final test performance achieved **WER=13.61% / CER=2.43%**.

Fine-tuning reduced CER to below 3%, placing Wav2Vec2-XLSR performance close to OmniASR and confirming that Ukrainian-centric SSL models adapt well to dialectal speech.

Model	Base checkpoint	Max steps	Per-dev BS	Grad acc.	Eff. BS	Best-by
Small	`openai/whisper-small`	8000	4	4	16	CER
Medium	`openai/whisper-medium`	8000	8	4	32	CER
Large-v3	`openai/whisper-large-v3`	8000	4	4	16	CER
Large-v2 (UK)	`arampacha/whisper-large-uk-2`	4000	4	4	16	CER

Table 3: Fine-tuning configuration for Whisper models.

6.3 Whisper Family

We have fine-tuned four Whisper variants: `whisper-small`, `whisper-medium`, `whisper-large-v3`, and `arampacha/whisper-large-uk-2`, the latter already adapted to Ukrainian using Common Voice 11.0(Ardila et al., 2022). We release our fine-tuned checkpoints on Hugging Face.[8] Training was performed in mixed-precision mode, using CER as the selection criterion, as character-level variation better captures dialectal orthography and subword differences. The inflectional nature of Ukrainian further motivates character-aware adaptation for dialect modeling.

Training setup. All Whisper models were trained for up to 8,000 steps (4,000 for the Common Voice–adapted checkpoint). Batch size, gradient accumulation, and max steps per model are summarized in Table 3. Learning rate schedules and optimization configurations were kept consistent across models to enable fair comparison.

Results. Final WER/CER performance is shown in Table 4. Both `whisper-large-v3` and the Ukrainian-adapted `whisper-large-v2` achieved sub-4% CER and approximately 13% WER on the test set, outperforming the medium and small variants.

Whisper fine-tuning consistently improved performance across all model sizes, with the large Ukrainian-adapted checkpoint achieving the best results.

6.4 Wav2Vec2-BERT

We additionally evaluated the Wav2Vec2-BERT architecture (Hsu et al., 2021; Baevski et al., 2020) using a Ukrainian-pretrained checkpoint [9]. The model was fine-tuned with adapters enabled, while freezing the feature encoder and BERT backbone to prevent catastrophic forgetting. Parameter-efficient training was selected following prior work on adapter-based optimization (Houlsby et al., 2019; Hu et al., 2021). A stronger augmentation pipeline was applied to increase robustness given the limited volume of dialectal speech data.

Results. Adapter-based fine-tuning achieved a test performance of **WER=18.24% / CER=3.47%**. In contrast, LoRA-based training converged less reliably, yielding **WER=43.08% / CER=9.93%** despite similar optimization settings. This suggests that adapter-based optimization is more suitable for this model family in low-resource dialect scenarios.

6.5 Overall Results Summary

Across all model families, fine-tuning led to large improvements in recognition quality on Hutsul speech. The strongest systems—OmniASR 1B, Whisper-Large-v3, Whisper-Large-v2-UK and

[8] https://huggingface.co/KSE-RESEARCH-Group/whisper-small-dido-yvanchyk-v2 https://huggingface.co/KSE-RESEARCH-Group/whisper-medium-dido-yvanchyk-v2 https://huggingface.co/KSE-RESEARCH-Group/whisper-large-v3-dido-yvanchyk-v2 https://huggingface.co/KSE-RESEARCH-Group/arampacha-whisper-large-v2-dido-yvanchyk-v2

[9] https://huggingface.co/Yehor/w2v-bert-uk-v2.1

Model	Val CER↓	Val WER↓	Test CER↓	Test WER↓	Best ckpt
Small	4.71	17.60	4.72	17.84	6500
Medium	4.06	13.80	3.96	14.61	7000
Large-v3	3.73	12.40	3.90	13.20	7000
Large-v2 (UK)	**3.73**	12.60	**3.69**	**13.03**	4000

Table 4: Whisper ASR performance after fine-tuning. Lower is better.

Metric	**Original**	**Finetuned**	**Improvement**
Word Error Rate (WER)	60.79%	14.32%	↓ 76.4%
Character Error Rate (CER)	17.86%	5.14%	↓ 71.2%
Perfect Transcriptions	2 (0.2%)	217 (25.8%)	↑ 108×
Samples with Errors	840 (99.8%)	625 (74.2%)	↓ 25.6%

Table 5: Overall Performance Comparison: Original vs Finetuned Whisper-Large-V3

Wav2Vec2-XLSR—achieved **CER < 3%** on the test set, showing that high-quality ASR for a low-resource dialect can be obtained using a relatively small aligned corpus.

Figure 1 and Figure 2 visualize convergence for Whisper models, where larger architectures show a faster drop in CER and more stable late-stage training.

