Felügyelt, önfelügyelt és gyengén felügyelt neurális technikák a magyar nyelvű beszédleiratozásban

Dr. MIHAJLIK Péter
mihajlik.peter@vik.bme.hu

TMiT

Speech Recognition Group

SmartLabs

Nyelvtudományi Kutatóközpont

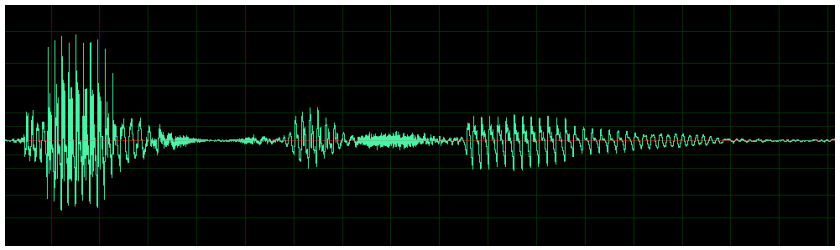
The evolution of ASR technology

A co-evolution with AI

mihajlik@tmit.bme.hu

Automatic Speech Recognition

- **Speech wave (acoustic time-pressure signal) → transcription (text)**



„I think …”

The beginning: electronic filters, rules-based algorithms

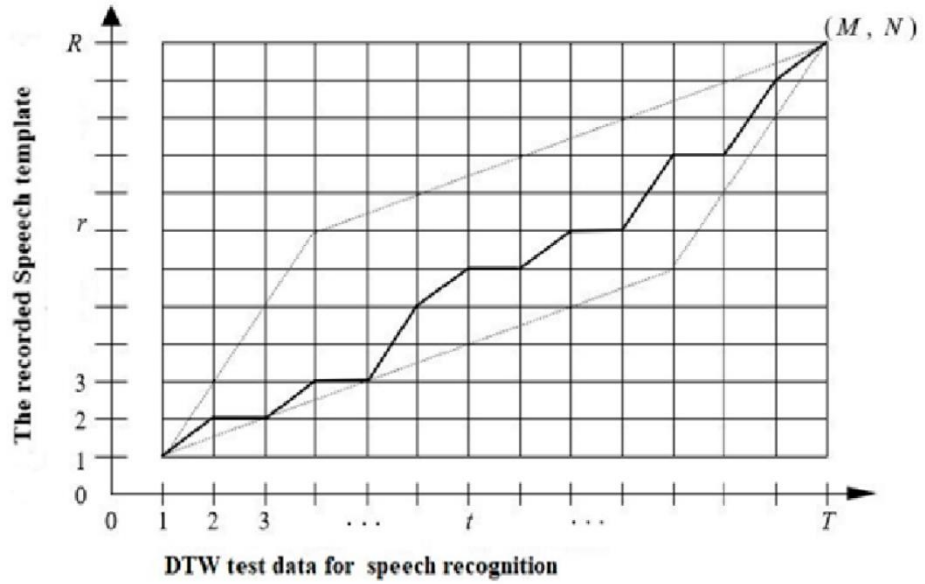
- 1950-52 Bell Laboratories:
  - Audrey (Automatic Digit Recognizer)
  - Numbers 1-9

- 1961 IBM
  - Shoebox
  - Numbers 0-9,
  - 6 basic arithmetic operations

!

Template based, isolated-word recognition

- From 1970
- Dynamic Time Warping



„apple”

„orange”

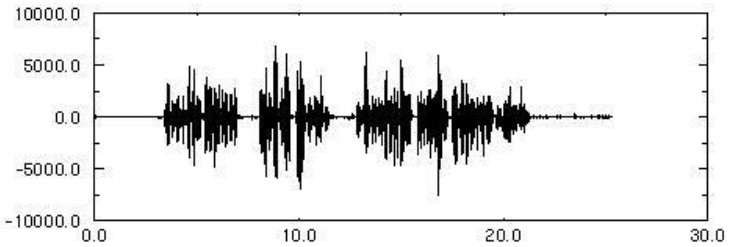
„ananas”

„nut”

Utterance unknown

Highest similarity?

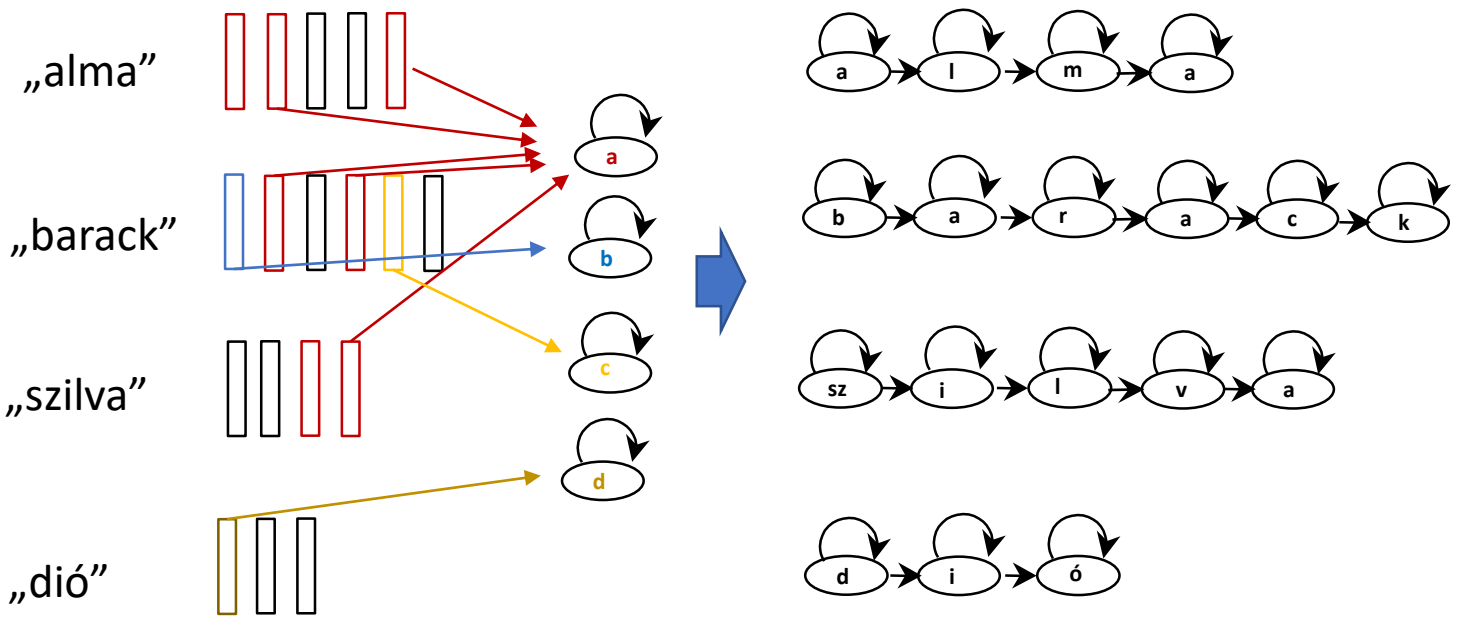
Acoustic feature extraction:



