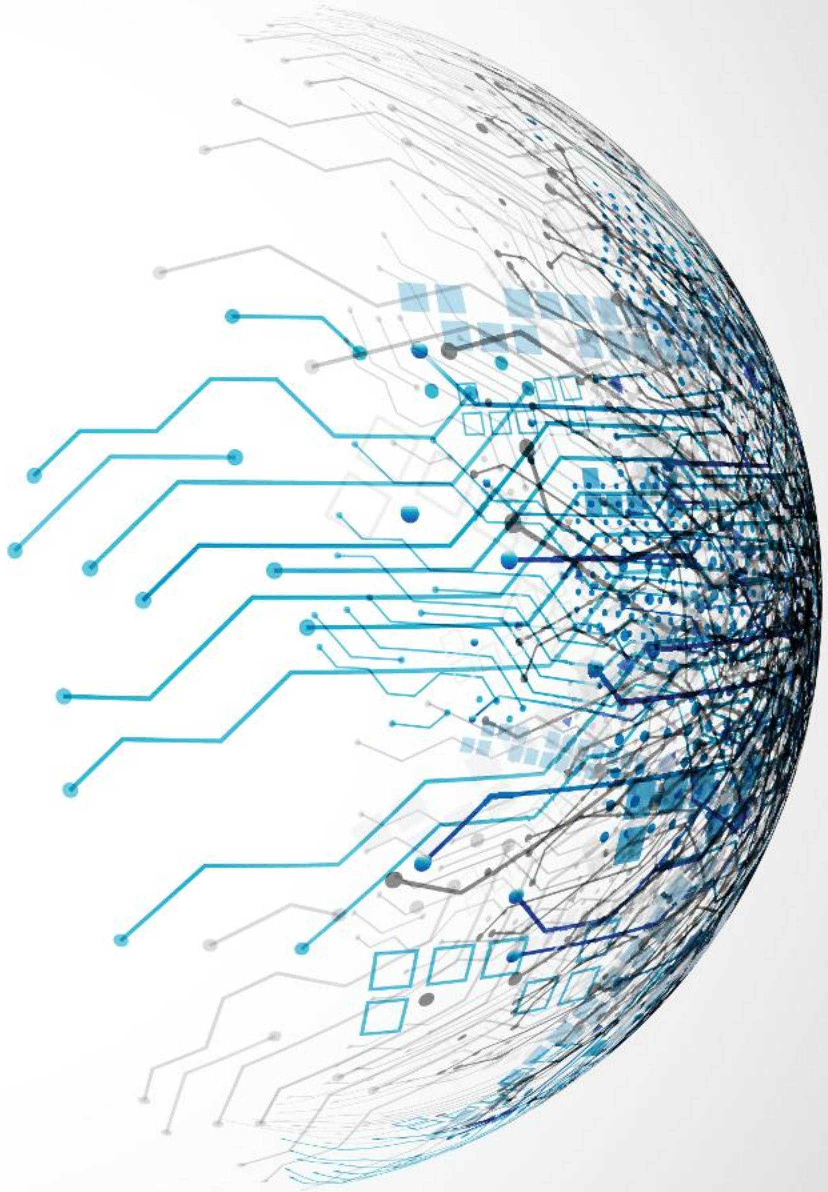
Deep Learning

End-to-End Automatic Speech Recognition

Dr. Mohammed Salah Al-Radhi

(slides by: Dr. Peter Mihajlik)



Copyright

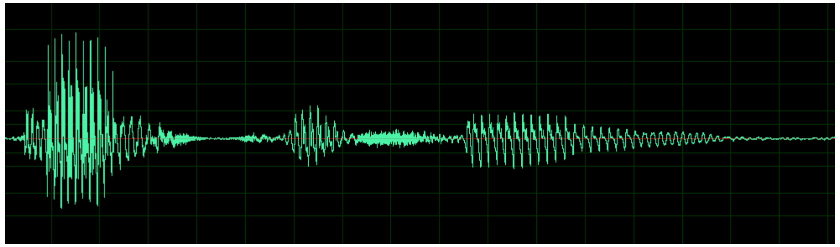
Outline

- 1. Introduction to ASR (Automatic Speech Recognition)
- 2. Speech-To-Text (STT) as a Seq2Seq task
- 3. Audio feature extraction
- 4. Training/Pre-training + Fine tuning
- 5. SOTA Architectures, results
- 6. Tools/Practice with NVIDIA NeMo

Introduction to ASR

What is ASR?

- **Speech-To-Text** (**STT**): acoustic pressure(time) signal) → text transcription



→ „I think …”

- Speaker recognition/diarization/verification
- Speech diagnostics
- Speech emotion recognition
- Etc.

} Auxiliary information

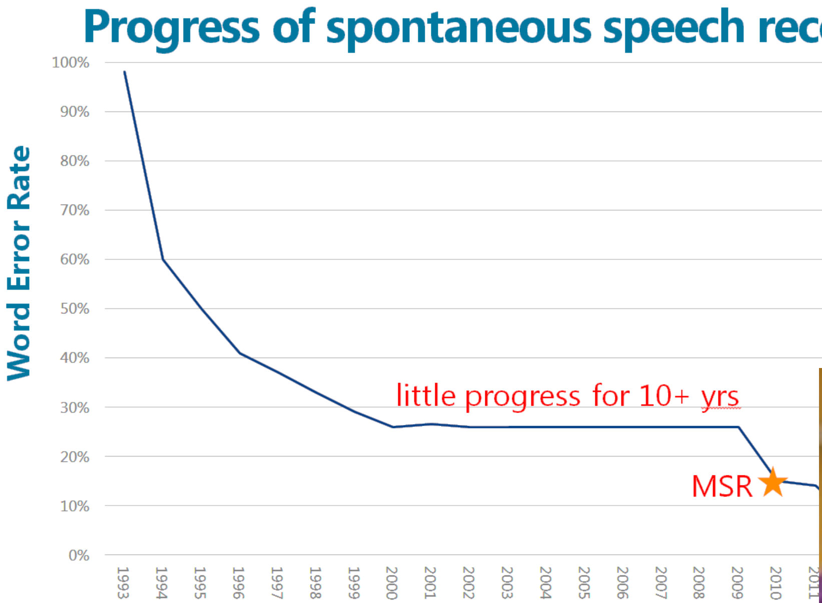
How is ASR related to deep learning?

Early attempts: limited success

- 1952, Bell Lab, Audrey
- 1961, IBM, Shoebox

First practical ASR systems: statistics and machine learning (HMM, GMM)

- 1975 CMU, IBM, … (Baker, Bahl, Jelinek) – 2010



Breakthrough: 2011, Florence Interspeech

Microsoft „Rosetta-stone”

Deep Neural nets for acoustic modeling

Frank Seide et al

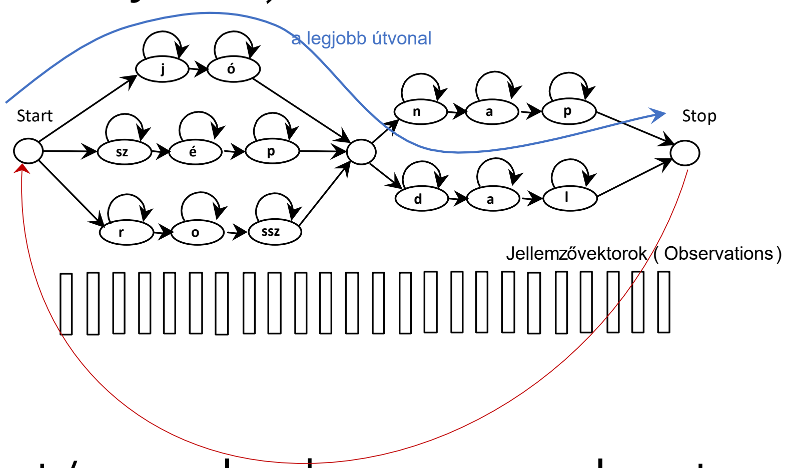
Deep learning is ASR

- „Classic” approach (HMM-DNN hybrid)

