Fine-tuning Whisper for Hungarian Speech Recognition

Dániel Varga

Second Semester Project Presentation
Eötvös Loránd University, Institute of Mathematics
Supervisors: Bence Bakos, András Lukács

June 4, 2026

Overview

1. Goal and Context
2. Dataset and Evaluation Setup
3. Pipeline Implementation
4. Experiments and Results
5. Conclusions and Future Work

From Foundations to Fine-tuning

- Long-term goal: develop an accurate Hungarian automatic speech recognition (ASR) system
- First semester: theoretical and experimental foundation
  - Whisper architecture
  - WER-based evaluation
  - preliminary robustness experiments
- Second semester: practical adaptation
  - Whisper fine-tuning pipeline
  - BEA-Base corpus (BEszélt nyelvi Adatbázis)
  - manual experiments and Optuna hyperparameter optimization (HPO)

Whisper Model Recap

- Whisper is a large pretrained multilingual ASR model trained on diverse speech data, making it robust to noise, accents, and speaking styles
- Since it already supports Hungarian, fine-tuning Whisper is a practical alternative to training an ASR model from scratch

Model size	Parameters	Relative role
Tiny	39M	very fast, limited capacity
Base	74M	lightweight baseline
Small	244M	good balance, used in this project
Medium	769M	higher accuracy, higher cost
Large	1550M	strongest variant, most expensive

BEA-Base: Hungarian ASR Benchmark

- BEA-Base is designed for Hungarian ASR training and evaluation
- Contains mostly spontaneous speech, but also includes repeated/read segments
- Realistic but challenging data: natural speech, hesitations, restarts, variable speaking styles, and some annotation noise
- Therefore, both word-level and character-level metrics are informative

Subset	Duration	Words
Train	71.20 h	555,322
Development total	4.67 h	32,049
Evaluation total	5.86 h	41,407

Main Research Question

Question

Can `openai/whisper-small` be substantially improved for Hungarian ASR through supervised fine-tuning?

- Baseline: Whisper-small on BEA-Base dev
- Adaptation: supervised fine-tuning on BEA-Base train
- Optimization: hyperparameter search with Optuna
- Evaluation: Word Error Rate (WER) and Character Error Rate (CER)

Evaluation Protocol

WER formula

$$\mathrm{WER} = \frac{S + D + I}{N}$$

- S: substitutions, D: deletions, I: insertions, N: reference words
- WER is basically word-level edit distance (i.e. Levenshtein distance)
- CER is the same on character level
- Lower WER and CER mean better recognition quality
- My experiments evaluate on the full BEA-Base development set: 577 examples

Fine-tuning Pipeline

Main stages:

- manifest preparation and train/dev/test split handling
- Whisper fine-tuning with HuggingFace `Seq2SeqTrainer`
- evaluation with WER and CER
- artifact export: metrics, predictions, logs, summary

Design goal

The same infrastructure can support larger Whisper models, reproducible measurements and a wide range of experiments.

Baseline Results and Fine-tuning

Run	WER	CER
Baseline	52.79%	22.52%
Manual best	**27.58%**	10.39%
HPO full-dev rerun	27.66%	**10.15%**

- The pretrained model was first evaluated without fine-tuning
- Whisper-small was not sufficient without fine-tuning on this Hungarian dev set
- Supervised fine-tuning produced a large improvement

Manual Fine-tuning Attempts

- The most important factor was the learning rate
- Too small learning rate: under-adaptation
- Too large learning rate: unstable or ineffective adaptation

LR	Eff. batch	Epochs	WER
$1.0 \cdot 10^{-6}$	16	12.2	37.49%
$1.0 \cdot 10^{-3}$	16	12.2	46.36%
$1.0 \cdot 10^{-4}$	16	12.2	**27.58%**
$1.0 \cdot 10^{-4}$	64	35.0	28.01%

Optuna Hyperparameter Optimization

- Manual experiments covered only a small set of configurations
- Optuna was used for a more systematic search
- Search space around the best manual setup:
  - learning rate: $5.0 \cdot 10^{-5}$ to $2.5 \cdot 10^{-4}$
  - epochs: approximately 10.5, 14.0, 17.5, or 21.0
  - warmup ratio: 0.05 to 0.20
- Effective batch size was fixed at 128
- In-training evaluation used a subset of the dev set for faster runtime

HPO Results

Trial	LR	Warmup ratio	Epochs	WER
0	$9.14 \cdot 10^{-5}$	0.159	21.0	28.39%
1	$1.31 \cdot 10^{-4}$	0.073	10.5	27.94%
2	$5.49 \cdot 10^{-5}$	0.140	21.0	29.94%
3	$1.56 \cdot 10^{-4}$	0.195	10.5	**27.74%**
4	$1.91 \cdot 10^{-4}$	0.076	10.5	27.77%

- Best trial: Trial 3
- The best HPO setting was rerun on the full development set
- Full-dev rerun: 27.66% WER and 10.15% CER

Interpreting the Error Distributions

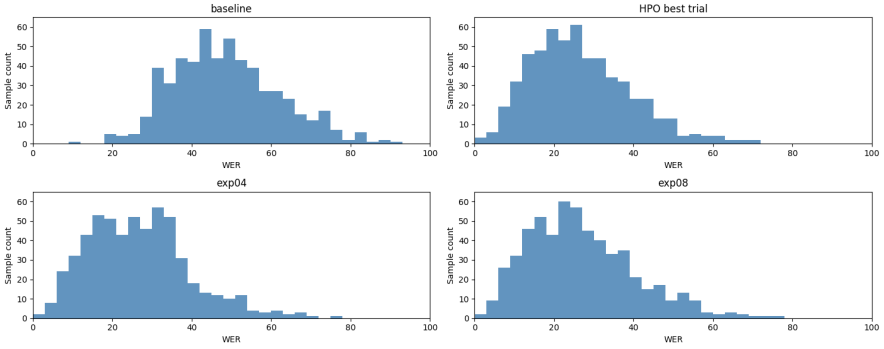

Figure: Sample-level WER distributions before and after fine-tuning.

- A sample is one evaluated audio segment, usually up to 30 seconds long
- Fine-tuning shifts the WER distribution strongly to the left
- However, some utterances remain difficult even after fine-tuning

Final Remarks

- Supervised fine-tuning substantially improved Hungarian ASR quality
- WER decreased from 52.79% to 27.58%
- CER decreased from 22.52% to 10.15%
- The pipeline is ready for future experiments
- Future work:
  - test larger Whisper variants
  - run broader HPO with more trials
  - analyze difficult utterances by speech type and error category
  - improve the quality of transcriptions using a text-to-text post-processing model

AI Usage During the Project

- Occasional support with designing and debugging the fine-tuning pipeline
- Assistance with structuring and wording the final report and presentation slides

Thank you for your attention.

Bibliography

- A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022.
- Mihajlik, P., Balog, A., Gráczi, T. E., Kohári, A., Tarján, B., & Mády, K. (2022). *BEA-Base: A Benchmark for ASR of Spontaneous Hungarian.* In LREC 2022, Thirteenth International Conference on Language Resources and Evaluation (pp. 1970–1977).