mihajlik.peter@vik.bme.hu

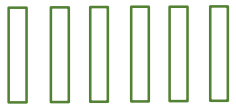
More data, phoneme-based ASR

- **Hidden Markov-modell** (HMM), from 1975…
- Similarity measure: by GMM



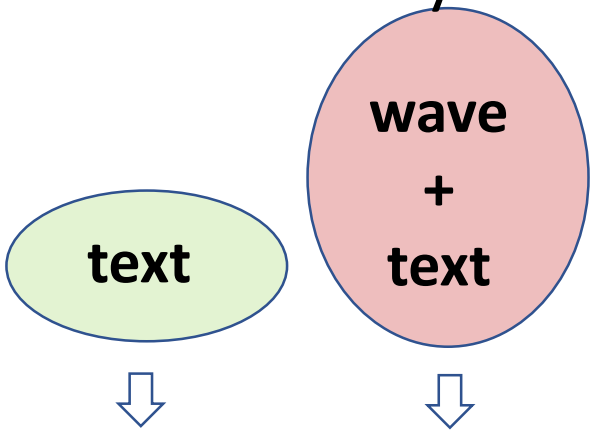
Utterance unknown

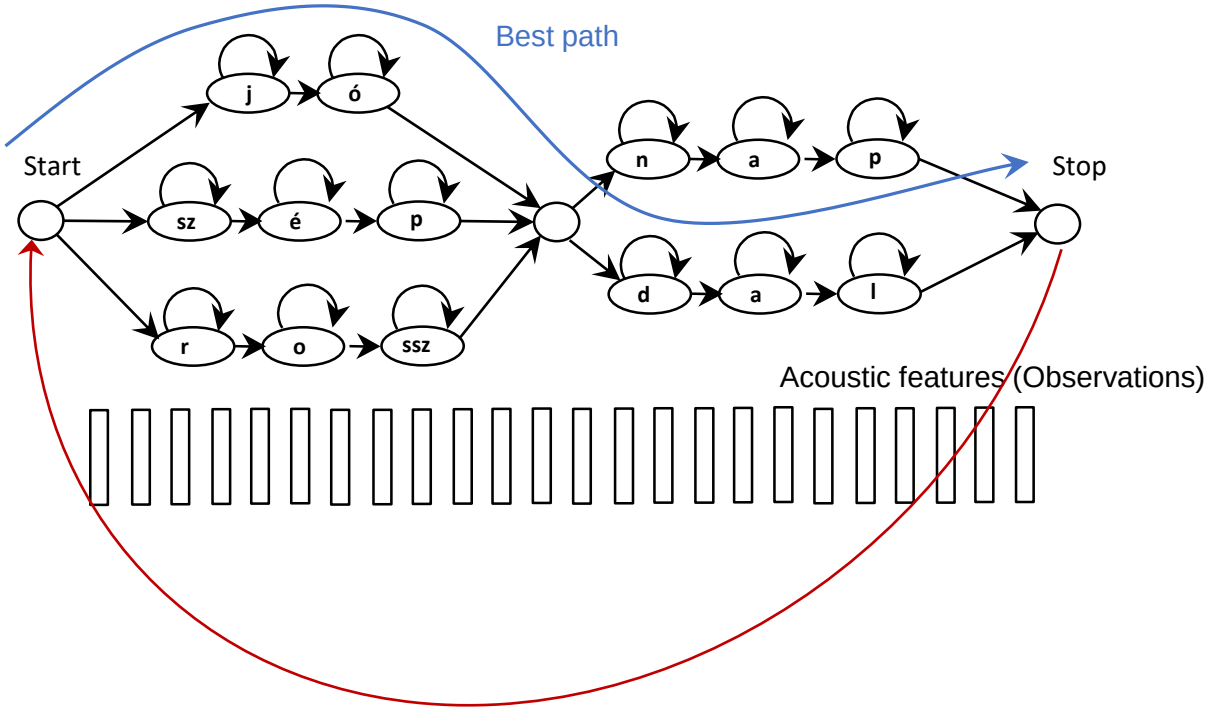
Highest similarity?



Adding text data and Language Model (LM)

- HMM: **Machine Learning in ASR**
- Data/statistics driven
- Pronunciation dictionary

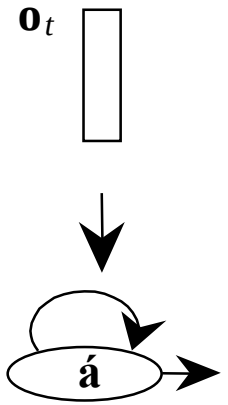


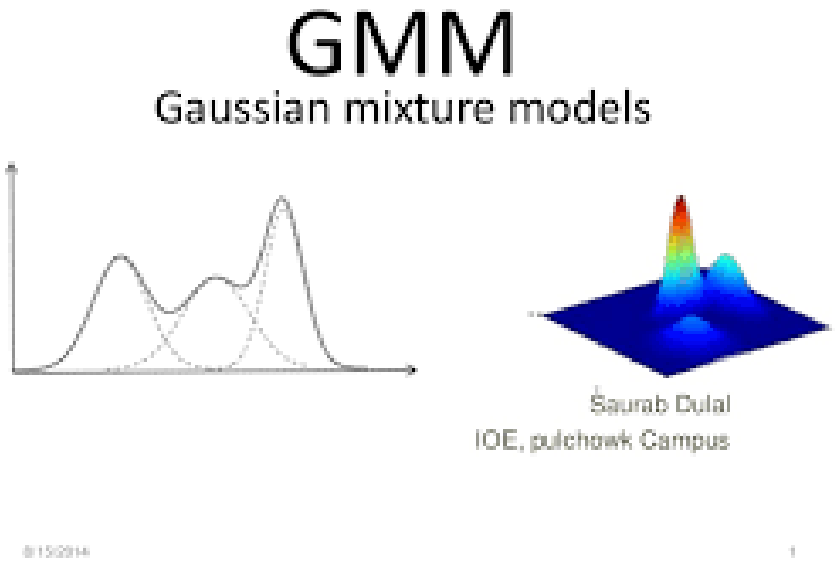


$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W)$$

Acoustic modeling

- Acoustic similarity measurement– based on the statistics of speech data





„The Deep Learning revolution”