text

Audio
+ transcription

$$\hat{W} = \arg\max_{W} P(W)\cdot P(O\,|\,W)$$



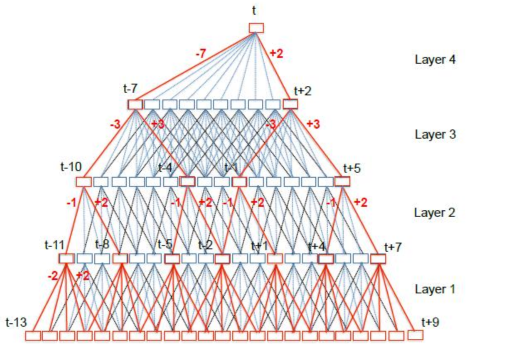
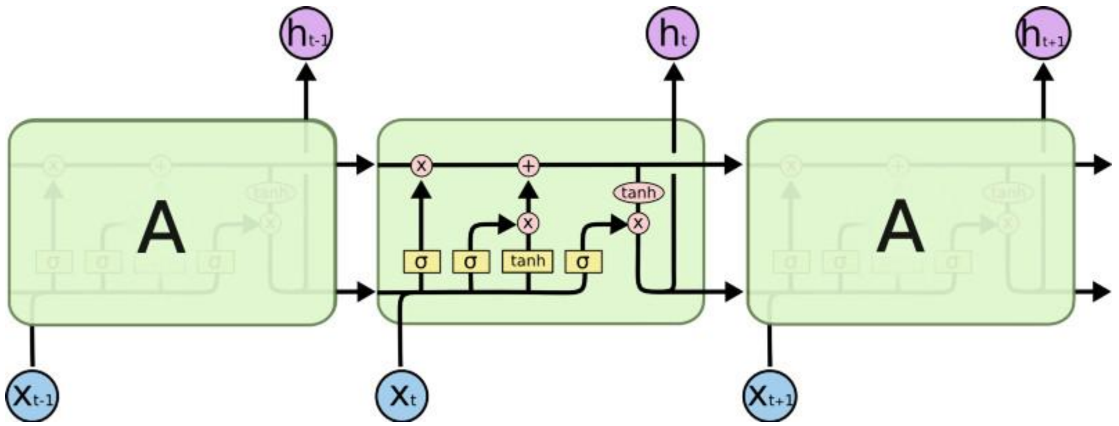


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

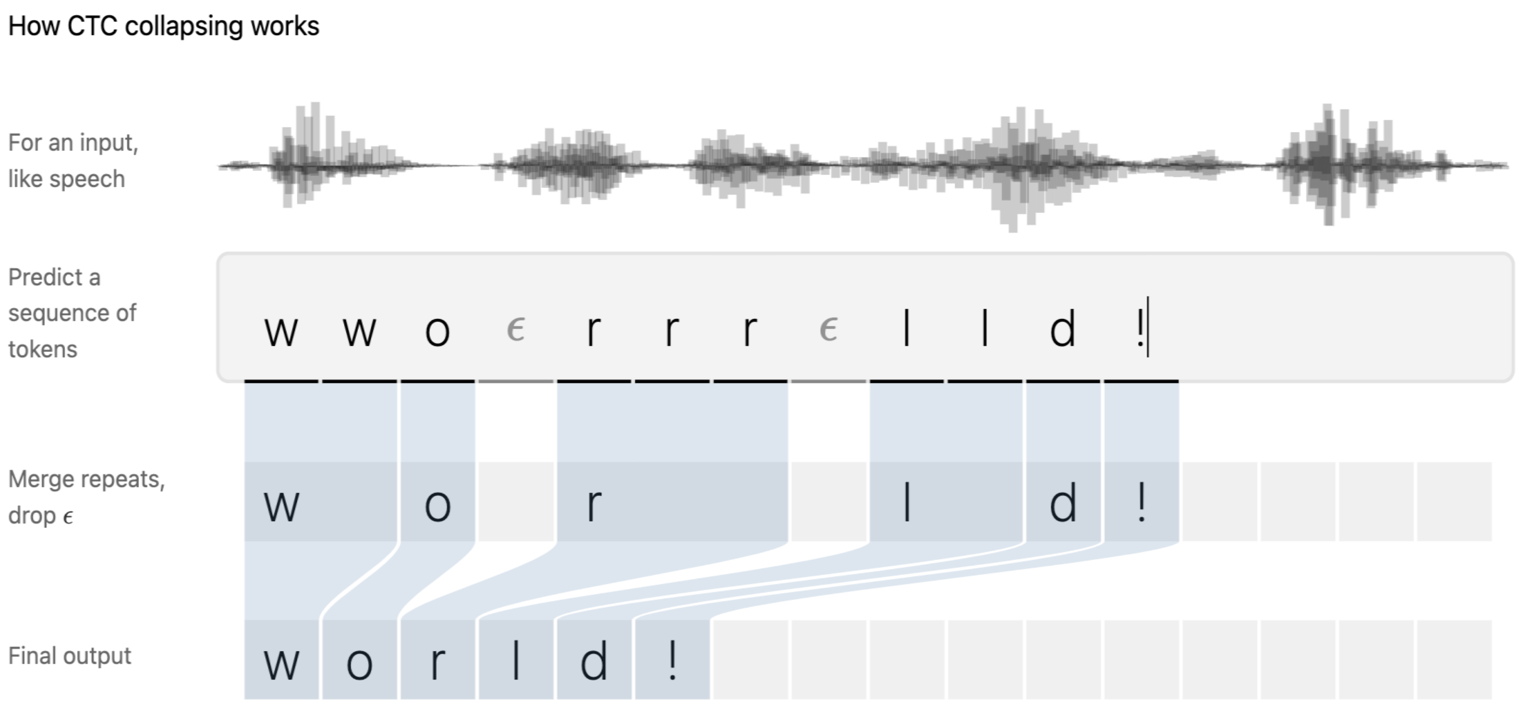
- End-to-end approach (/almost/ purely deep neural network)



Speech-To-Text (STT/ASR) as a Seq2Seq task

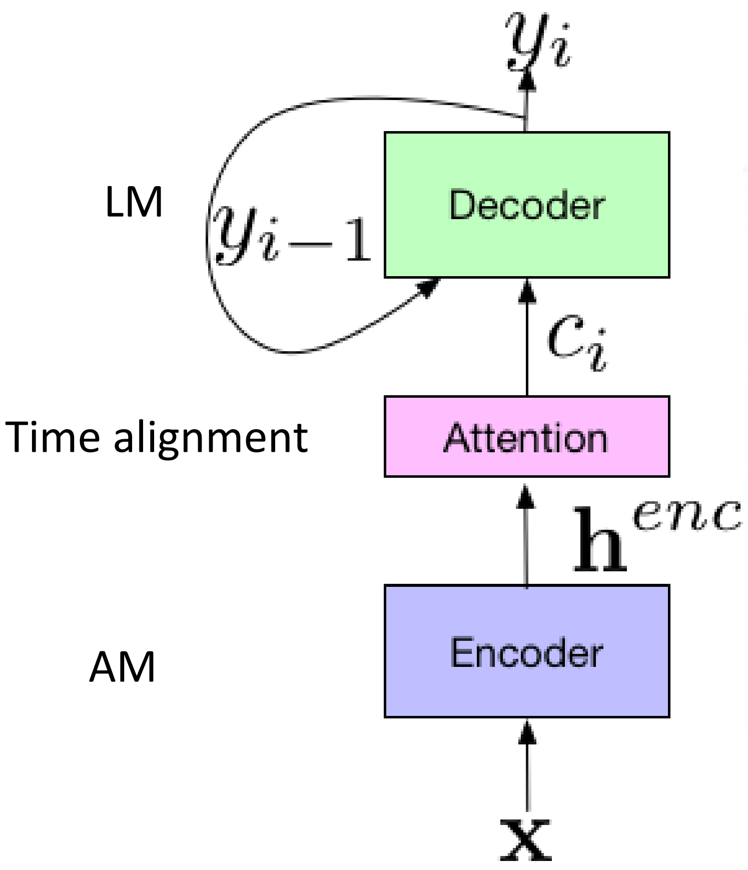
ASR with S2S

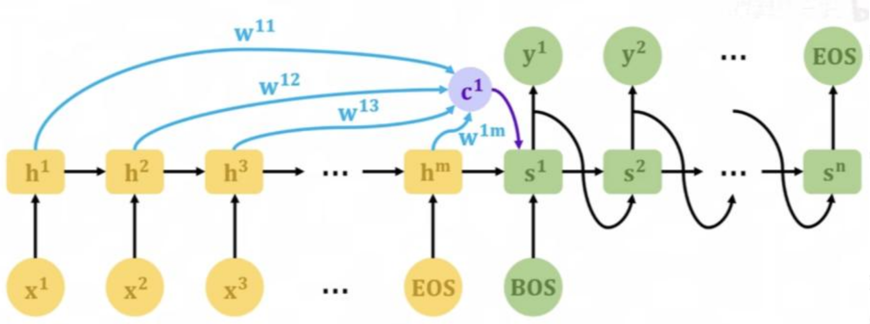
- Encoder-Decoder structures
  - Conv1D (NVIDIA Jasper/QuartzNet)
  - RNN (DeepSpeech)
  - Transformer (SOTA: Google Conformer, META wav2vec2, OpenAI Whisper)
- Special loss function: CTC
  - Why we need it:
    time alignment problem



