Speech Recognition Systems Using Wav2Vec Models

November 24, 2023

Jan Lehečka, jlehecka@kky.zcu.cz

DEPARTMENT OF CYBERNETICS

NTIS

AI TEAS

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

TABLE OF CONTENTS

01

ASR

Introduction to Automatic Speech Recognition

02

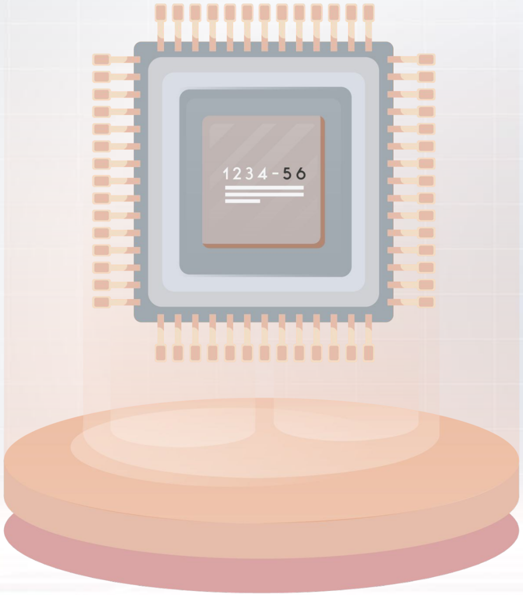
Wav2Vec Model

How does it work?

03

Our Models

Our recent successes, experiences, and challenges



01

ASR

Introduction to
Automatic Speech Recognition

Automatic Speech Recognition

The task: automatically transcribe speech into text

Essential part of many AI tools:

- dialogue systems
- virtual assistants
- video captioning / subtitling
- speech translation

Transcript quality is critical – errors propagate into subsequent tasks

A Brief History of ASR

1952

“Audrey” – the first digit recognizer

1970s

“Harpy” – 1000 words recognizer

1980s

Move from pattern matching to probabilistic modeling (HMM)

1990s, 2000s

Fast processors arrived → Large Vocabulary Continuous Speech Recognition (LVCSR)

2010s

Collecting speech data, cloud-based ASR services (Google Voice Search, Siri)

2020s (for now)

The age of Transformers, huge datasets and self-supervised learning

Current Trends in ASR

What is changing:

- **size of models** (up to billions of trainable parameters): larger models → better models, but more expensive training
- **size of datasets** (up to millions of hours of speech)
- adding more input/output **modalities** (text, speech, images, videos)
- adding more **languages** → powerful translation models
- increasing **GPU performance** and memory

What is NOT changing (as for now):

- the core **architecture** of the model – still almost the same Transformers as presented in 2017 [1]

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Top ASR Models Today

Wav2vec 2.0

[June 2020, Facebook AI]

Pioneer end-to-end ASR model

→ English, 100 million params
→ extension XLS-R: 128 languages, 300M params

Whisper

[September 2022, OpenAI]

Large multi-lingual translation model

→ can solve ASR+translation
→ 1.55 billion params (large)
→ 99 languages

SpeechT5

[October 2021, Microsoft]

Multi-modal extension of text model T5

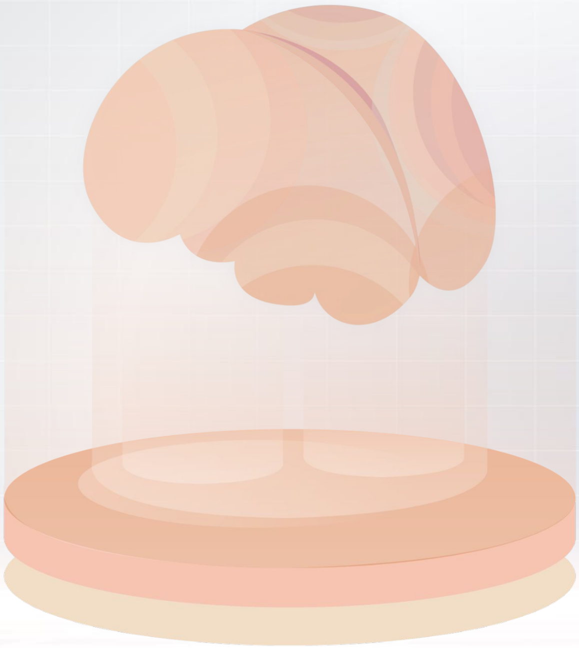
→ can solve ASR, speech-to-speech, TTS and text-to-text tasks
→ 153 million params
→ only English

SeamlessM4T

[August 2023, Meta AI]

Massively Multilingual & Multimodal Machine Translation model ("Babel Fish")

→ can solve any-to-any translation task between many languages
→ 600 million params
→ 100 languages



02

Wav2Vec Model

How does it work?