2011 -

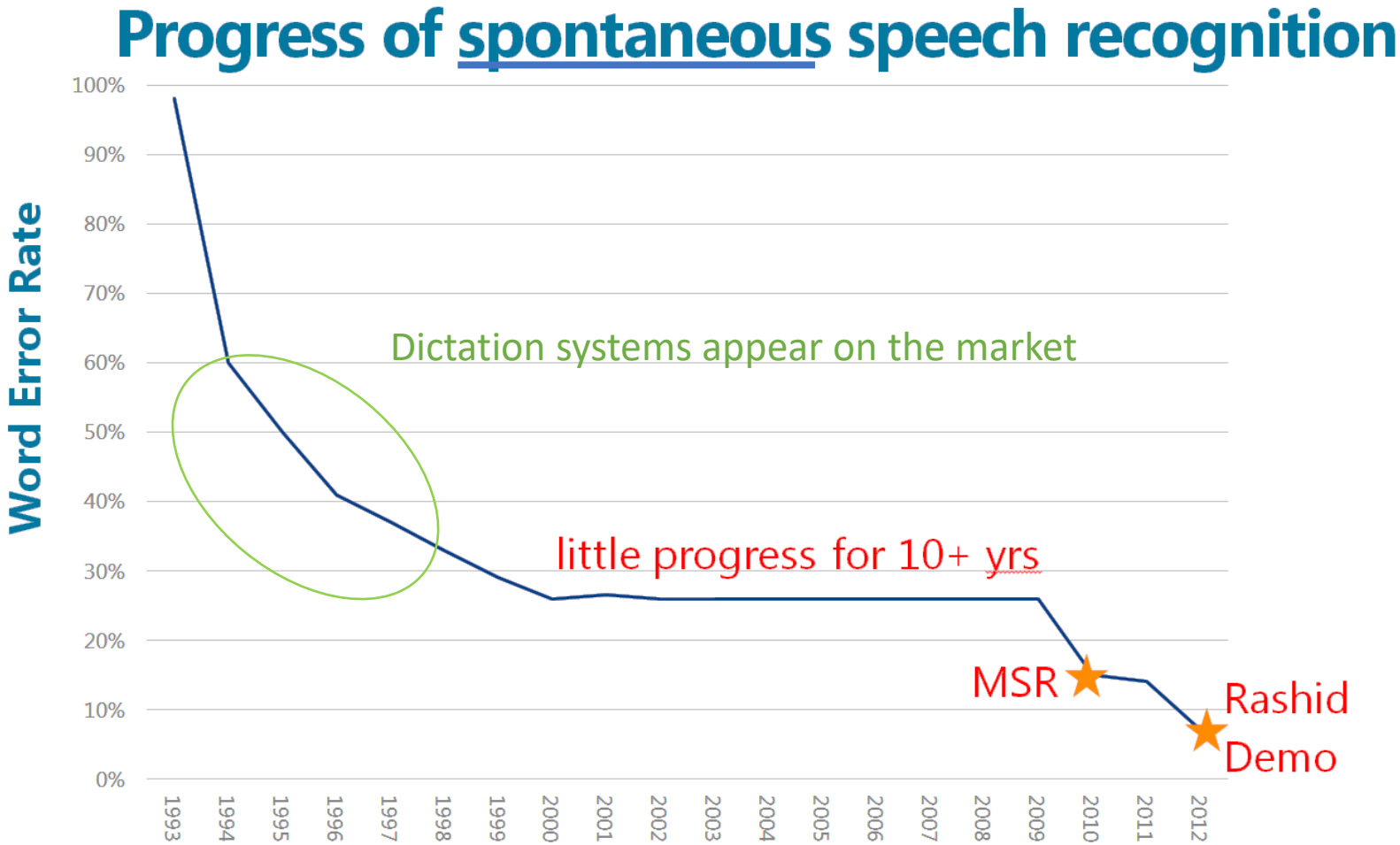
mihajlik.peter@vik.bme.hu

Microsoft and the rosetta-stone of ASR

After no improvement for 10+ years by the research community…

Dahl, Yu, Deng, and Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Trans. ASLP,* Jan. 2012 (also ICASSP 2011)

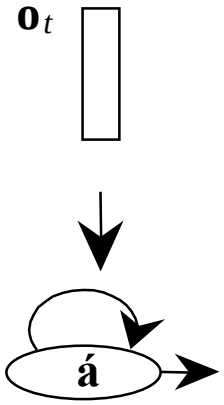
Seide et al, Interspeech, 2011.