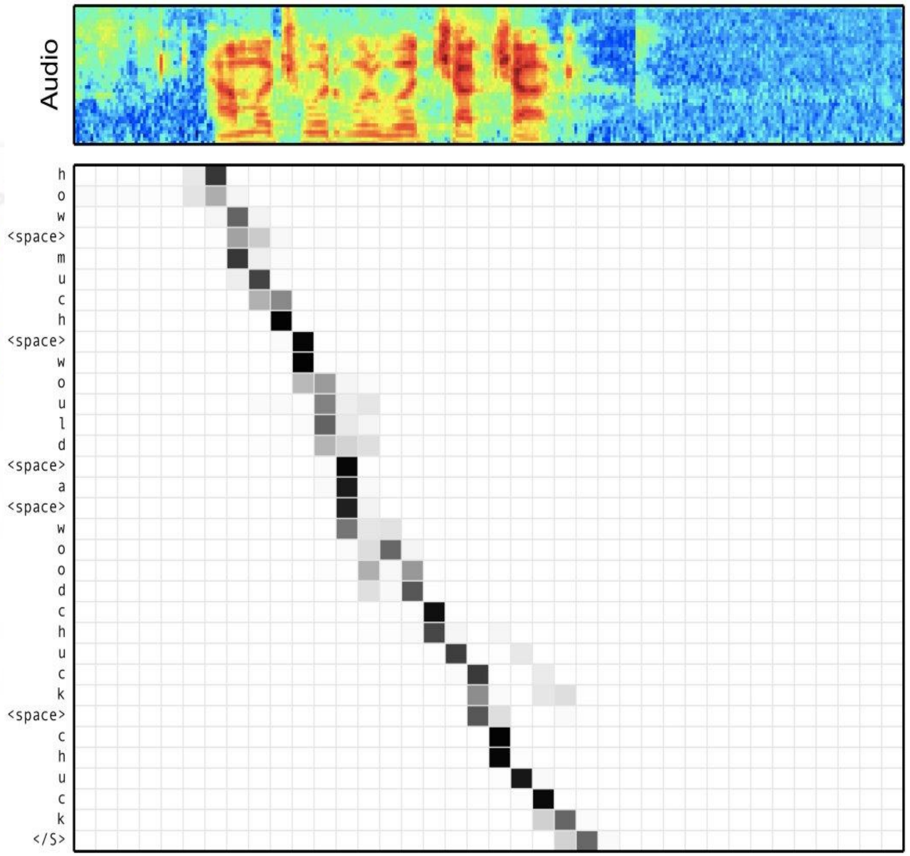
Listen-Attend-Spell (LAS) model

- Alternative to CTC alignment ➡ best practice: attention + CTC (Watanabe et al)





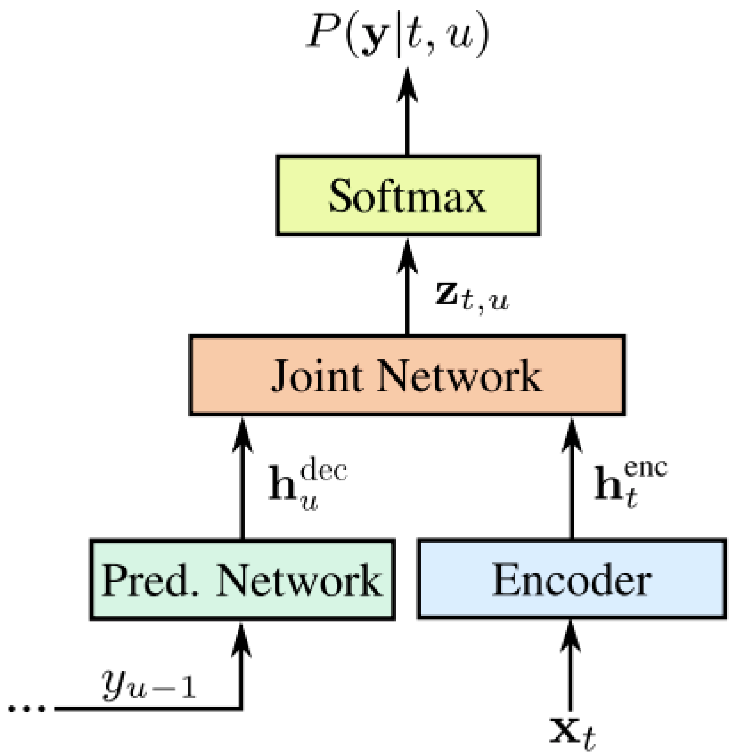
$$\mathbf{c}^1 = \mathbf{h}^1 \cdot \mathbf{w}^{11} + \mathbf{h}^2 \cdot \mathbf{w}^{12} + \dots + \mathbf{h}^m \cdot \mathbf{w}^{1m}$$