Models and resources. All trained models and datasets described are publicly released on Hugging Face: https://huggingface.co/KSE-RESEARCH-Group.

6.6 Qualitative Example Analysis

To illustrate model behavior beyond aggregate metrics, we provide qualitative recognition examples for OmniASR models, including Cyrillic text with transliteration. These examples highlight typical recognition patterns, including character alignment with occasional lexical substitutions and vowel variation.

This shows that most errors relate to dialect-specific morphology and orthographic variation rather than acoustic confusion. The complete samples are provided in Appendix B.

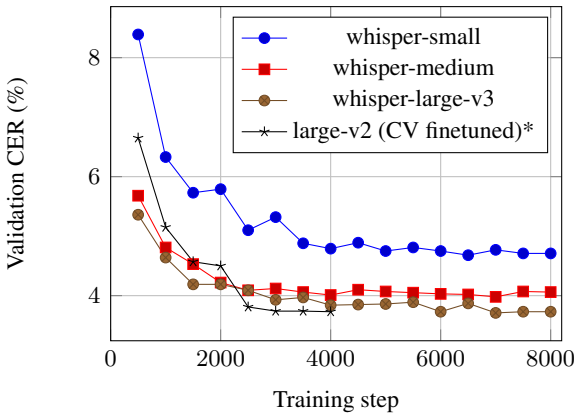

Figure 1: Validation CER over fine-tuning steps (evaluated every 500 steps). * – arampacha/whisper-large-v2, previously fine-tuned on Common Voice 11.0.

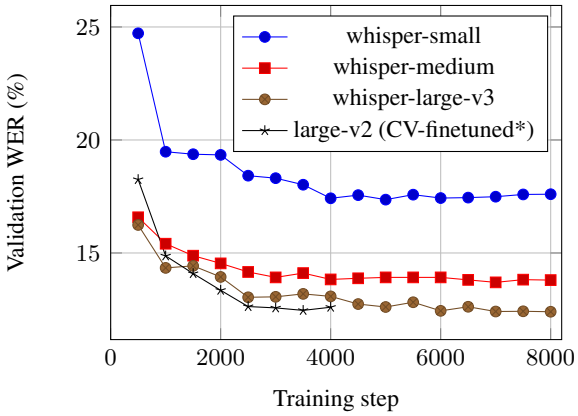

Figure 2: Validation WER over fine-tuning steps (evaluated every 500 steps). * – arampacha/whisper-large-v2, previously fine-tuned on Common Voice 11.0.

7 Discussion

Our results show that modern ASR models can adapt well to Ukrainian dialectal speech when fine-tuned on even relatively small aligned data. All tested model families improved substantially compared to zero-shot recognition, and (Omni-ASR 1B, Whisper-Large-v3, Whisper-Large-v2-UK, Wav2Vec2-XLSR) reached CER below 3%. This suggests that large multilingual and Ukrainian-centric self-supervised models can be effectively used for dialect ASR with limited supervision (Baevski et al., 2020; Shas, 2024; team et al., 2025).

During qualitative inspection, we observed that most remaining errors come from dialect-specific morphology, rare lexemes, vowel reduction, and non-standard spelling. Small changes in endings or palatalization often result in substitutions. This indicates that future work could explore phoneme-aware decoding, lexicon extension, or LM rescoring to reduce errors caused by dialectal variation, as also noted for other Slavic speech systems (Sussex and Cubberley, 2006).

Although this work focuses on the Hutsul dialect, the methodology is general and can be applied to other Ukrainian regional varieties. The alignment pipeline, training scripts and evaluation setup can serve as a template for expanding research to other dialects. In future, combining several dialect corpora may allow dialect classification, cross-dialect transfer, and speech-style robustness studies.

The released dataset and pipeline represent a first step toward more complete Ukrainian dialect ASR resources. As new recordings are collected, it will be important to ensure balanced speaker representation, informed consent, and fair coverage across regions.

8 Conclusion

We present the first speech corpus with aligned text for the Hutsul dialect and evaluated four modern ASR model families on this data. Fine-tuned Whisper, OmniASR and Wav2Vec2-XLSR models achieved strong accuracy with CER below 3%, showing that high-quality recognition of Ukrainian dialectal speech is possible even with a single-speaker corpus when alignment and augmentation are applied effectively. Adapter-based training also improved Wav2Vec2-BERT, although it remained behind the top systems.

While our experiments focus on Hutsul, the same training pipeline can be applied to other dialects, enabling work on multi-dialect modeling, dialect identification and cross-dialect transfer. In future we plan to extend the corpus with more speakers, spontaneous speech and additional regions, as well as explore phoneme-level decoding and LM rescoring for better handling of dialect-specific forms.

