Automatic speech recognition for Hungarian

Dániel Varga

Eötvös Loránd University, Faculty of Science
Supervisors: Bence Bakos, András Lukács

January 8, 2026

Overview

1. Goal and Context
2. Audio Processing
3. Whisper Model
4. My Practical Work This Semester
5. Challenges and Next Steps

Overall Goal

- Long-term goal: develop an accurate Hungarian ASR system
- Main objectives this semester:
  - learn the basics of modern neural ASR
  - exploring the Whisper model
  - build an initial evaluation and experimentation pipeline

ASR in One Sentence

High-level idea

Audio in $\rightarrow$ neural model $\rightarrow$ text out

- Deep neural networks learn from speech–text pairs
- Transformers are the standard architecture today

Log-Mel Spectrograms

Key question: what representation of audio do we feed to the model?
$\rightarrow$ log-mel spectrograms

- Raw audio is not a suitable direct input for neural networks
- Log-mel spectrograms transform speech into a time–frequency representation
- The mel scale emphasizes perceptually relevant frequencies
- The power is converted to a logarithmic scale (decibels)

Takeaway

Log-mel spectrograms provide a compact, informative representation that modern ASR models can learn from efficiently.

Whisper: What It Is

- OpenAI Whisper: large pre-trained multilingual ASR model
- Trained on very large, diverse data
- Robust to noise, accents, and speaking styles
- Advantages of fine-tuning Whisper:
  - Already supports Hungarian
  - Cheaper than training from scratch
  - Robustness helps with realistic, non-studio recordings
  - Many existing fine-tuning results to learn from

Whisper Pipeline: Big Picture

- Input: log-mel spectrogram
- Encoder:
  - produces high-level acoustic representations
- Decoder:
  - generates text token-by-token
  - autoregressive: at each step, predict next token given past tokens and audio

Initial Experiments

Evaluation: Word Error Rate (WER)

- Standard ASR metric

WER definition

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

- S: substitutions, D: deletions, I: insertions, C: correct

Noise Augmentation

- Goal: test (and later improve) robustness
- Method: mix clean speech with additive noise at different SNR levels
- Then evaluate with WER on clean vs. noisy inputs
- Expectation: more noise $\Rightarrow$ higher WER

Initial Observation from Noisy Tests

- Trend: WER increases as noise increases
- Note:
  - these are early sanity-check results
  - WER is sensitive to normalization

Takeaway

Whisper is robust to noise and consistent with large-scale, diverse training data.

SNR (dB)	Medium	Turbo
10	4.26%	4.81%
5	10.11%	5.35%
1	16.49%	6.42%
-1	11.17%	9.09%
-5	37.77%	7.49%
-10	34.04%	32.62%

Table: WER comparison across noise levels using different Whisper models on an English sample

Challenges

- Finding suitable Hungarian speech corpora
  - aspects to consider: size / quality / accessibility
- Evaluation reliability
  - WER is sensitive to normalization and preprocessing
- Practical constraints
  - large models $\Rightarrow$ careful GPU/resource management

Plans for the Upcoming Semester

- Find or construct a Hungarian ASR dataset
- Learn and apply Whisper fine-tuning practices
- Run controlled experiments:
  - compare model sizes / settings
  - measure improvements with WER
  - analyze typical Hungarian error patterns

Thank you for your attention.

AI Usage During the Project

- Clarification of theoretical concepts in deep learning and ASR
- Assistance with wording and structuring presentation slides
- Occasional support for checking implementation details and evaluation logic