BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

# Czech Speech Emotion Recognition

Bc. Adam Rajko

**Abstract**

**Problem / goal:** High-quality Czech speech emotion recognition (SER) data is scarce. Manual collection and annotation is expensive, and available corpora often lack speaker diversity and scale. Our goal is to build a reproducible pipeline that automatically converts long-form Czech online recordings into short, speaker-homogeneous speech segments with transcripts, while keeping emotion labels provided by human annotators.
**Methodology:** As an audio source we use YouTube platform, preprocess it, perform speaker diarization with NVIDIA NeMo, transcribe with Whisper (turbo) with word timestamps, and align words to diarized speaker intervals. The pipeline's primary purpose is to automate segmentation and transcript generation; emotion/sentiment labels are assigned manually by annotators.
**Results:** The pipeline produces segment-level audio and aligned transcripts together with a manifest suitable for training audio-based SER models (arousal/valence regression and multilabel emotion classification).
**Impact:** The proposed workflow lowers the barrier to creating Czech SER datasets by reducing the effort needed to obtain clean, speaker-homogeneous segments and consistent transcripts, while preserving label quality through human annotation. In addition, we provide a baseline fine-tuning setup for common audio backbones to validate usability of the resulting dataset and to enable quick iteration on modeling choices.

*xrajko00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

## 1. Introduction

**Motivation.** Speech Emotion Recognition (SER) benefits from large and diverse corpora, but Czech resources are limited compared to high-resource languages. A major bottleneck is turning long recordings into clean, speaker-homogeneous segments that are practical to annotate; segmentation and transcript preparation are time-consuming and error-prone when done manually. At the same time, long-form Czech speech in the wild (e.g., interviews, podcasts, debates) is abundant and contains varied speaking styles and emotional states, making it a valuable source once it is properly segmented.

**Problem definition.** We address the problem of building an *end-to-end Czech speech emotion recognition (SER) system*: given short speech segments with human-provided emotion annotations, train and evaluate an audio-based model that can generalize to unseen speakers and real-world recordings.

As a supporting contribution (bonus), we develop a reproducible pipeline that turns long-form Czech online recordings into clean, speaker-homogeneous audio segments with aligned transcripts, making manual annotation practical at scale and producing an ML-ready manifest for training.

**Existing solutions.** SER research traditionally relies on curated, manually annotated corpora, often recorded in controlled conditions or with acted emotions; commonly used resources include the acted Emotional Speech Database (EMO-DB) and acted corpora such as RAVDESS-style datasets where applicable, while many modern systems also evaluate on multilingual benchmarks (e.g., IEMOCAP) [1]. For representation learning, state-of-the-art SER systems frequently reuse self-supervised speech encoders such as wav2vec 2.0 or HuBERT and fine-tune them for emotion targets, which improves data efficiency but remains sensitive to domain and language mismatch [2, 3].

However, emotion annotation remains the main bottleneck due to subjectivity, inter-annotator disagreement, and the difficulty of capturing prosody and context from text alone.

**Our approach.** We implement an end-to-end pipeline that downloads audio, preprocesses it to a standard format, runs speaker diarization, transcribes Czech speech with word timestamps, and aligns transcripts

to speaker segments. The pipeline is designed to automate the creation of high-quality segments (and their transcripts) to make subsequent manual emotion annotation by human annotators efficient and scalable.

**Contributions.** (i) A reproducible automated pipeline for Czech SER dataset creation focused on segmentation, with caching of intermediate artifacts; (ii) alignment logic that combines diarization segments with word-level ASR timestamps to form clean, speaker-homogeneous samples; (iii) a dataset format (manifest + segment metadata) that supports efficient manual emotion annotation by human annotators; (iv) baseline fine-tuning setup for audio backbones (wav2vec2/Hubert) with windowed inference and multi-head outputs.

## 2. Dataset Creation Pipeline

The dataset creation process (see poster flow diagram) transforms raw YouTube videos into structured emotional speech samples.

**Step 1: Audio Acquisition** Audio is downloaded using `yt-dlp` and cached locally.

**Step 2: Preprocessing** Audio is converted to mono 16kHz and optionally processed with Demucs to isolate vocals. Long recordings are split into manageable chunks.

**Step 3: Speaker Diarization** NVIDIA NeMo clustering diarizer segments audio into speaker-homogeneous regions with timestamps.

**Step 4: Transcription** Whisper (turbo) produces Czech transcripts with word-level timestamps.

**Step 5: Alignment** Transcriptions are aligned with diarized segments and filtered based on duration and textual quality.

**Step 6: Emotion Labeling** Each segment is analyzed using:

- Czech Roberta classifier (high confidence threshold)
- Local LLM (fallback for uncertain cases)

## 3. Model and Training Pipeline

The training focuses on learning robust emotional representations from audio.

**Windowing Strategy** Audio is split into overlapping windows (e.g., 5s with 1s overlap) to handle variable-length inputs.

**Feature Extraction** Each window is processed by a pretrained transformer backbone:

- Wav2Vec2
- HuBERT

**Model Architecture** A shared encoder is followed by three task-specific heads:

- Regression: arousal + valence
- Multilabel classification: emotion leaf labels
- Multilabel classification: emotion categories

**Aggregation** Window-level predictions are averaged to obtain clip-level outputs.

## 4. Experiments

We evaluate the effect of annotator quality on model performance.

**Experimental Setup** We compare two pretrained backbones:

- HuBERT (`hubert-base-ls960`)
- Wav2Vec2 (`wav2vec2-base`)

Two dataset variants are used:

- **Full dataset**
- **Filtered dataset** (excluding annotator MJ)

**Metrics**

- Accuracy (Acc)
- F1 score

**Results**

| Model | Dataset | Target | Acc | F1 |
|---|---|---|---|---|
| HuBERT | Full | Arousal | 0.39 | 0.22 |
| HuBERT | Full | Valence | 0.43 | 0.26 |
| HuBERT | w/o MJ | Arousal | 0.64 | 0.64 |
| HuBERT | w/o MJ | Valence | 0.63 | 0.52 |
| Wav2Vec2 | Full | Arousal | 0.60 | 0.59 |
| Wav2Vec2 | Full | Valence | 0.40 | 0.40 |
| Wav2Vec2 | w/o MJ | Arousal | 0.54 | 0.53 |
| Wav2Vec2 | w/o MJ | Valence | 0.61 | 0.57 |

**Table 1.** SER performance comparison across models and dataset variants.

**Discussion**

Overall performance is inconsistent and in several cases weak, particularly on the full dataset. The results vary significantly between models and tasks, and improvements are not uniform.

A key observation is that removing a single annotator (MJ) leads to substantial performance changes, especially for HuBERT.

## 5. Conclusions

We presented an automated, reproducible pipeline for creating a Czech SER dataset from long-form online recordings. By combining standardized preprocessing, speaker diarization, word-timestamped transcription,

and alignment, the pipeline produces short speaker-homogeneous segments in a manifest format suitable for training. Emotion labels are provided by human annotators; the automated pipeline focuses on producing clean, speaker-homogeneous segments and aligned transcripts to make manual annotation practical at scale.

The main bottleneck is not the model architecture, but the **quality of annotations**.

## Acknowledgements

I would like to thank my supervisor Ing. Igor Szőke Ph.D.

## References

[1] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.