T structures

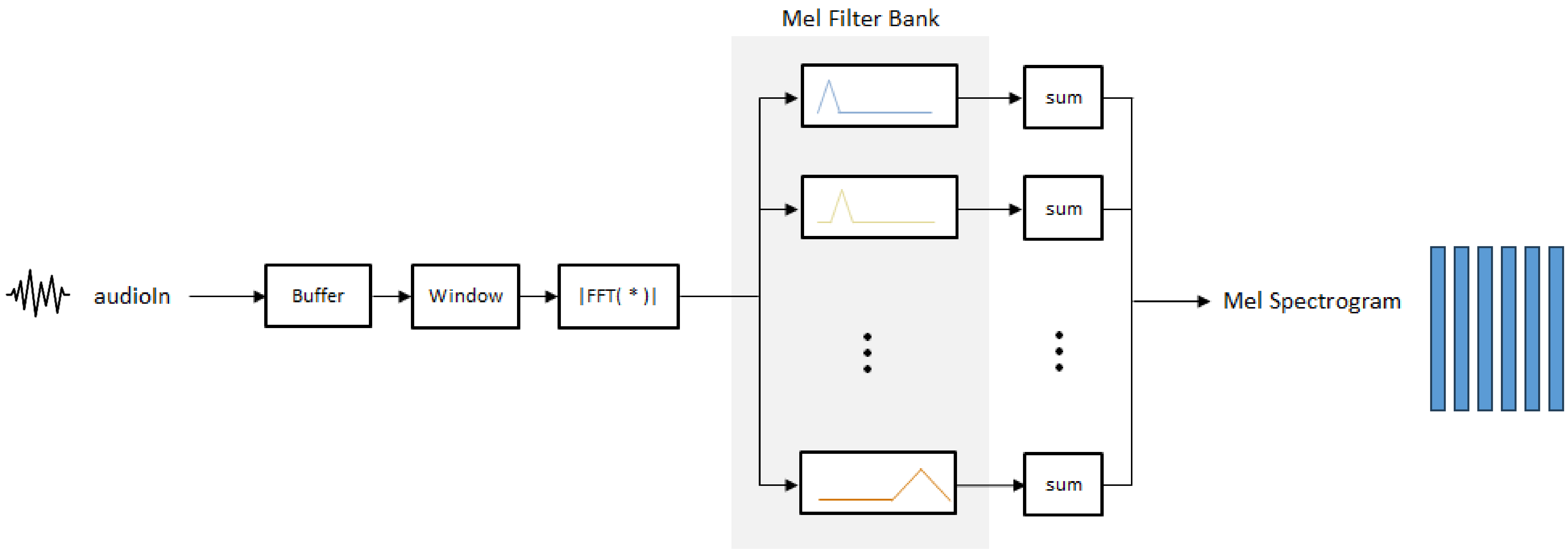
RNN-T (Alex Graves): RNN-Transducer

Conformer-T, Transformer-T, …

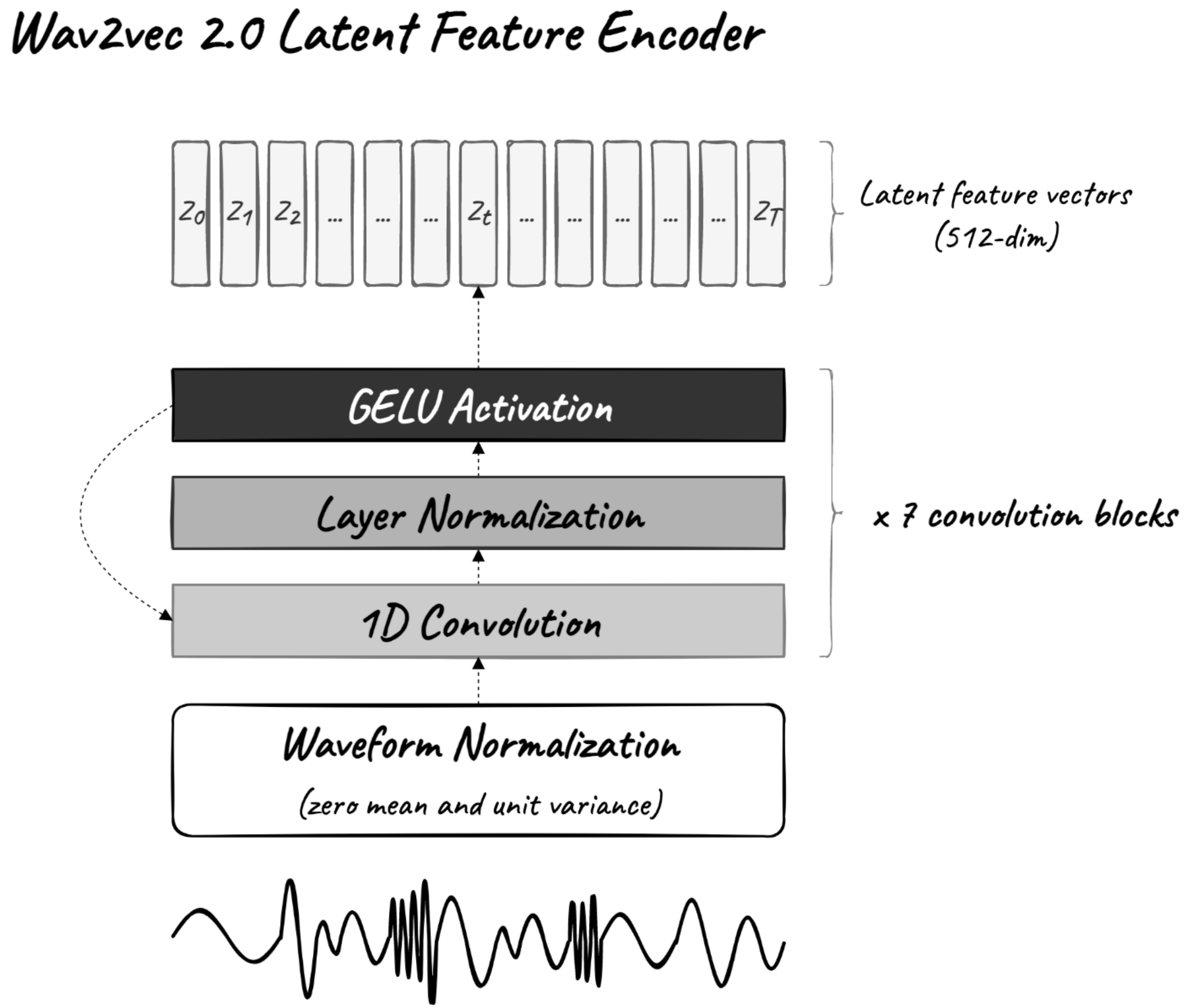


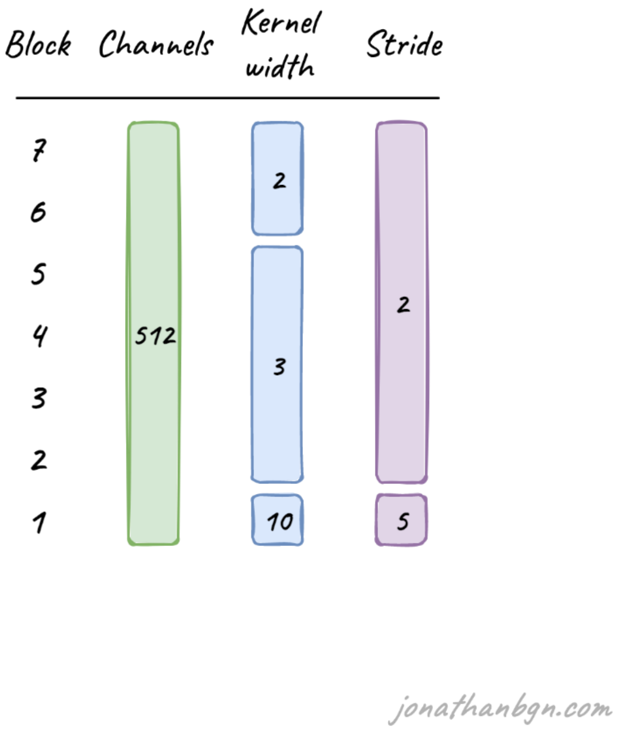
Audio Feature Extraction

Mel-Spectrogram



Feature extraction with Conv1D





https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html

ASR training

Why and how?

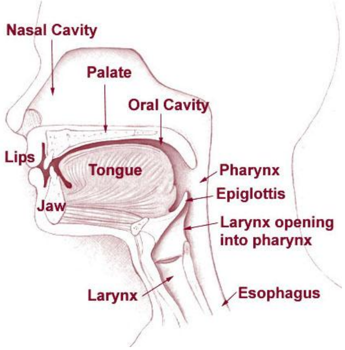
- https://huggingface.co/
- https://catalog.ngc.nvidia.com/
- https://modelzoo.co/
- …

Hugging Face

NVIDIA

M

- Multitude of pre-trained models
- Just not what you really need …
- Training from scratch?
  - At least 200 hours but 2k or 20k performs significantly better
  - Needs lot of GPU's (multi GPU, multi node – PyTorch DDP at least…)
- Transfer learning?! Yes!



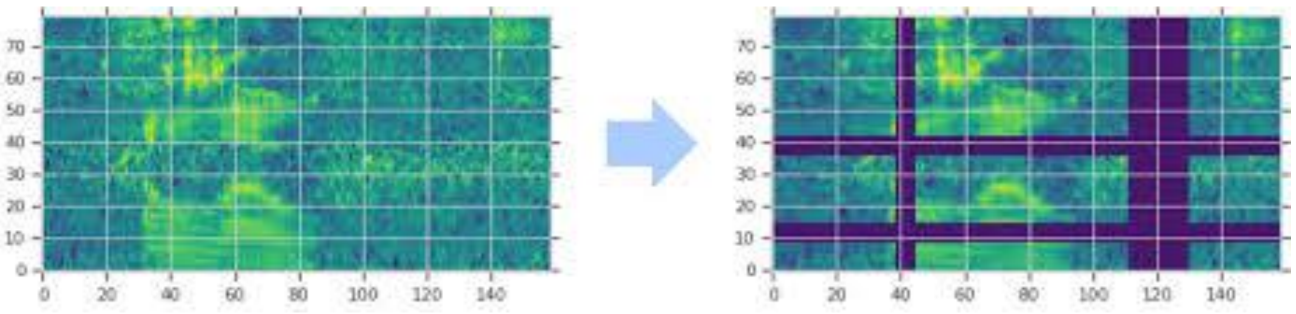
- Always use pre-trained model + fine-tuning!
  - Data efficient: fine-tuning works even for 1 hours of supervised data

