First Semester Project Report: Development of a Hungarian Speech-to-Text System

Daniel Varga

December 14, 2025

1 Introduction and Project Goals

The objective of this project is to establish a solid technical foundation for developing an automatic speech recognition (ASR) system tailored to the Hungarian language. The work focuses on the mathematical and algorithmic components of modern ASR pipelines, with particular emphasis on Whisper, an encoder–decoder transformer trained on large-scale multilingual data.

By the end of the semester, I gained a working understanding of end-to-end ASR principles, implemented an audio preprocessing and evaluation pipeline, and conducted initial robustness experiments with Whisper to assess model behaviour under noisy conditions. These components provide the basis for future fine-tuning efforts on Hungarian speech data.

2 Whisper Architecture and Input Representation

Speech recognition involves mapping a continuous, highly variable audio signal to a discrete sequence of text tokens. This task is challenging due to differences in speakers, speaking styles, recording conditions, and background noise. Contemporary approaches address these challenges by learning this mapping directly from data using large neural models, rather than relying on handcrafted processing pipelines. In this project, the Whisper model is used as a representative example of such modern speech recognition systems, providing the basis for the methods and experiments presented in the following sections.

Whisper uses log-Mel spectrograms as its input representation, which are computed from raw audio through a sequence of standard signal processing steps. The audio waveform is first sampled at a fixed rate and normalized, then divided into short, overlapping time frames to capture local temporal structure. For each frame, a short-time Fourier transform is applied to obtain a frequency-domain representation, which is then projected onto a perceptually motivated Mel frequency scale using a filter bank. The resulting Mel-band energies are converted to a logarithmic, decibel-based scale, which compresses the dynamic range of the signal and reduces sensitivity to variations in loudness and recording conditions. The final log-Mel spectrogram can be interpreted as a two-dimensional time–frequency representation that preserves both spectral and temporal information in a compact form suitable for neural speech recognition models such as Whisper.

The log-Mel spectrogram serves as the sole input to the Whisper model. Architecturally, Whisper is a sequence-to-sequence model based on the transformer framework. It consists of an encoder and a decoder, both implemented as stacks of transformer blocks. The encoder processes the input spectrogram and produces a sequence of high-level acoustic representations that capture both local and long-range temporal dependencies through self-attention mechanisms.

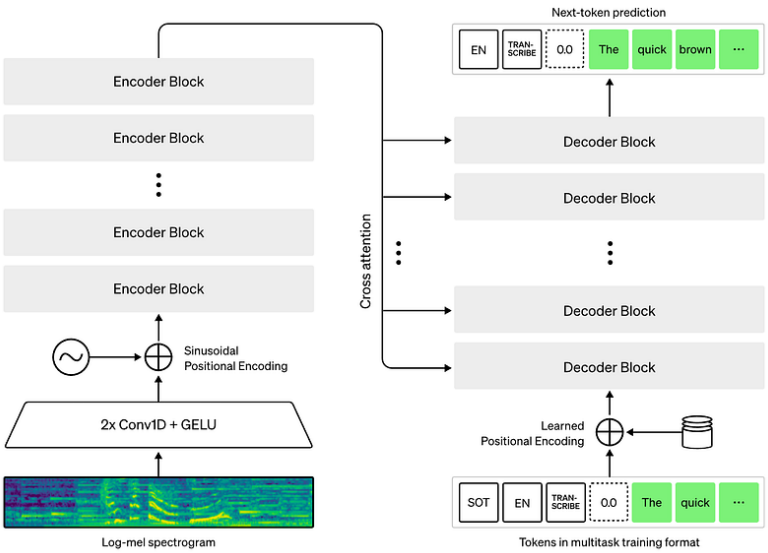

Figure 1: Overview of the Whisper model architecture.

The decoder is also a transformer and generates the output text autoregressively, one token at a time. At each decoding step, the model conditions explicitly on the previously predicted tokens in addition to attending to the encoder representations, allowing it to model linguistic context while aligning segments of the input audio with the corresponding textual output. This autoregressive formulation is essential for maintaining grammatical consistency and capturing long-range dependencies in the generated transcription.

A defining characteristic of Whisper is its multitask training setup. The model is trained on large-scale multilingual data covering multiple speech-related tasks, such as speech recognition and speech translation. Due to its scale, multilingual coverage, and multitask design, Whisper learns robust acoustic and linguistic representations that generalize well across domains and noise conditions. These properties make it a strong foundation for subsequent adaptation and fine-tuning toward Hungarian automatic speech recognition.

3 Experimental Pipeline and Evaluation Setup

For experimentation, I implemented a lightweight preprocessing pipeline aligned with Whisper's input requirements. Raw audio is transformed into log-Mel spectrograms using `torchaudio` with configurable STFT parameters, allowing analysis of different time–frequency resolutions.

Evaluation relies on word error rate (WER), computed from the minimal Levenshtein distance between reference and predicted word sequences:

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

where S, D, and I denote substitutions, deletions, and insertions, and $|R|$ is the number of reference words. Text normalization (lowercasing, punctuation removal, whitespace standardization) is applied prior to comparison to ensure consistency.

4 Noise Robustness Experiments

To assess Whisper's robustness to additive noise, I conducted controlled experiments on an English speech sample extracted from a publicly available podcast recording. Multiple noisy variants of the audio were generated at signal-to-noise ratios (SNR) ranging from 10 dB to –10 dB using additive white noise. For each noisy input, WER was computed separately for the `medium` and `turbo` Whisper models, using a manual transcript as reference.

SNR (dB)	**Medium WER**	**Turbo WER**	**Difference**
10	4.26%	4.81%	-0.56%
5	10.11%	5.35%	4.76%
1	16.49%	6.42%	10.07%
-1	11.17%	9.09%	2.08%
-5	37.77%	7.49%	30.28%
-10	34.04%	32.62%	1.42%
		Average improvement	8.01%

Table 1: WER comparison of Whisper-medium and Whisper-turbo across noise levels.

These measurements are preliminary and based on a single English podcast segment; therefore, they should not be interpreted as statistically representative. Their purpose is to validate the evaluation pipeline and to illustrate qualitative differences between model variants under controlled noise conditions.

5 Future Work

Future work will focus on preparing and curating Hungarian speech datasets suitable for supervised fine-tuning. The main objectives for upcoming semesters include:

- Finalizing the choice of one or more Hungarian speech corpora, cleaning the transcripts, and aligning them with audio segments suitable for training.
- Fine-tuning Whisper on Hungarian speech data using established frameworks.
- Evaluating the resulting models on standardized Hungarian benchmarks where possible, and analyzing common error types.

6 Conclusion

This semester established the theoretical and computational foundation required for developing a Hungarian ASR system. The work included studying Whisper's architecture, implementing essential preprocessing and evaluation tools, and conducting initial robustness experiments. These components lay the groundwork for future fine-tuning and more comprehensive experimentation aimed at building a high-quality Hungarian speech-to-text model.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, 2023, pp. 28492–28518.