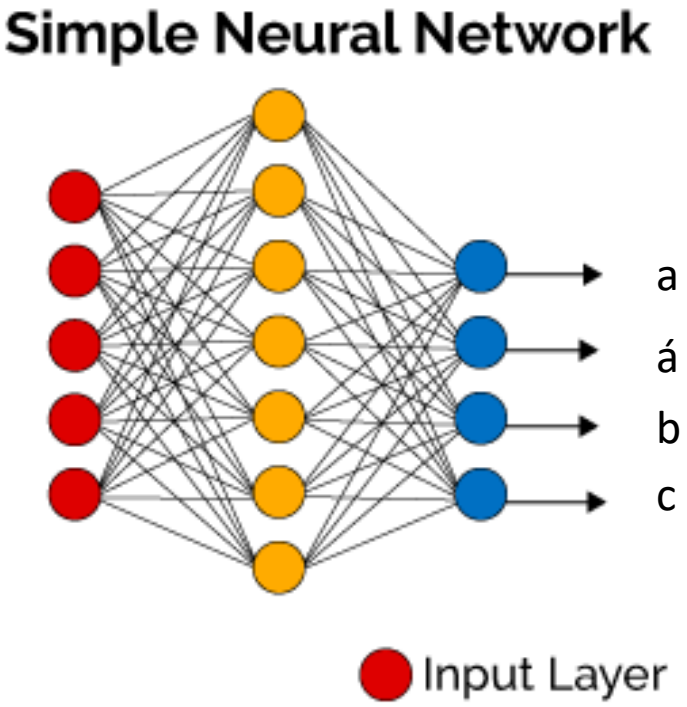


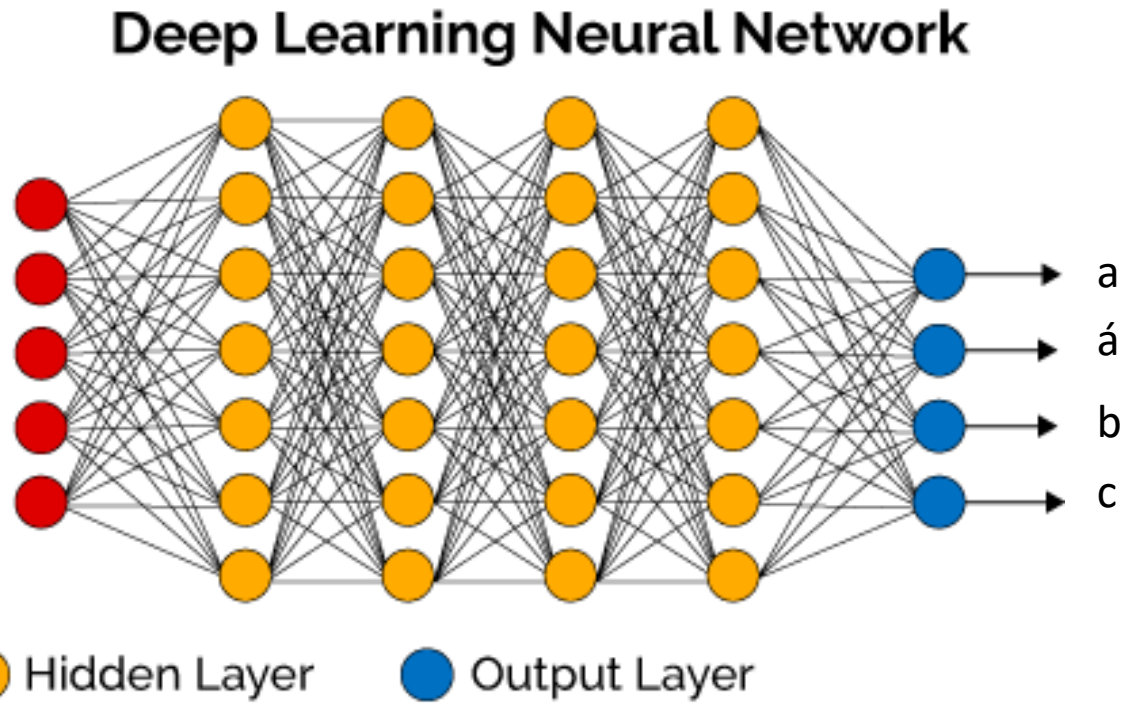
10

Deep learning acoustic models

- Deeper structures – higher abstraction

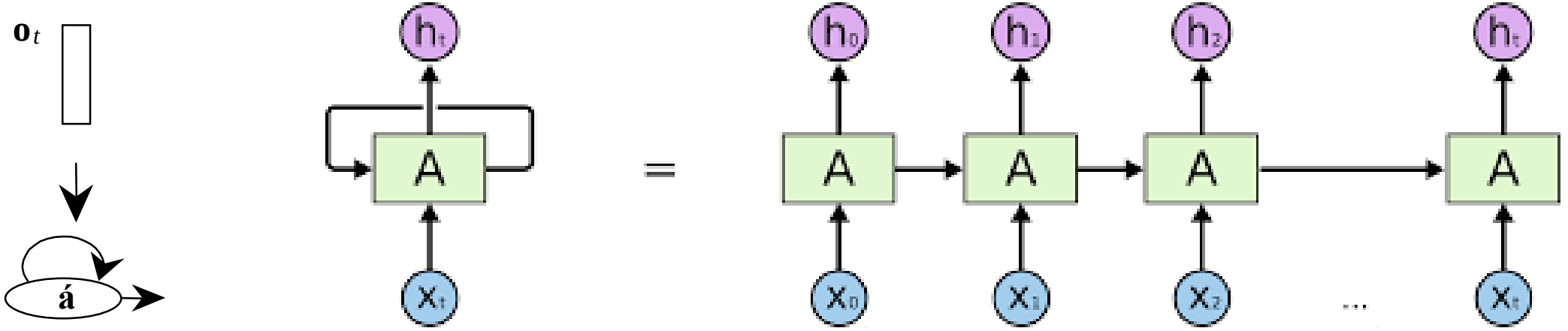




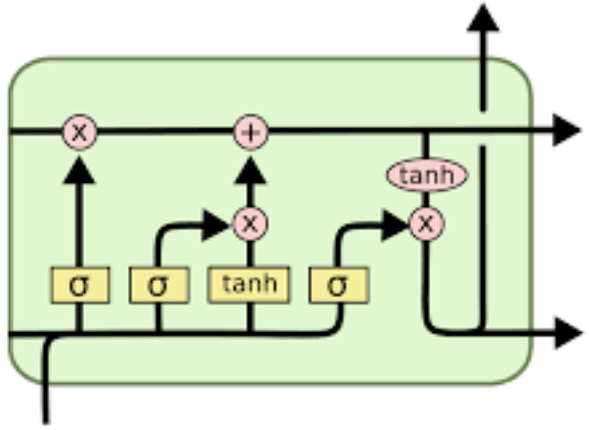


Deep learning acoustic models (2)

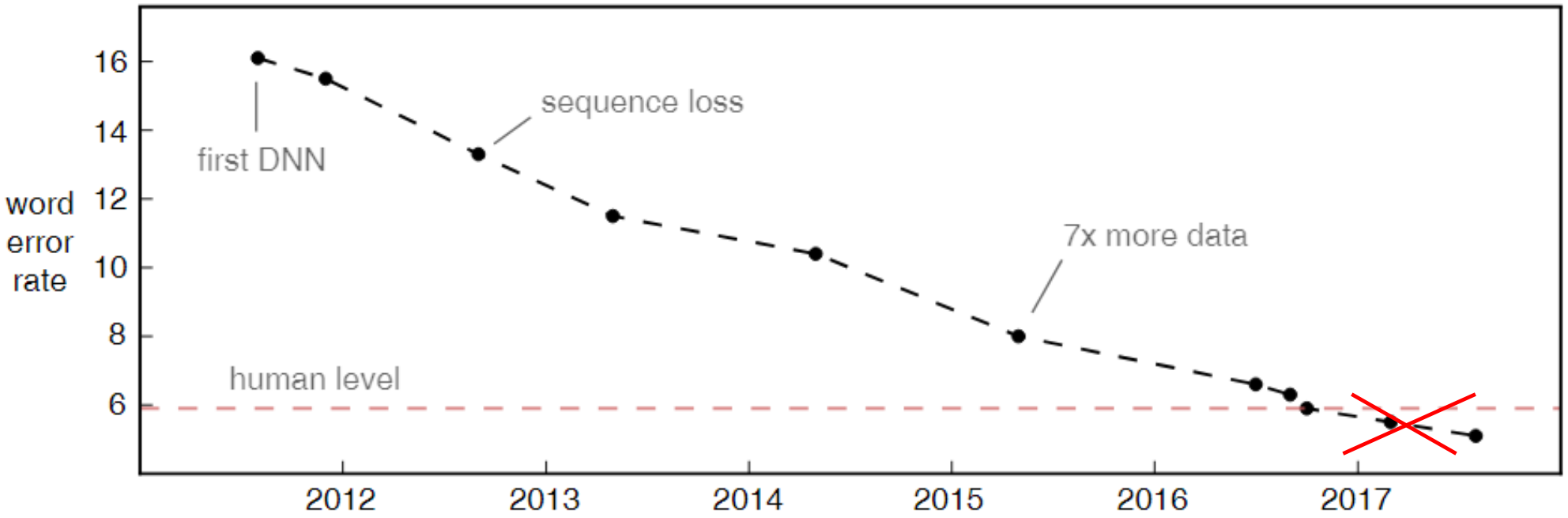
- Recurrent structure – „we don’t forget what has happened before”



LSTM (Long Short-Term Memory)



The effect of Deep Learning on WER



Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.

End-to-end deep neural net based ASR

End-to-end automatic speech recognition

mihajlik.peter@vik.bme.hu

Basic idea: sequence 2 sequence modeling using recurrent nets

- LSTM

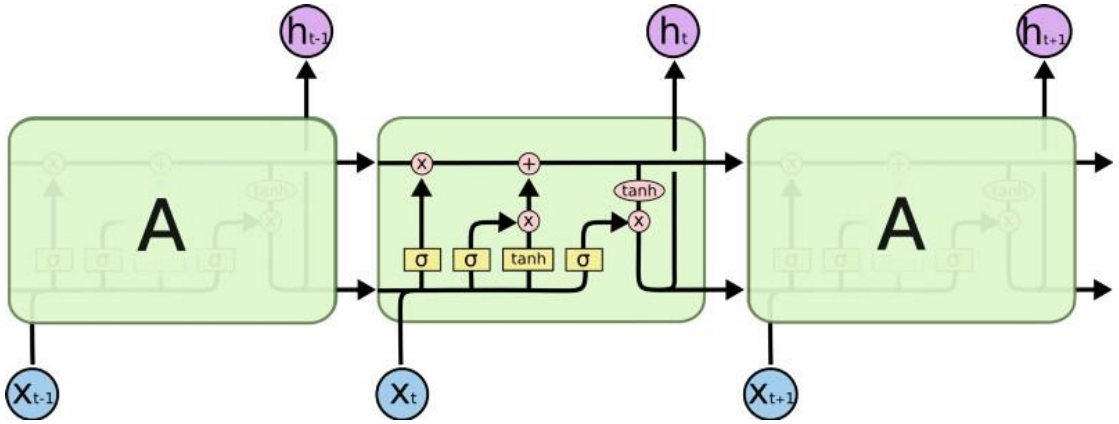
Highest probability?