(Pre-)/(post-) training

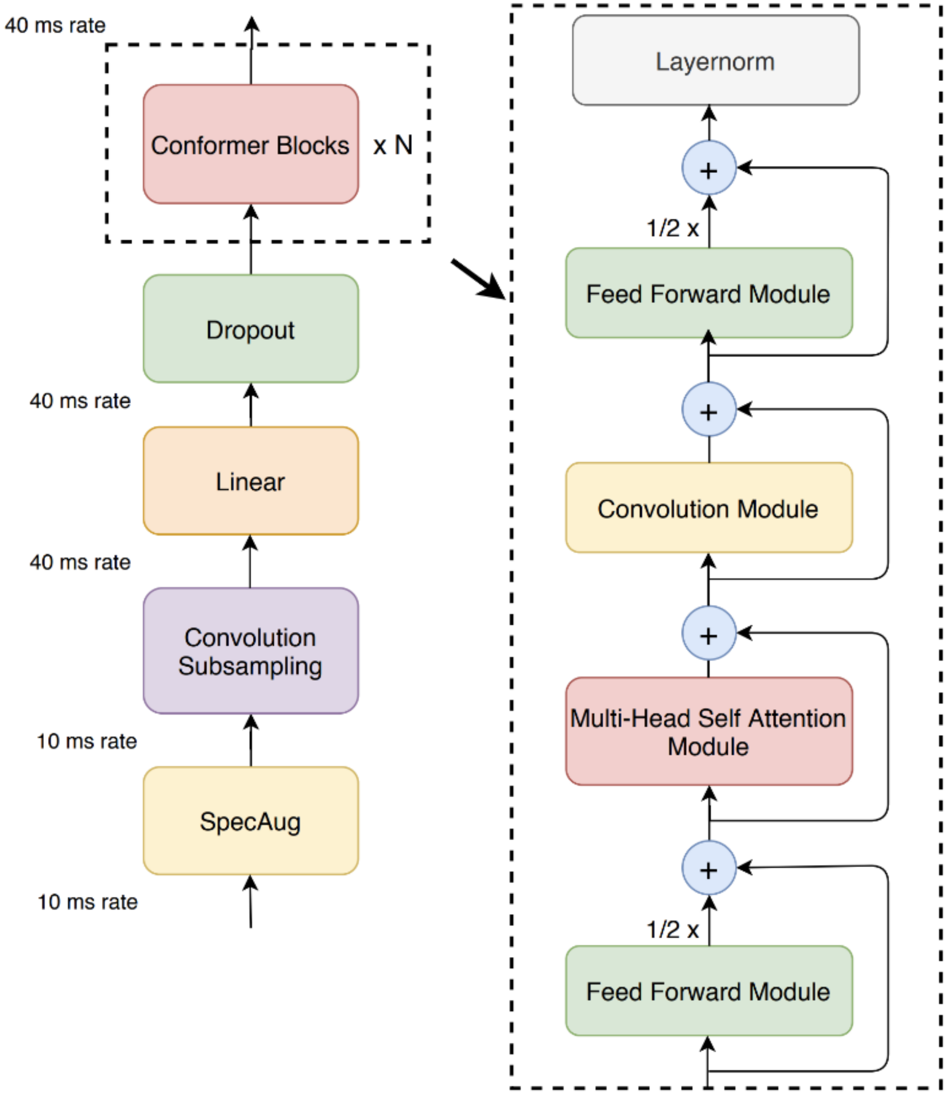
- Pre-training:
  - Supervised (audio + exact transcription)
  - Weakly supervised (audio + edited/simplified transcription)
  - Self-supervised (audio only!)