We release the dataset, code and splits to facilitate reproducibility and further research on Ukrainian dialect ASR. We believe this work provides a starting point for building larger dialect resources and contributes toward speech technologies for low-resource Slavic varieties.

Limitations

This work represents an initial step toward ASR for Ukrainian dialects and carries a lot of limitations.

First, the current corpus is based on recordings from a small number of speakers, from a single Hutsul region, which may restrict dialectal and acoustic diversity. Broader demographic (age, gender, speaking style, recording conditions) will be necessary to ensure robust generalization.

Second, our training pipeline relies on automatic forced alignment. Although it is effective, alignment errors sometimes occurs in the fine-tuning data and may influence on achievable performance.

Third, we evaluate only end-to-end CTC and encoder–decoder architectures, leaving LM-rescoring, shallow fusion, and hybrid ASR systems for future study.

Finally, evaluation is limited to WER and CER, without semantic or intelligibility assessments, which would provide a more complete measure of transcription quality.

As such, results should be interpreted as a strong baseline rather than a fully comprehensive solution for Hutsul or broader Ukrainian dialect recognition.

Ethics

We fully acknowledge the ACL Ethics Policy and commit to responsible research practice, including careful consideration of consent, potential harms, and responsible data release.[10]

Data provenance and consent. The Dido-Yvanchyk corpus is derived from publicly available audio recordings of a single narrator reading *"Dido Yvanchik"*, published on YouTube.[11] The literary work *"Dido Yvanchik"* by Petro Shekeryk-Donykiv

[10] https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics

[11] https://www.youtube.com/@didoyvanchik7322

is a canonical text of Hutsul cultural heritage and is available in the public domain. The recordings were produced by the Ukrainian Cultural Fund,[12] which releases its materials under an open license permitting use for research purposes.[13] Accordingly, the use of these recordings complies with applicable consent and data usage requirements for academic research.

Acknowledgments

We thank the Kyiv School of Economics (KSE) for institutional support and for providing computational resources used in this work. We also thank Vasyl Zelenchuk for his work on the reading "Dido Yvanchyk" and providing it to the public domain.[14].

References

Rifqi Naufal Abdjul and et al. 2025. Indonesian speech content de-identification in low resource transcripts. In *Proceedings of SEALP 2025*.

Oliver Adams and et al. 2020. A study of multidialect speech recognition with massively multilingual models. *Preprint*, arXiv:2007.03001.

Ahmed Ali and et al. 2014. Arabic speech recognition for iwslt: Dialectal challenges and system description. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Natalia Meyer, Joelle Stevens, Michael Henretty, Lindsay Morais, Cătălina Saunders, Wan Xi Chua, and 1 others. 2022. Mozilla common voice corpus 11.0. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. ELRA. A large-scale multilingual speech dataset.

Arun Babu and et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Max Bachmann. 2020. Rapidfuzz: A fast fuzzy string matching library. https://github.com/maxbachmann/RapidFuzz.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Li Chen and et al. 2025. Fuzzyseg: Segmentation and alignment for long-form speech. *Preprint*, arXiv:2505.15646.

Albert Coll and et al. 2023. Character-based or subword-based sequence labeling? a comparative study. In *Proceedings of CoNLL 2023*.

Alexis Conneau and et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.

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

ElevenLabs. 2024. Meet scribe: Word error rates across 102 languages. https://elevenlabs.io/blog/meet-scribe#fleurs-word-error-rate-102-languages. Accessed: 2025-01-XX.

Xing Gao and et al. 2025. Dynamic alignment for speech and text in low-resource settings. *Preprint*, arXiv:2509.24478.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, and Ruslan Salakhutdinov. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Chen, Yu Li, Sijia Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint*, arXiv:2106.09685.

Ondřej Klejch and et al. 2025. A practitioner's guide to building asr models for low-resource languages: A case study on scottish gaelic. *Preprint*, arXiv:2506.04915.

Ashish Kumar and et al. 2025. Hindi dialect speech recognition in low-resource settings. *Preprint*, arXiv:2507.15272.

Michael McAuliffe, Michaela Socolof, Steven Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*, pages 498–502.

Davide Michelsanti and et al. 2019. Asr for unwritten languages: A survey. *Preprint*, arXiv:1907.13511.

Hieu Nguyen and et al. 2023. Investigating the impact of data augmentation techniques for low-resource asr. *Preprint*, arXiv:2307.07948.

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

[12] https://ucf.in.ua/

[13] https://ucf.in.ua/storage/docs/06012025/CUW%202025%20%D0%9B%D0%9E%D0%A2%201_4ca00d7c2239c56085a20d4f869de9cd4cb2b9fe.pdf

[14] https://www.youtube.com/@didoyvanchik7322

Stefan Pugh and Ian Press. 2005. *Ukrainian: A Comprehensive Grammar*. Routledge.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

ReadBeyond. 2015. Aeneas: Audio-text forced alignment toolkit. GitHub repository.