A B C ··· Z

···

···

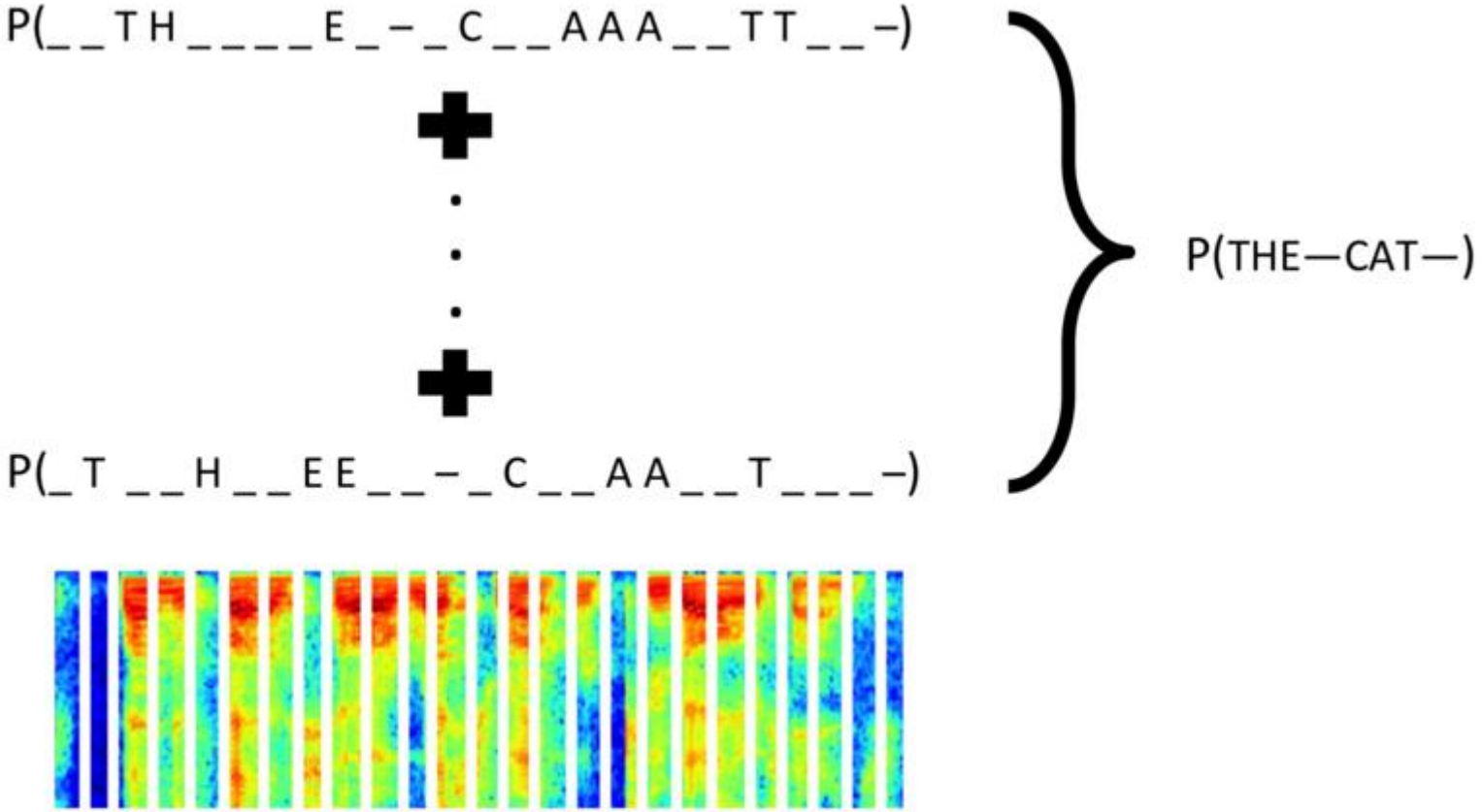
Text (chars, words, word fragments …)



Acoustic feature vectors

The challenge: time alignment

„Connectionist Temporal Classification”