- „Post training": noisy student – training on ASR pseudo labels

- „ASR" augmentations:
  - SpecAugment (2019)
  - Speed perturbation
  - …



State of The Art



Conformer

/as encoder with a light decoder/

Transformer,
Self-attention + Convolution + FF

Google

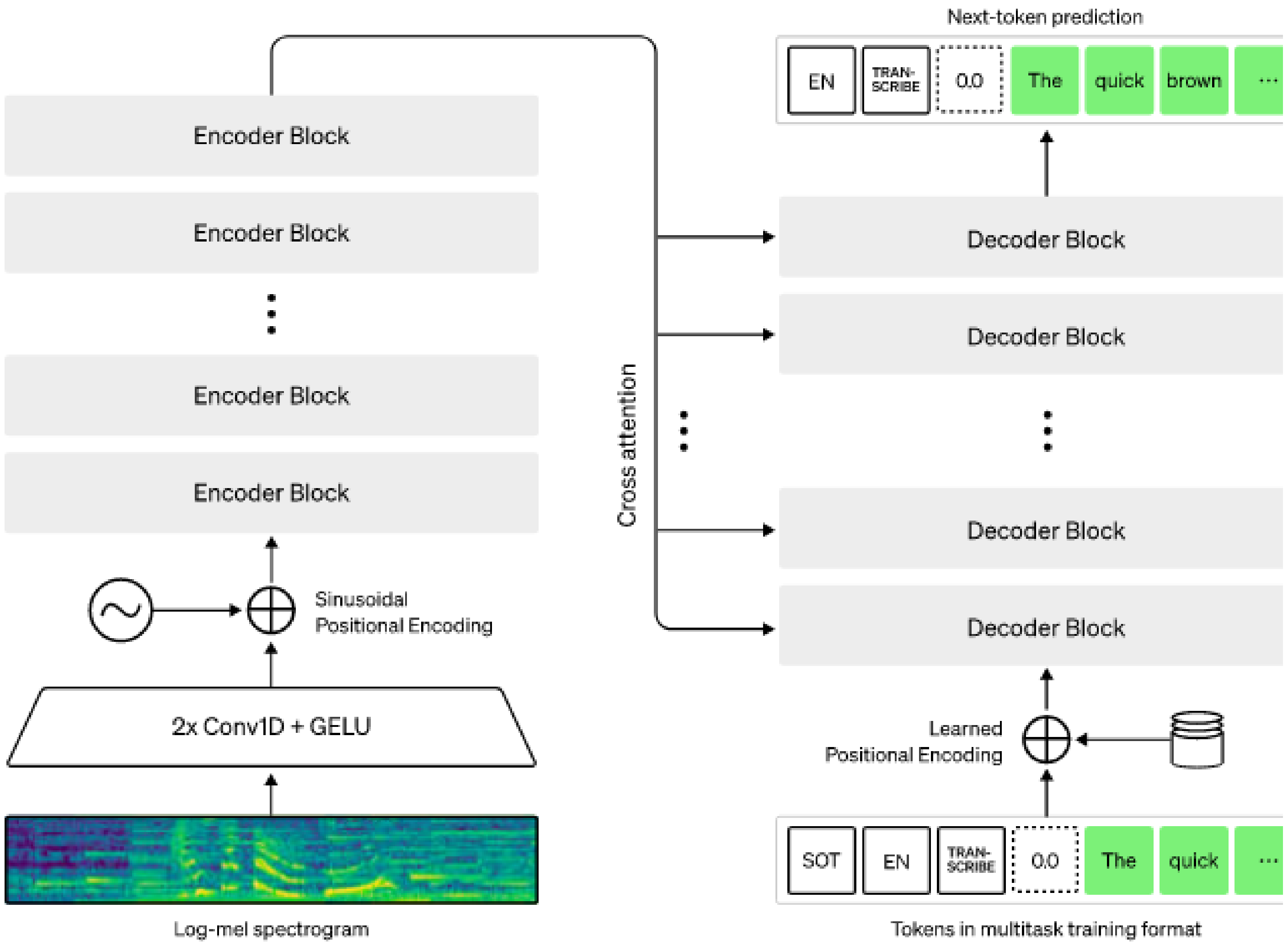
Whisper: Encoder + decoder transformer

Weak supervision =
not exact transcriptions

- Multilingual, multitask learning
- Language ID
- Punctualization

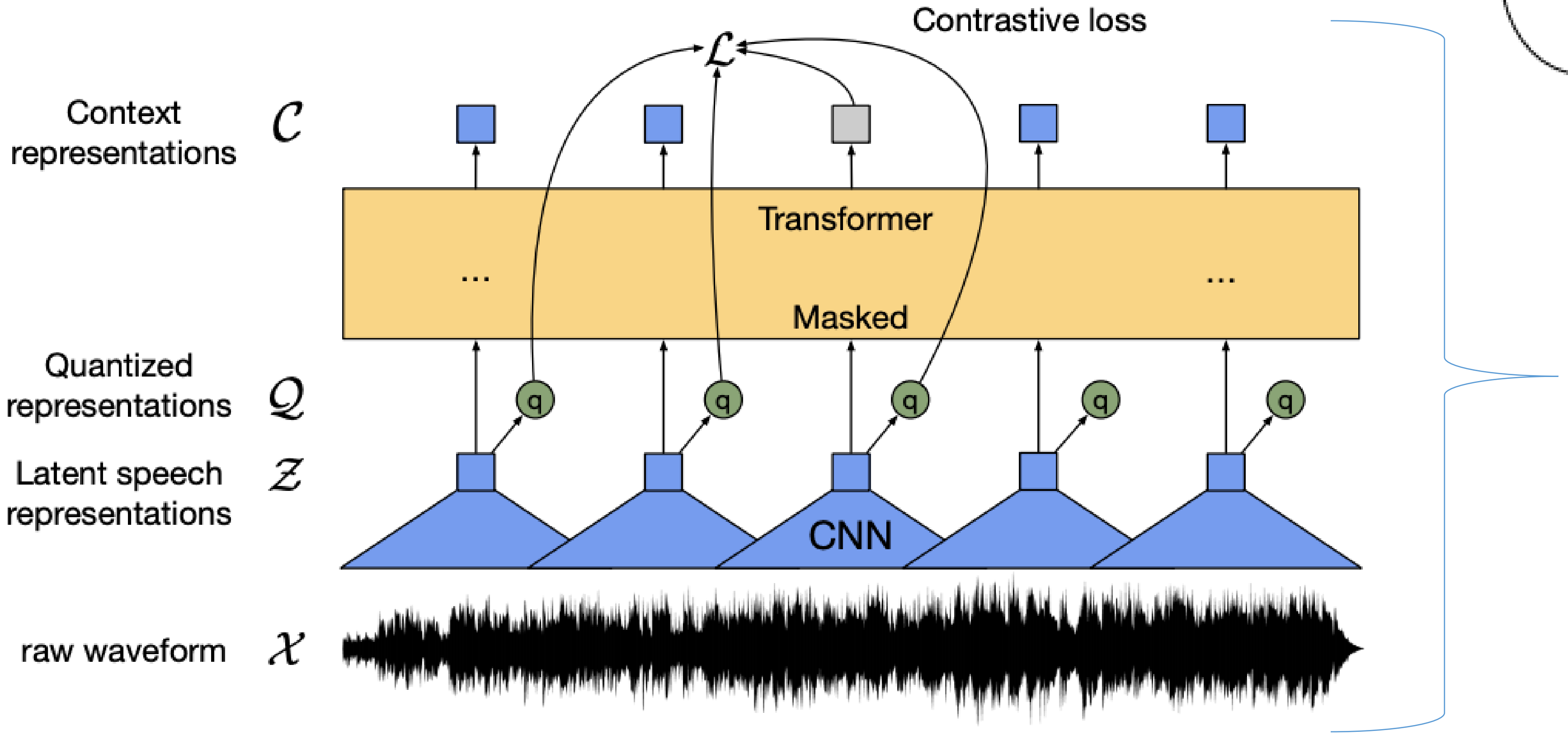
BUT

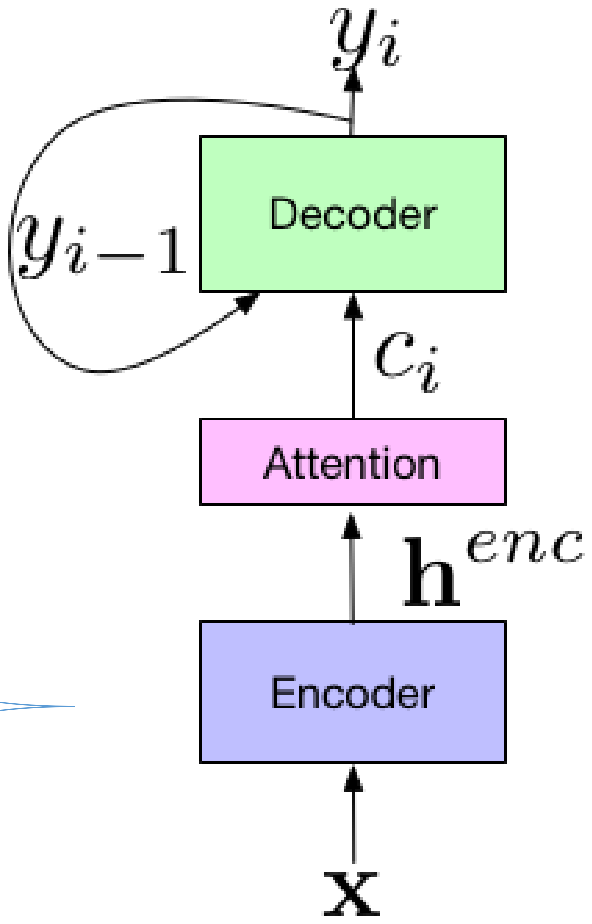
- Slowww…
- Non-streamable
- „Input audio is split into 30-second chunks”→ latency
- Non-English Accuracy?



Wav2vec2.0: transformer encoder

Self-supervised pre-training: no labels are needed





Case-study

Hungarian ASR on studio quality spontaneous and read/repeated speech

Experimental data (Hungarian)

Table 1: *Main characteristics of data sets used in the experiments.*

	HGC SPOK	train-114	dev-repet	BEA-Base dev-spont	eval-repet	eval-spont	CV test
Length [hours]	-	71.2	0.65	4.02	0.95	4.91	6.8
Num of speakers	-	114	10	10	16	16	220
Num of segments	-	76 881	568	4 893	858	5 693	4 871
Num of characters	516.84M	3.1M	28 467	154 994	43 448	197 738	250 709
Num of words	56.13M	0.56M	4 110	27 939	6 229	35 178	35 485
3-gram PPL	-	-	924	771	846	857	2 387
OOV rate [%]	-	-	1.6	1.9	1.4	1.7	3.1

LM training

AM training

Evaluation

Evaluation Metrics

- Accuracy

English/Hungarian : Word Error Rate (WER)
Mandarin: Character Error Rate (CER)

$$WER(CER) = \frac{S + D + I}{N} * 100\%$$

S: Substitution. D: Deletion I: Insertion C: Correct
N: Total numbers of characters, N = S + D + C

$$RTF = \frac{T_{process}}{T_{audio}}$$

- Real-time Factor (RTF)

If RTF < 1, indicating the system can transcribe faster than real-time.

N-gram Language Model (LM)

- **Language Model (LM)**

  A **statistical model** used to predict the **likelihood of a sequence of words.**

- **N-gram**

  An N-gram refers to a continuous sequence of **N items**.

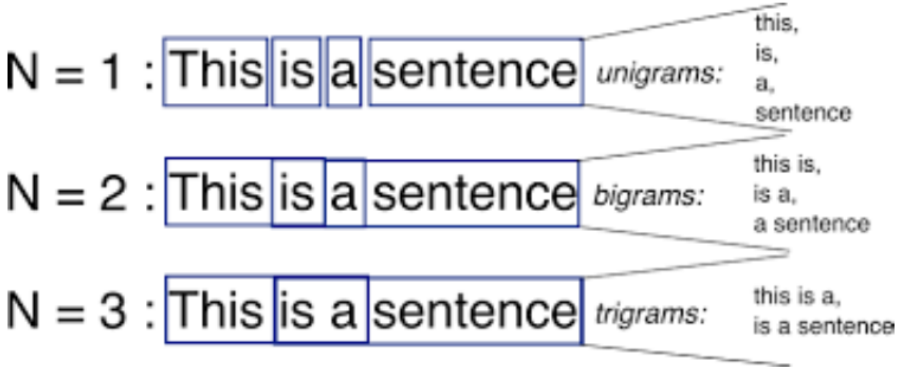


Fig 4. The example of N-gram LM

- **N-gram LM**

  Transcribe spoken words by predicting the **most likely word sequences**.

Training from scratch vs. cross-lingual transfer learning

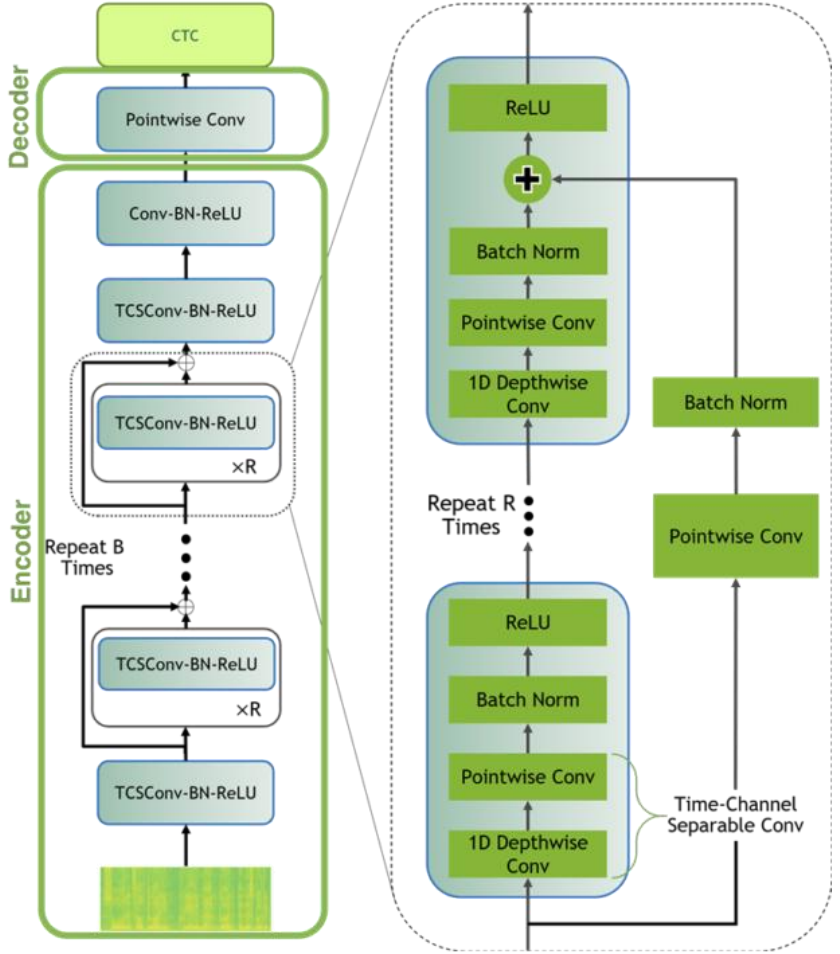
Conformer – NVIDIA NeMo implementation, supervised pre-training