Codruț Rotaru, Nicolae Ctlin Ristea, and Radu Tudor Ionescu. 2024. RoDia: A new dataset for Romanian dialect identification from speech. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 279–286, Mexico City, Mexico. Association for Computational Linguistics.

Taras Sereda. 2024. Transcribe, align and segment: Creating speech datasets for low-resource languages. *Preprint*, arXiv:2406.12674.

Anurag Shas. 2024. Whisper-large-v2 fine-tuned for ukrainian speech recognition. HuggingFace model.

Bowen Shi and et al. 2023. Whisperx: Time-accurate speech transcription with word-level alignment. *Preprint*, arXiv:2303.00747.

Yehor Smoliakow and et al. 2023. Ukrainian xls-r speech model. Ukrainian adaptation of XLS-R pretrained on Common Voice.

Yehor Smoliakow and et al. 2024. Wav2vec2-bert-uk-v2.1: Ukrainian wav2vec2 + bert integrative model. HuggingFace model.

SeatGeek Open Source. 2011. Fuzzywuzzy: Fuzzy string matching in python. `https://github.com/seatgeek/fuzzywuzzy`.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press.

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 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Jesin James Thennal D K. 2024. Advocating character error rate for multilingual asr evaluation. `https://arxiv.org/abs/2410.07400`. ArXiv:2410.07400.

Tianyu Zhong, Ziqi Yang, Zhen Liu, Rui Zhang, Yiheng Liu, Hanqi Sun, Yujia Pan, Yiming Li, and Yifan Zhou. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

Split	Mean	Median	Std	Min	Max	P95
Train	8.31	6.64	6.70	0.11	78.84	21.13
Validation	8.27	6.63	6.36	0.62	64.44	19.24
Test	8.47	6.76	6.72	0.11	42.98	21.79
Overall	8.35	6.68	6.59	0.11	78.84	20.72

Table 6: Audio duration statistics (in seconds)

Split	Characters	Words	Unique	Mean len
Train	736,512	131,142	21,816	108.1
Validation	81,715	14,582	4,936	107.9
Test	92,900	16,480	5,495	110.3
Total	911,127	162,204	32,247	108.8

Table 7: Text transcription statistics

A Dataset Characteristics

This section provides additional descriptive statistics and qualitative error analysis for the Hutsul speech corpus used in our experiments.

Tables 6 and 7 provide supplementary statistics for the dataset.

Table 7 reports transcription-level statistics for each data split, including character and word counts, vocabulary size, and average segment length.

A.1 Error Pattern Analysis

Character-level errors are dominated by vowel substitutions. The most frequent patterns include i(i)→и(y) (81),и(y)→i(i) (51), у(u)→в(v) (26), and е(e)→є(ye) (16), reflecting dialectal vowel variation and non-standard orthography.

Consonant-level errors are less frequent and include alternation between у(u) and в(v), voicing changes such as з(z)↔с(s) and д(d)↔т(t), as well as substitutions involving the Hutsul-specific consonant ґ(g).

Vowel substitutions represent the most frequent error type overall. The dominant pattern is alternation between и(y) and i(i), reflecting the characteristic phenomenon in Hutsul speech. Other frequent changes include е(e)↔и(y), е(e)→є(ye), and а(a)↔о(o), consistent with known dialectal variation.

B Qualitative Recognition Examples

Below we show sample recognition outputs for OmniASR 300M and 1B models. Cyrillic text is shown with English transliteration.

OmniASR 300M

- “Жеріт, жер би вас гад, йкісте бешшєсно проїсні.”
  Zheryt, zher by vas had, ykiste beshshesno projisni.
  “жеріт жер би вас гадий кісте бешєсно про єсні”
  Zheryt zher by vas hadyj kiste beshesno pro jesni.
  WER=0.63, CER=0.11

- “Нима страху, ни завіситси».”
  Nyma strachu, ny zavisytsi.
  “нима страху низавісиц”
  Nyma strachu nyzavisyts.
  WER=0.50, CER=0.17

OmniASR 1B

- “Жеріт, жер би вас гад, йкісте бешшєсно проїсні.”
  Zheryt, zher by vas had, ykiste beshshesno projisni.
  “жеріт жерби вас гади йкійсте бешєсно проєсні”
  Zheryt zherby vas hady ykiiste beshesno projesni.
  WER=0.75, CER=0.11

- “Нима страху, ни завіситси».”
  Nyma strachu, ny zavisytsi.
  “нима страху ни завіситс”
  Nyma strachu ny zavisyts.
  WER=0.25, CER=0.04