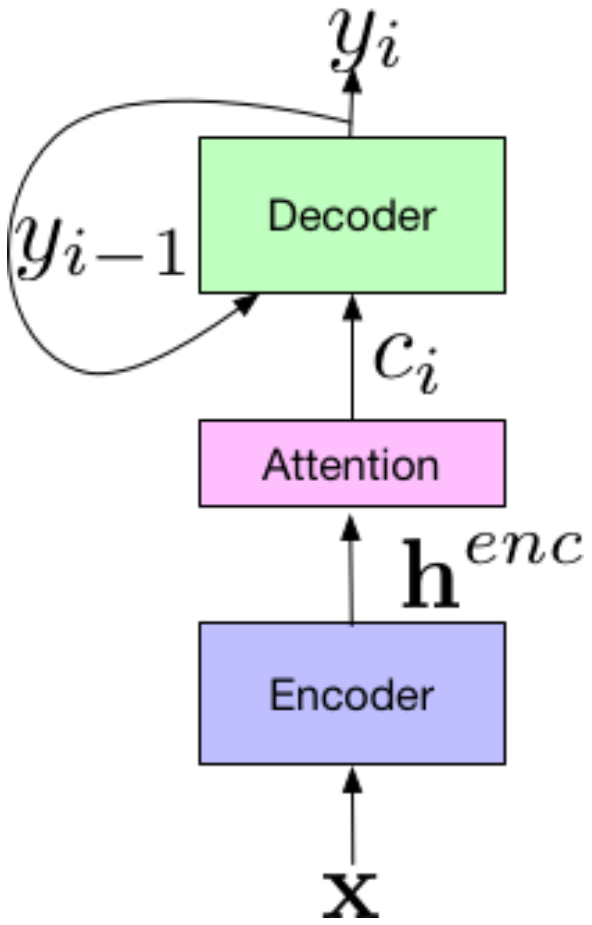


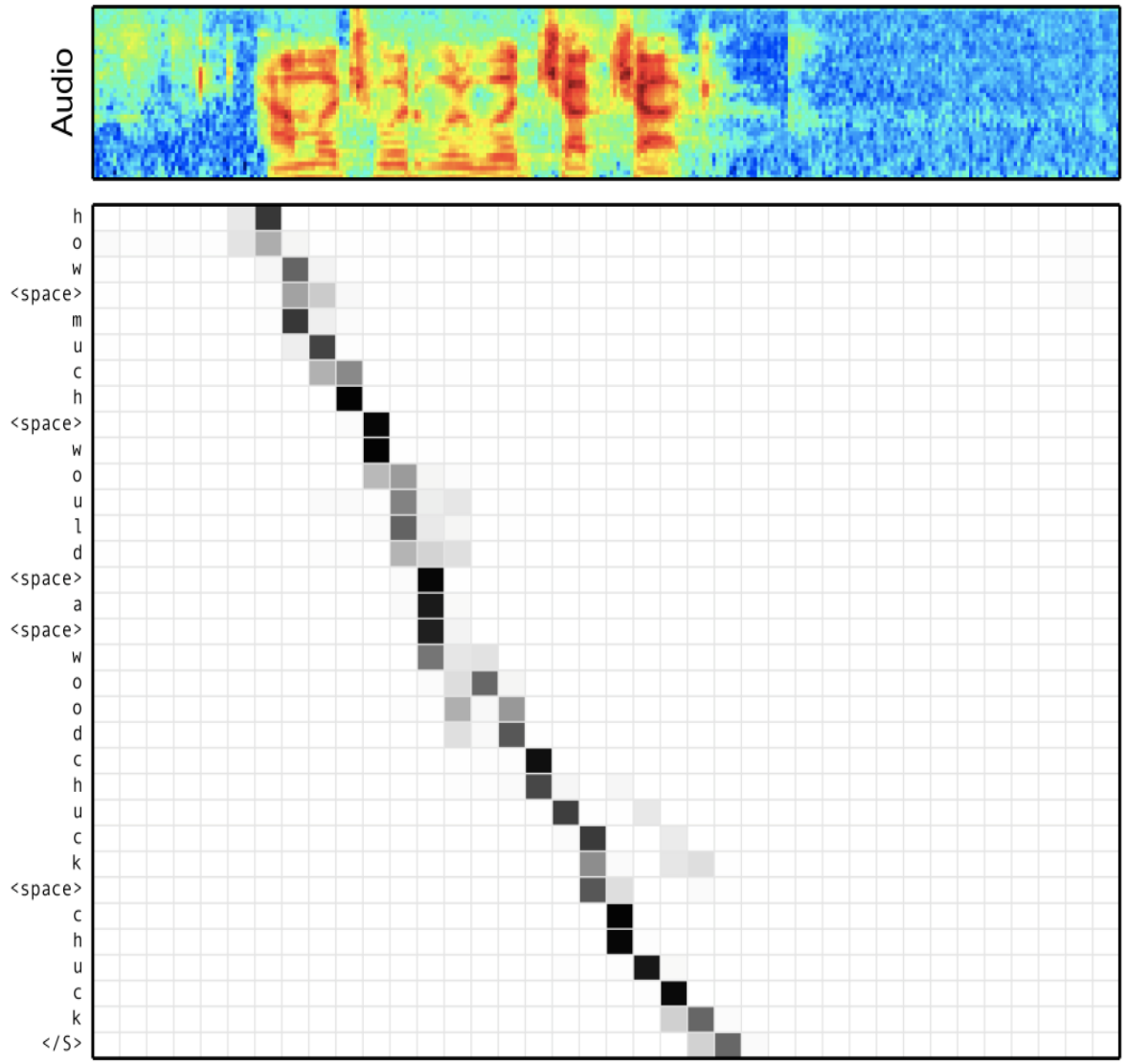
Listen – Attend – Spell (LAS) end-to-end (2016)

LM

Time alignment

Acoustic preprocessing





Transformer Acoustic Models

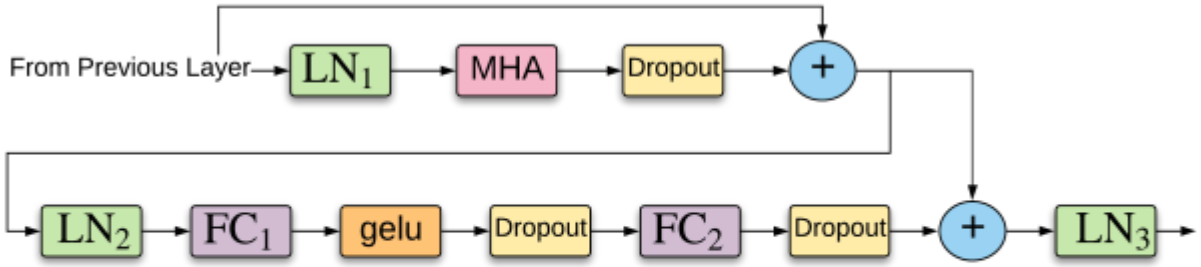
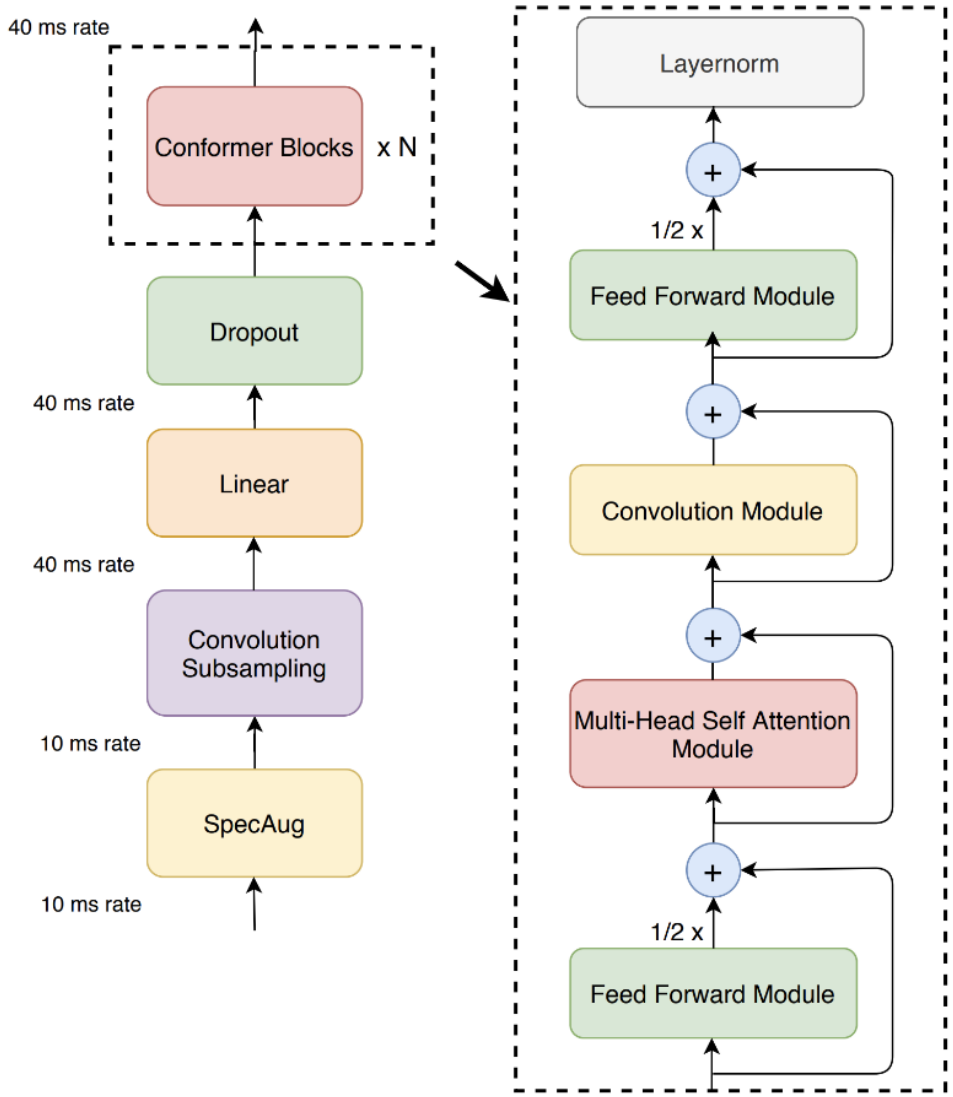


Fig. 1: Architecture of one transformer layer. "LN" means layer normalization [25]; "FC" means fully connected linear transformation; "gelu" means the gelu nonlinear activation [26].