Transformer

Deep neural network introduced in 2017 by Google [1]

- became the core architecture of many successful AI models across various domains (text, speech, video, music, ...)
- sequence-to-sequence model
- originally designed for the **text domain**
- **encoder** – encodes the input sequence into contextualized embeddings
- **decoder** – auto-regressively decodes contextualized embeddings and generates the output sequence

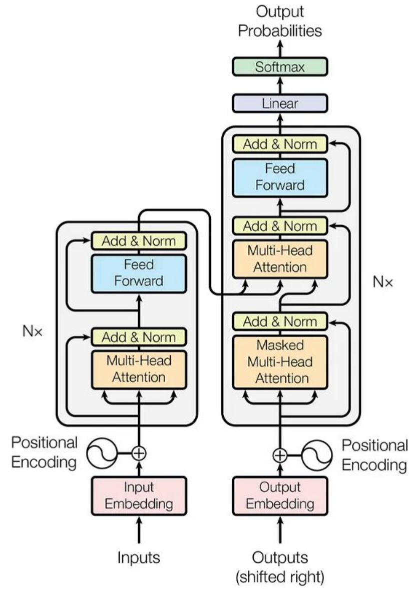


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Deep neural network introduced in 2017 by Google [1]

- became the core architecture of many successful AI models across various domains (text, speech, video, music, ...)
- sequence-to-sequence model
- originally designed for the **text domain**
- **encoder** – encodes the input sequence into contextualized embeddings
- **decoder** – auto-regressively decodes contextualized embeddings and generates the output sequence

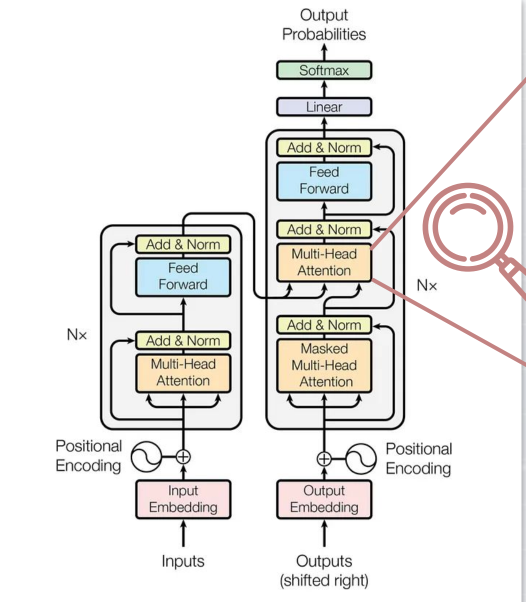
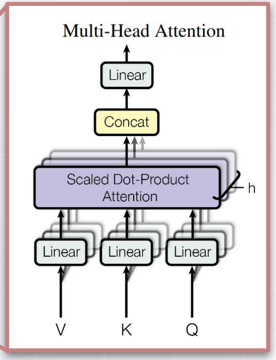


Figure 1: The Transformer - model architecture.



[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

Deep neural network introduced in 2017 by Google [1]

- became the core architecture of many successful AI models across various domains (text, speech, video, music, ...)
- sequence-to-sequence model
- originally designed for the **text domain**
- **encoder** – encodes the input sequence into contextualized embeddings
- **decoder** – auto-regressively decodes contextualized embeddings and generates the output sequence