From scratch, supervised end-to-end ASR with convolutional/Conformer acoustic models

WER[%] results on BEA-Base (starting from scratch)

Structure/ num of parameters	LM	eval-repet	eval-spont	CV
QuartzNet	–	11.56	26.70	–
15x3 / 12.7M	3-gram	6.86	26.83	–
Conformer-	–	12.73	25.31	49.8
Small / 13M	6-gram	7.98	22.78	**42.7**
Conformer-	–	10.98	24.93	49.8
Medium / 30M	6-gram	**5.65**	**21.01**	42.9

QuartzNet (baseline)



Supervised pre-training* (En) + fine-tuning (Hu)

QuartzNet (baseline) 15x5 (18M) based transfer learning WER[%] results on BEA-Base

Pre-training data size [hours]	LM	eval-repet	eval-spont	CV
3k	–	10.63	**24.87**	–
	3-gram	**5.83**	25.23	–

Conformer Small (13M) / Large (121M) transfer learning WER[%] results on BEA-Base

Pre-training data size [hours]	LM	eval-repet	eval-spont	CV
~10k	–	11.22	21.39	40.8
	6-gram	4.96	17.77	34.8
	–	5.2	17.24	34.8
	6-gram	**3.66**	**16.25**	**30.8**

End-to-end deep learning approach – **weakly-supervised** training

Whisper in zero-shot and fine-tuning setups

Whisper results: zero-shot vs. fine-tuning

- (Pre-)training data size: 680k hours
- Num of languages: 97
- Composition: 83% English … 0.03% Hungarian

Whisper Medium/Large_v2 WER [%] results of BEA-Base

Model	Fine-tuning	Num of parameters	eval-repet	eval-spont	CV
Whisper-Medium	–	769M	22.33	38.67	27.6
Whisper-Large	–	1550M	18.04	32.76	**20.4**
Whisper-Medium	Decoder (456M)	769M	4.90	20.60	27.9
Whisper-Large	Decoder (906M)	1550M	**4.37**	**18.69**	23.7

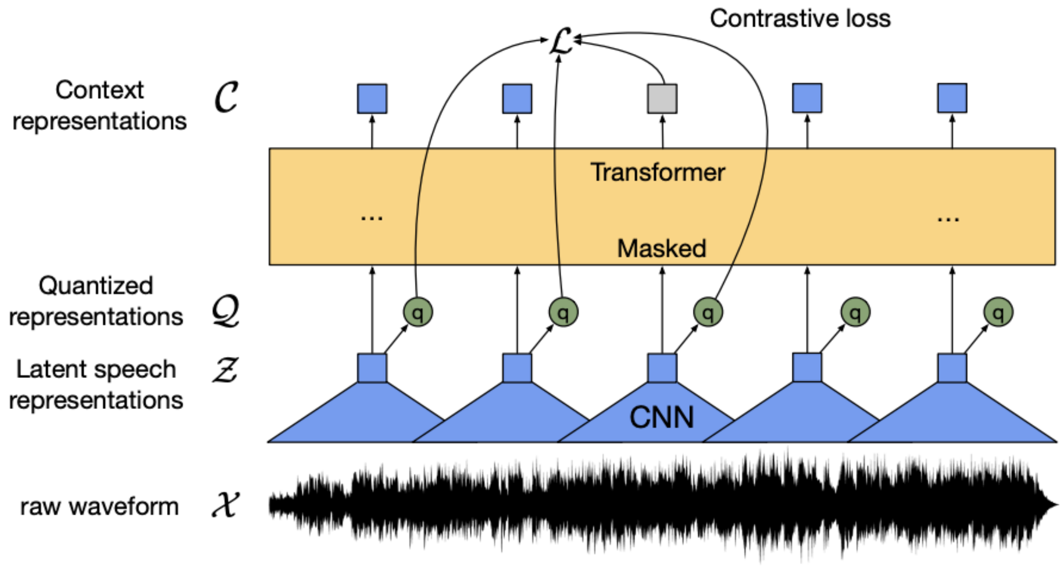
End-to-end deep learning approach – **self-supervised** pre-training

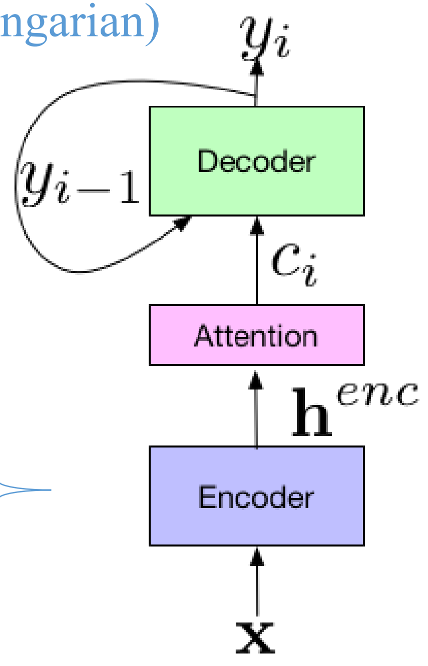
Transcription-free SSL pre-training + wav2vec2 encoder + attentional decoder

Self-Supervised Pre-training based Transfer Learning

- SSL pre-training
- All pre-trained models are downloaded from HuggingFace
  - wav2vec2-large-lv60: LibriVox (English)
  - wav2vec2-large-xlsr-53: CommonVoice + BABEL + Multilingual LibriSpeech
  - wav2vec2-xls-r-300m: CV + BABEL + MLS + VoxPopuli + VoxLingua107 (0.04% Hungarian)
  - wav2vec2-mms-300: MMS-lab-U + VoxPopuli + … (Massively multilingual incl. Hungarian)
  - wav2vec2-uralic: VoxPopuli (3 languages, 41% Hungarian)

- Units: BPE (600)
- Loss: CTC + NLL
- Decoder: GRU
- Encoder: wav2vec2.0-large, 300M
- LM: –/Transformer





Wav2vec2 (SSL)-based Transfer Learning Results

wav2vec2-large+GRU+CTC+Attention+BPE_600 based transfer learning WER[%] results on BEA-Base

Model	SSL Pre-training languages	Pre-training data size [hours]	LM	eval-repet	eval-spont	CV
wav2vec2-large-lv60	1*	53k	–	8.46	19.17	36.5
wav2vec2-large-xlsr-53	53	56k	–	5.81	16.62	34.2
wav2vec2-xls-r-300m	128	440k	–	6.16	15.61	30.5
wav2vec2-mms-300	1406	491k	–	6.56	18.82	34.9
wav2vec2-uralic	3**	42k	–	4.24	11.55	21.3
			Transformer	**2.42**	**10.50**	**17.2**

* = English
** = Estonian + Finnish + Hungarian

Wav2vec2 (SSL)-based Transfer Learning Results

wav2vec2-large+GRU+CTC+Attention+BPE_600 based transfer learning CER[%] results on BEA-Base