State-of-the-art in ASR: Conformer end-to-end



Transformer with Self-attention + Convolution

Google

End-to-end Deep Learning approach

- No phonemes
- No dictionaries
- No language experts
- Still good to have LM

Fully data driven

2020: the beginning of a new era in ASR

Paradigm shift from fully supervised learning to **unsupervised pre-training** + supervised fine tuning

Any better idea than initializing NN weights with random numbers?

Transfer learning: use English model weights to initialize Hungarian (end-to-end) ASR training

- We still need a lot of manually transcribed data (in English)!

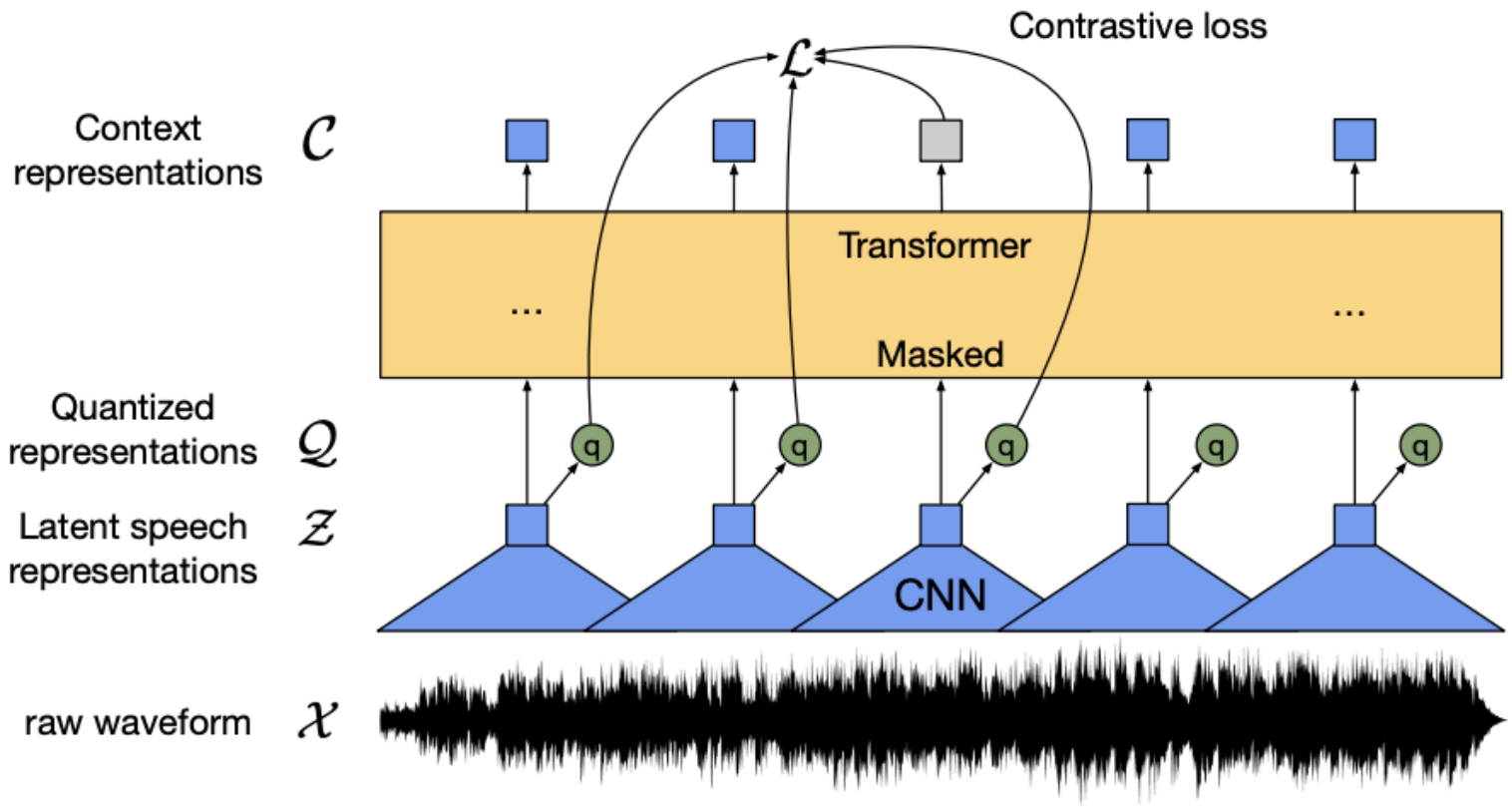
Unsupervised pre-training on pure acoustic data?

- Retsricted Boltzmann-machines (outdated)

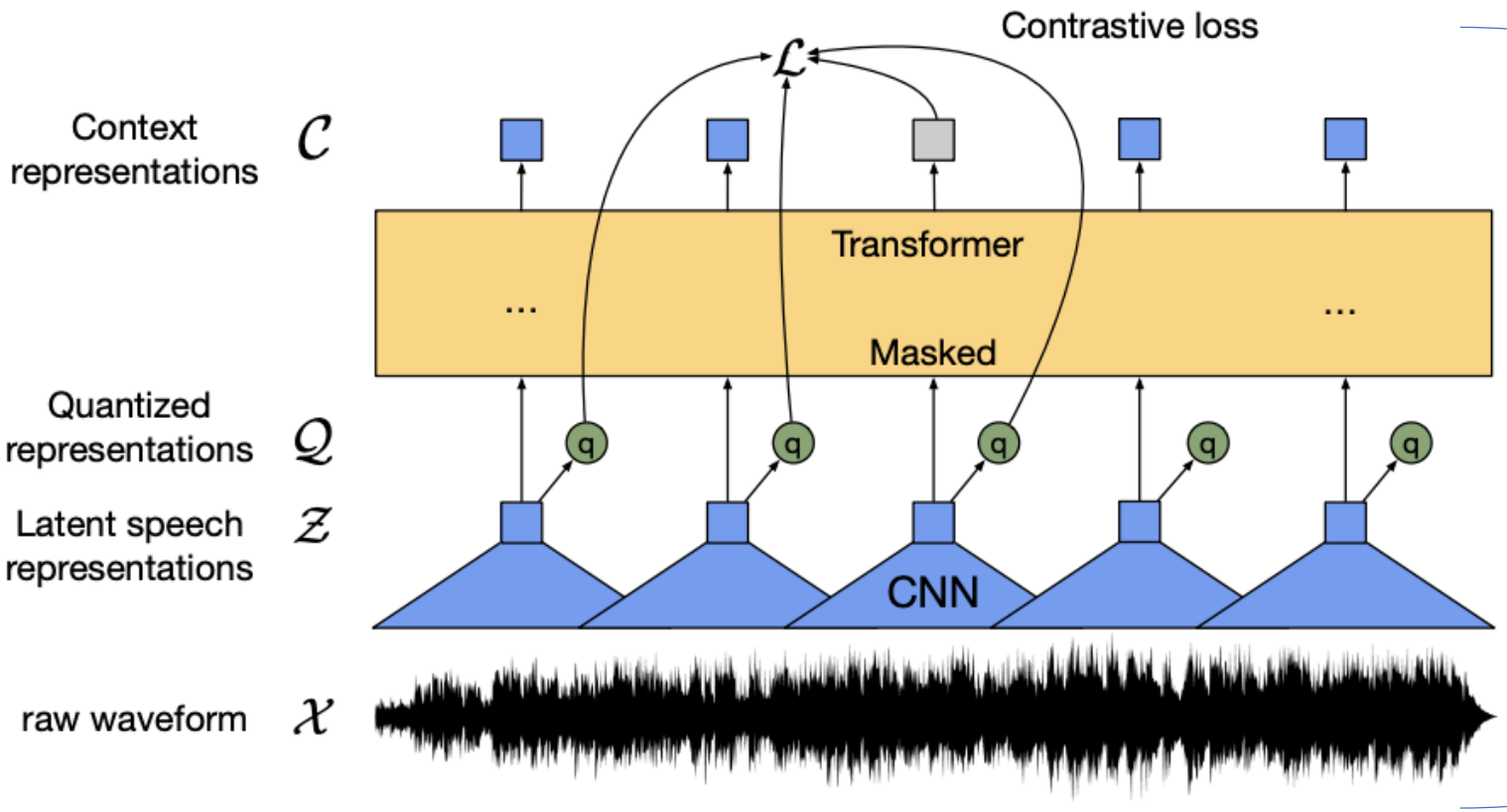
Self-supervised pre-training!