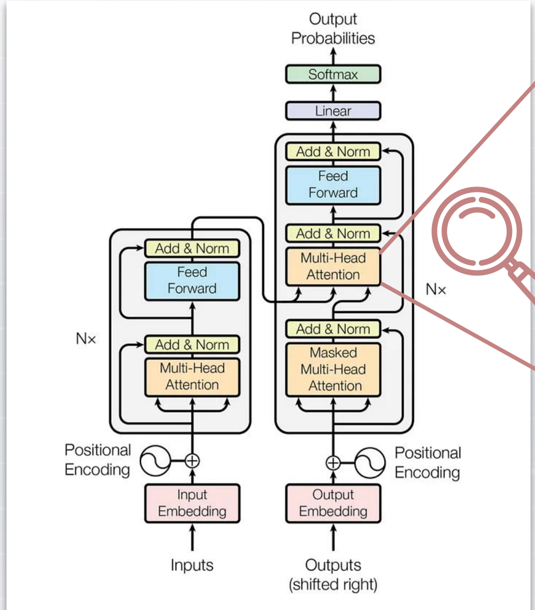
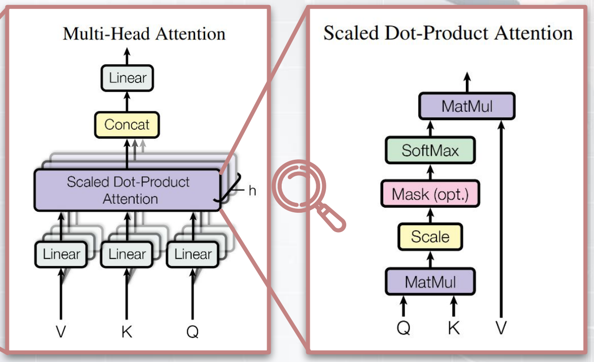


Figure 1: The Transformer - model architecture.



No magic inside!

Just basic matrix operations allowing the model to decide which part of input data to attend in every step

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Model Types

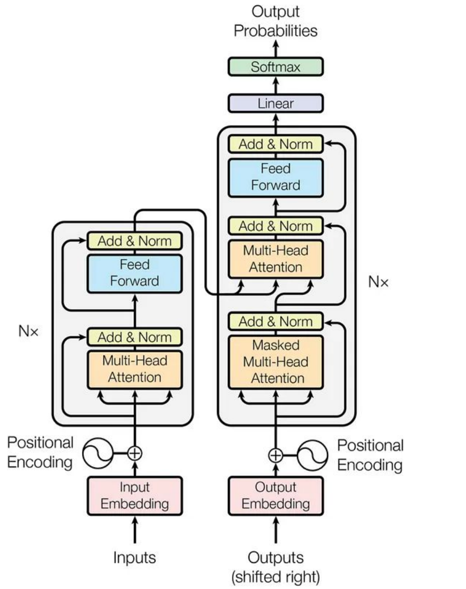


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Model Types

Encoder-only models

- BERT (text)
- RoBERTa (text)
- Wav2Vec (speech)
- ...

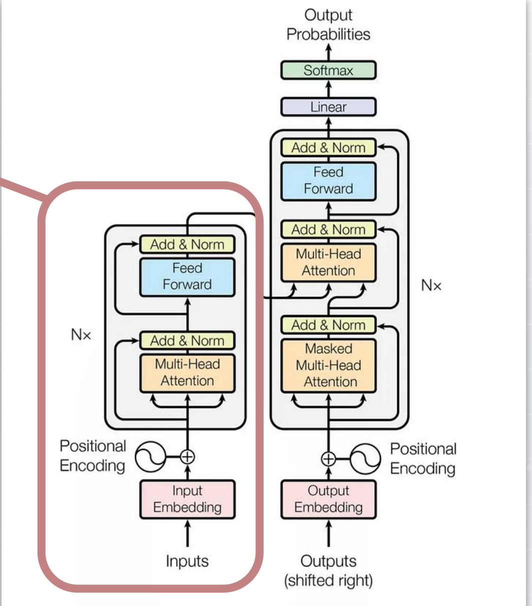


Figure 1: The Transformer - model architecture.

Transformer Model Types

Encoder-only models

- BERT (text)
- RoBERTa (text)
- Wav2Vec (speech)
- ...

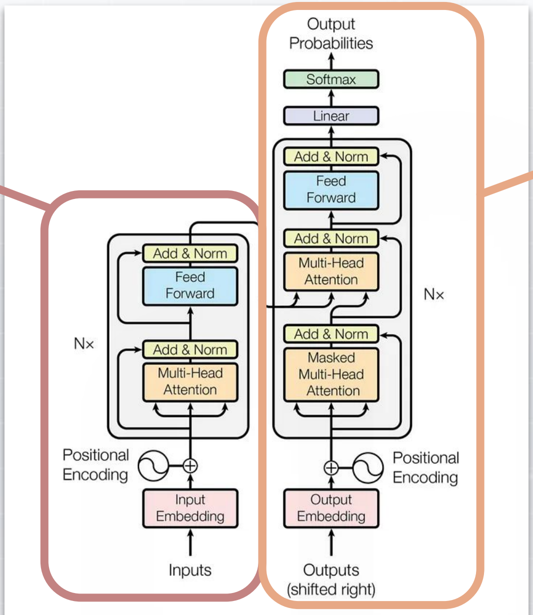


Figure 1: The Transformer - model architecture.