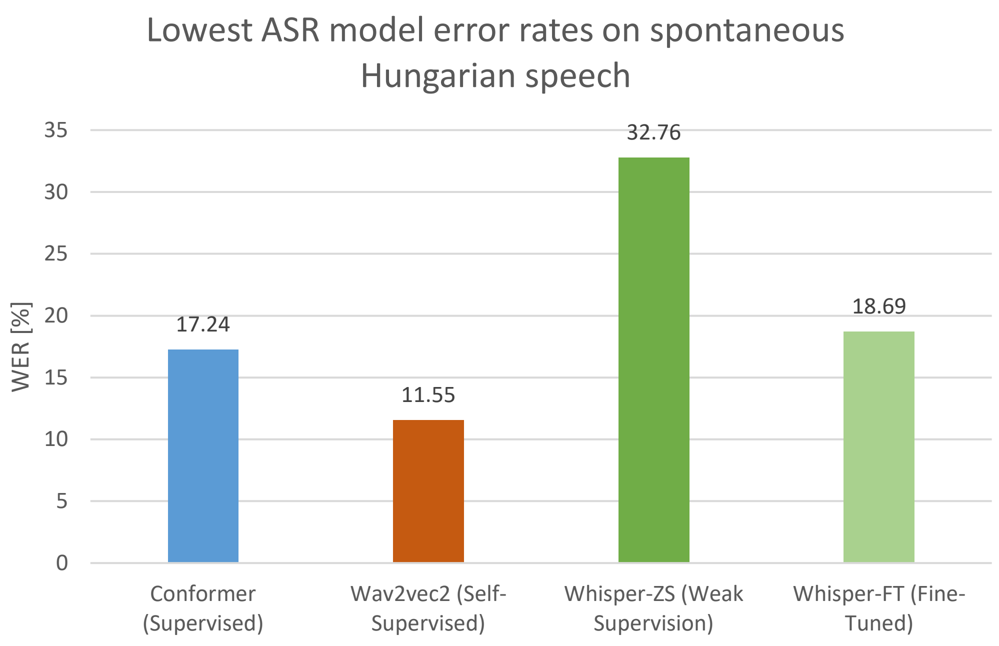
Model	SSL Pre-training languages	Pre-training data size [hours]	LM	eval-repet	eval-spont	CV
wav2vec2-large-lv60	1*	60k	–	2.6	5.9	11.2
wav2vec2-large-xlsr-53	53	56k	–	2.1	5.5	10.5
wav2vec2-xls-r-300m	128	440k	–	2.4	5.1	8.6
wav2vec2-mms-300	1406	491k	–	2.2	5.8	9.1
wav2vec2-uralic	3**	42k	–	1.7	3.7	5.8
			Transformer	**0.7**	**3.3**	**4.5**

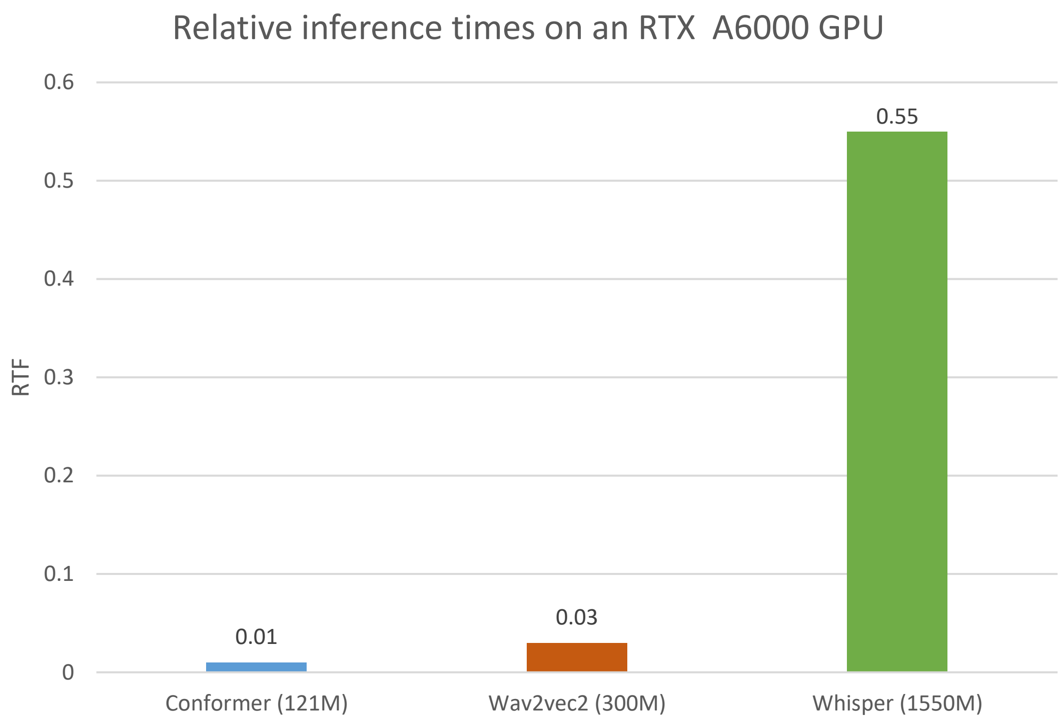
* = English
** = Estonian + Finnish + Hungarian

Final comparison – with respect to RTF

Best of Conformer vs. Wav2vec2 vs. Whisper

Best ASR results on spontaneous Hungarian (without LM) vs. inference times





/Word Error Rate: the lower the better/

/Real-Time Factor: the lower the better/

Recommended end-to-end ASR tools

- https://github.com/espnet/
- https://github.com/facebookresearch/fairseq
- https://github.com/k2-fsa/k2
- https://github.com/lhotse-speech/lhotse
- https://speechbrain.github.io/
- https://github.com/openai/whisper
- https://github.com/NVIDIA/NeMo
- https://github.com/wenet-e2e

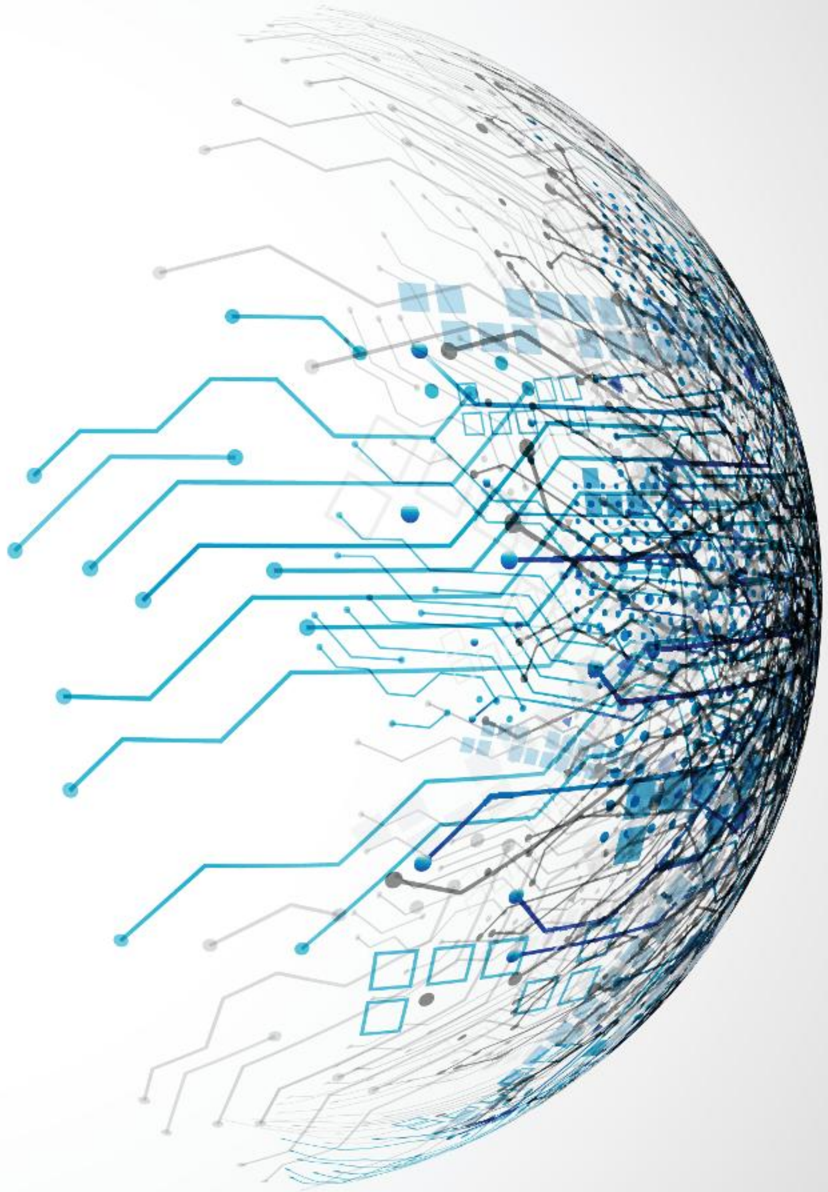
References

References

- Stanford Lecture on ASR
- "An Intuitive Explanation of Connectionist Temporal Classification"
- Explanation of CTC with Prefix Beam Search
- Listen Attend and Spell Paper (seq2seq ASR model)
- Explanation of the mel spectrogram in more depth
- Jasper Paper
- QuartzNet paper
- SpecAugment Paper
- Explanation and visualization of SpecAugment
- Cutout Paper
- Transfer Learning Blogpost

Please, don't forget to send feedback: https://bit.ly/bme-dl

Thank you
for your attention



Dr. Mohammed Salah Al-Radhi
malradhi@tmit.bme.hu

(slides by: Dr. Peter Mihajlik)

05 November 2024