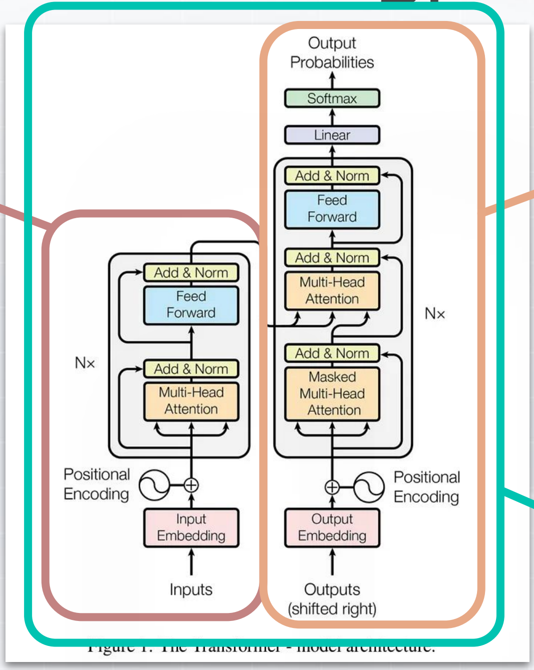
Decoder-only models

- Generative models
- GPT family
- Large language models (LLMs)

Transformer Model Types

Encoder-only models

- BERT (text)
- RoBERTa (text)
- Wav2Vec (speech)
- ...



Decoder-only models

- Generative models
- GPT family
- Large language models (LLMs)

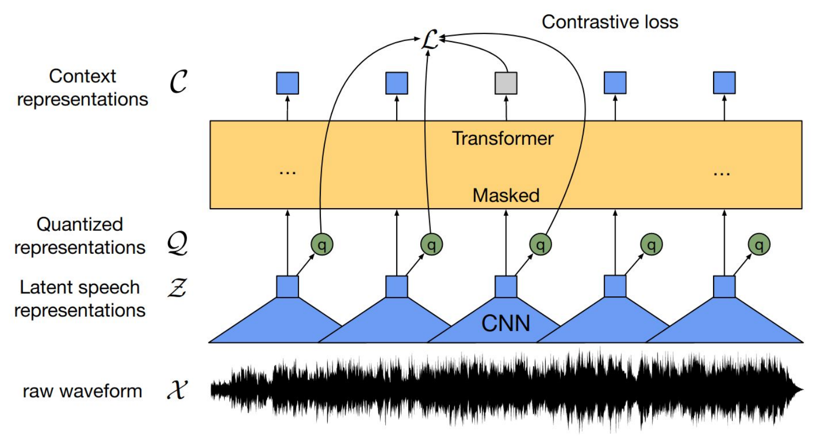
Encoder-Decoder models

- T5, BART (text)
- SpeechT5, Whisper, SeamlessM4T (speech&text)

Wav2Vec 2.0

Established new ASR paradigm – fine-tuning the model on only one hour of labeled speech data could beat the previous state-of-the-art systems trained on 100 times more labeled data [1].

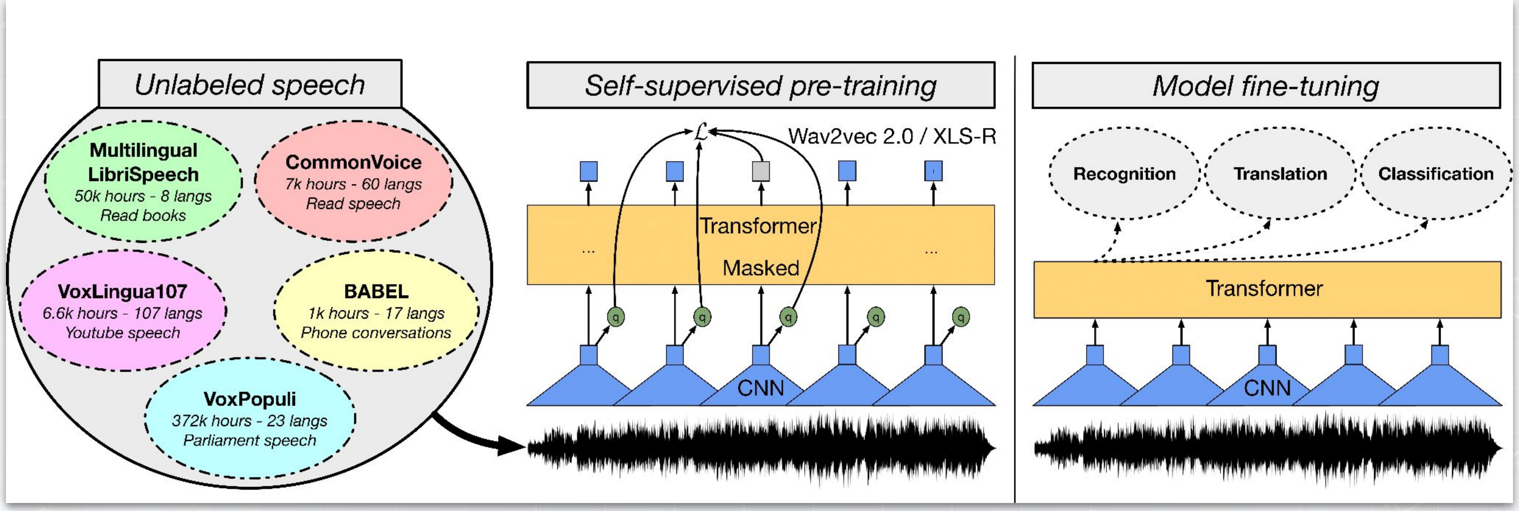
- end-to-end neural network
- **encoder-only** – do not suffer from hallucinating
- speech signal is sliced into small (20ms) frames and each frame is encoded
- output is a sequence of speech embeddings



Wav2vec solved some problems LVCSR had (and introduced some new problems LVCSR didn't have...)

[1] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in Neural Information Processing Systems 33 (2020): 12449-12460.

Wav2Vec Training



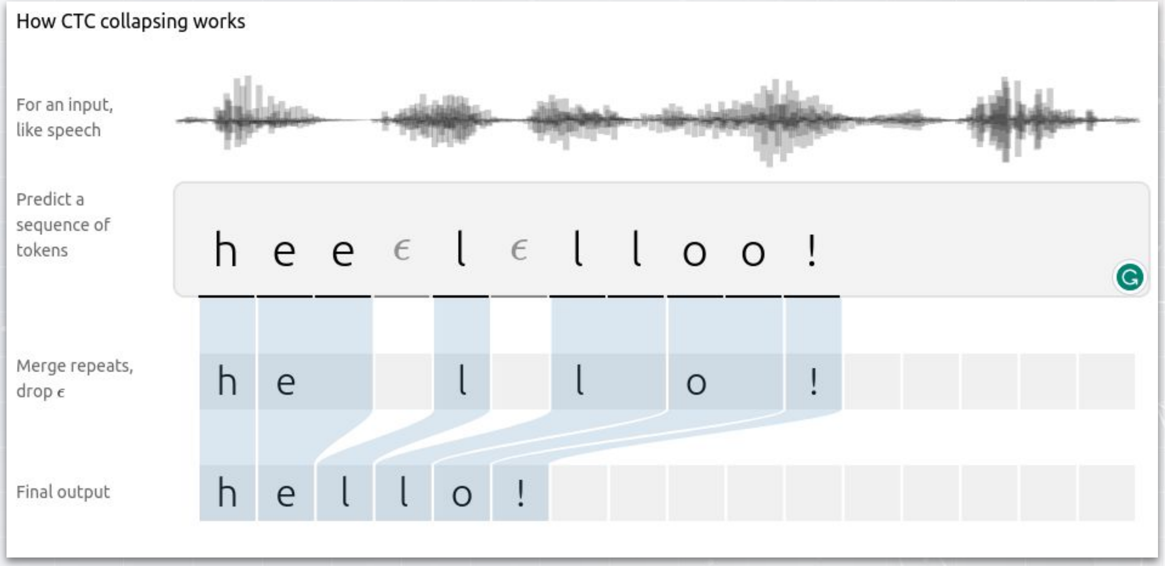
[1] Babu, Arun, et al. "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv preprint arXiv:2111.09296 (2021).

CTC – The Final Classification Layer

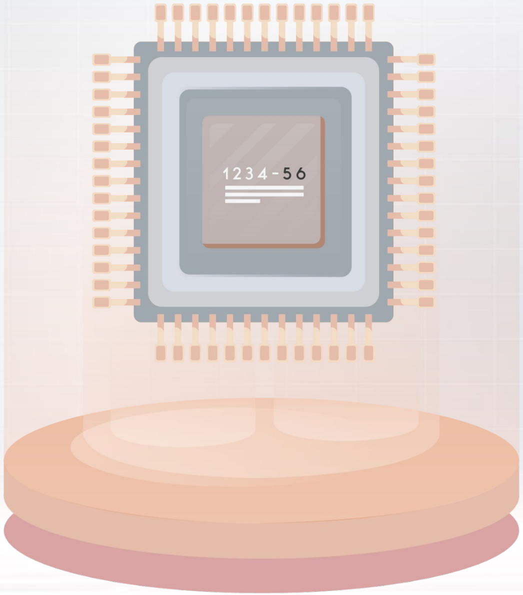
ASR is implemented by adding Connectionist Temporal Classification (CTC) layer on top of Wav2vec's encoder.

CTC algorithm:

1. assign the most probable output token to each audio frame
2. group sub-sequences with the same token into a single token
3. remove blank tokens



https://distill.pub/2017/ctc



03

Our Models

Our recent successes, experiences, and challenges

Motivation

Common approach: adopt a multilingual pretrained model and fine-tune it on own labeled ASR data.

We were not satisfied with results of public models

- Czech is minority language among worldwide training data
- we sit on large speech datasets
- we have access to high-end GPUs

So why not pre-train our own Czech state-of-the-art model from scratch?

82 401

hours of Czech speech used for pre-training
(~9.5 years of non-stop listening)

10 TB

dataset size on disk

14 days

pre-training time on a high-end machine with
4xA100 GPUs

ClTRUS

Czech **l**anguage **TR**ansformer from **U**nlabeled **S**peech

–

the Czech version of Wav2Vec 2.0 model developed on FAV/KKY

–

pre-trained model released publicly for non-commercial use:
https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS

Fine-tuning on ASR

Training time: 12 hours on a machine with 4xA100

We benefit from decades of research at our department – we have collected a lot of various labeled ASR data → **2-phase fine-tuning:**

1. 6 thousand hours of high-quality mixed-domain Czech labeled speech
2. smaller amount of in-domain Czech labeled speech

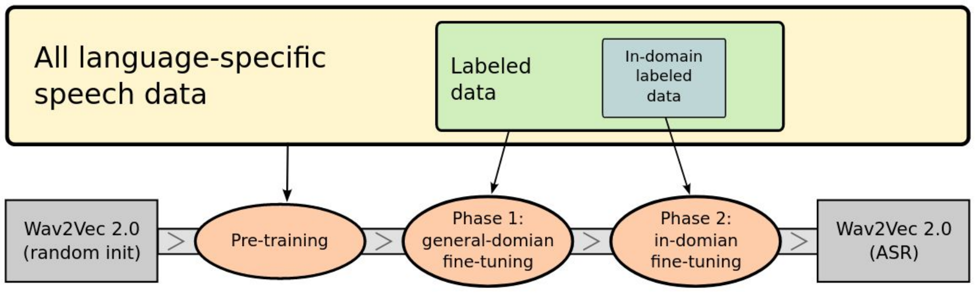


Figure 1: *The scheme of 2-phase fine-tuning.*

Results, successes

CZECH PUBLIC DATASETS

CommonVoice

- crowd-sourced project Mozilla Common Voice

VoxPopuli

- large-scale multilingual speech corpus
- collected from 2009-2020 European Parliament event recordings

	CommonVoice		VoxPopuli	
	no LM	LM-C5	no LM	LM-C5
5 epochs (default)	11.17	6.10	12.37	10.13
10 epochs (2xUP)	9.30	5.04	11.00	9.42
10 epochs (2xBS)	9.23	4.82	11.18	9.52
20 epochs (2xBS, 2xUP)	8.49	4.59	10.75	9.30
20 epochs (4xBS)	8.39	4.62	10.82	9.31
40 epochs (4xBS, 2xUP)	7.68	4.29	10.23	8.81
+ ASRSpec	**5.41**	**3.80**	**10.07**	**8.80**
Whisper-large	21.63		19.49	

Word Error Rates (WER) [%] [1]

[1] Lehečka, J., Švec, J., Pražák, A., Psutka, J.V.: Exploring Capabilities of Monolingual Audio Transformers using Large Datasets in Automatic Speech Recognition of Czech. In: Proc. Interspeech 2022. pp. 1831–1835 (2022). https://doi.org/10.21437/Interspeech.2022-10439

Results, successes

MALACH – memories of Holocaust survivors

- unique oral history archive
- audiovisual interviews in 32 languages
- very challenging dataset:
  - natural speech with emotional outpourings and heavy accents
  - old people (75 years old in average)

Method name	# of params	Year	English	Czech	German	Slovak
GMM-HMM diag	7.6M	2004	39.60	40.23		
GMM-HMM full	19M	2013	34.27	26.00		
DNN-HMM	31M	2017	30.91	23.18		
TDNN-LF-MMI	8.1M	2020	20.79	17.44		
CNN-TDNN-LF-MMI	6.8M	2021	17.85	14.65		
Wav2Vec 2.0	**95M**	**2023**	**12.88**	**8.43**	**17.08**	**11.57**
Wav2Vec-XLS-R	300M		14.31	9.50	22.52	12.17
Whisper-large	1550M		17.34	25.95	22.99	27.49

Word Error Rates (WER) [%]

Results, successes

UWebASR – simple public web interface to our ASR models

https://uwebasr.zcu.cz

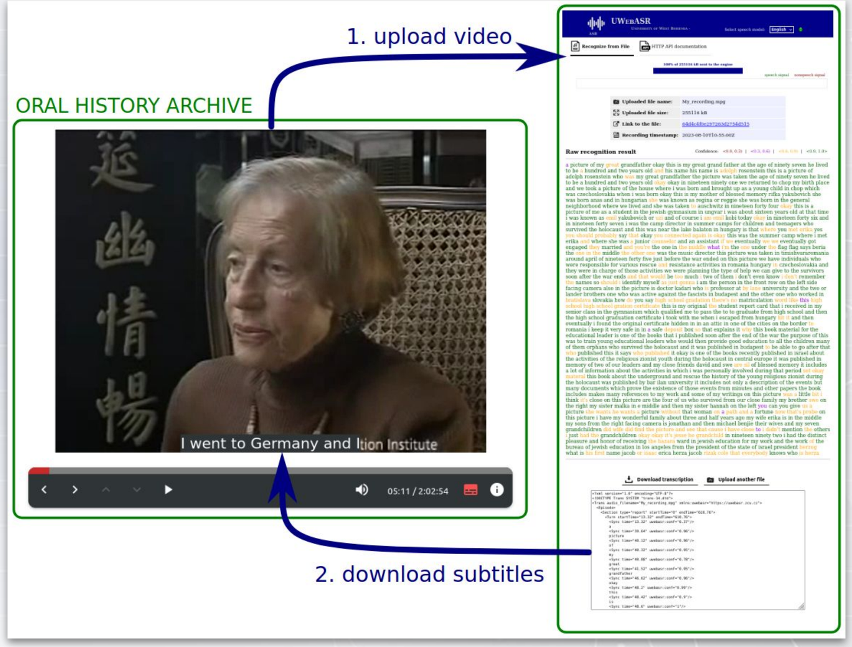
contact author: Jan Švec, honzas@kky.zcu.cz

Supported languages:

- English
- German
- Czech
- Slovak

Result formats:

- plaintext
- subtitles (XML, WebVTT)
- JSON (words, timestamps, confidence)



Challenges

High-end GPUs

expensive and scarce technology;
we train in Metacentrum and IT4I

Storage

TBs of data must be stored as close to GPUs as possible; operations with such a dataset take a long time

Fixing Errors

hard to find (and fix) cause of faulty outputs; hard to change transcription of specific words

ASR Latency

the nature of attention mechanism leads to high latency of models → hard to use as online ASR

Future plans

Czech SpeechT5 model

- multi-speaker TTS
- speech-to-speech tasks (edit audio, remove noise, change speaker, add emotions, ...)
- multi-modal inputs (speech and text prompt)

Multilingual models for oral history archives

- include translation among languages
- multilingual interactions

Enrich ASR output

- SCD task – speaker change detection
- SID task – speaker identification

THANK YOU FOR ATTENTION!

Jan Lehečka
jlehecka@kky.zcu.cz

DEPARTMENT OF CYBERNETICS

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

NTIS