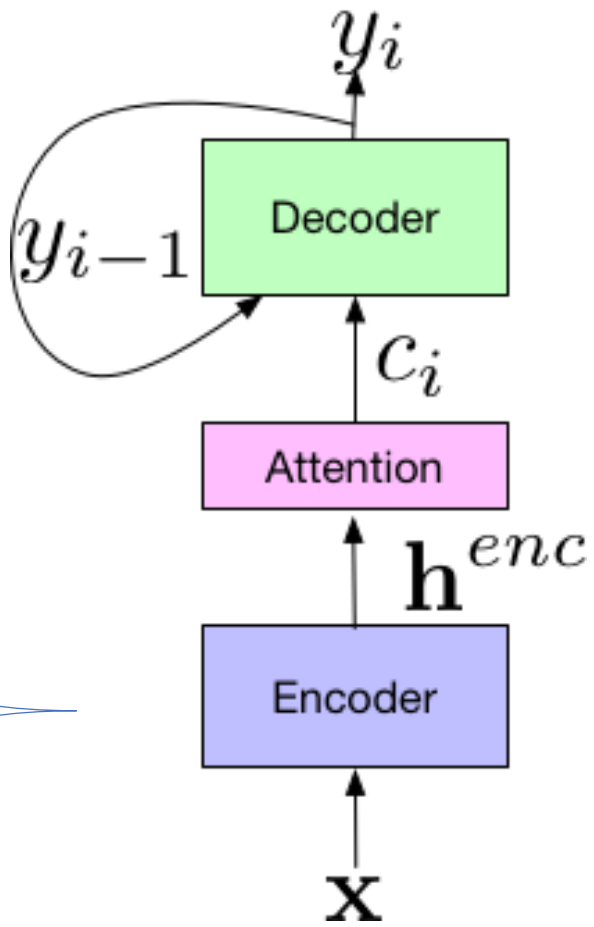
- Based on the very successful BERT training…

Wav2vec2.0


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
Masked
Quantized representations
$\mathcal{Q}$
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

Wav2vec2.0 + fine tuning





2022: Surprise from OpenAI: Whisper

Robust Speech Recognition via Large-Scale Weak Supervision

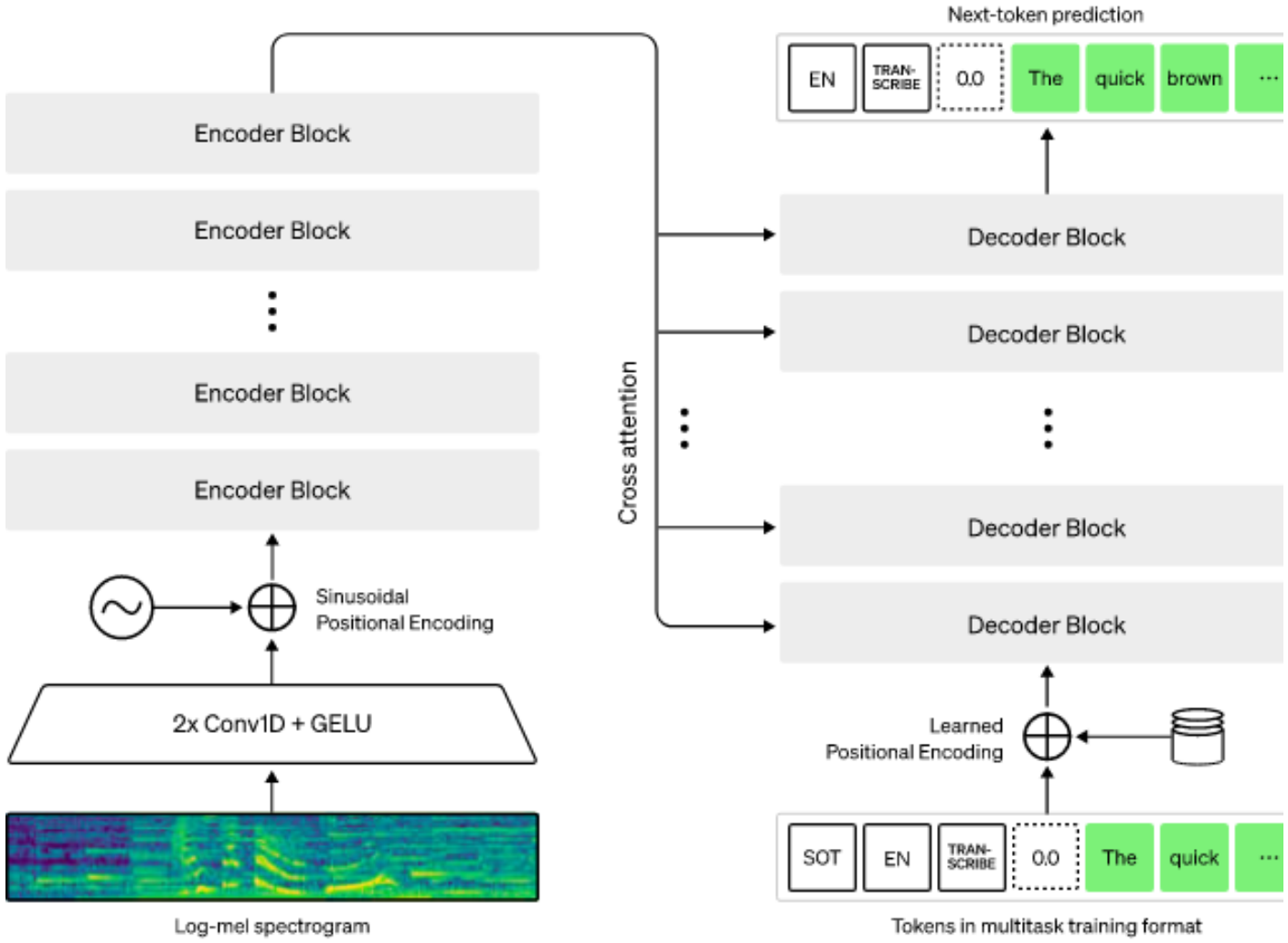
Whisper

Weak supervision = not exact transcriptions

- Encoder + decoder transformer
- Multilingual, multitask learning
- Language ID
- Punctualization

BUT

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



Case-study: Hungarian

Automatic transcription of **spontaneous** vs. **read** Hungarian speech

mihajlik.peter@vik.bme.hu

Composition of BEA-Base datasets

Modules

repetitional (same text)

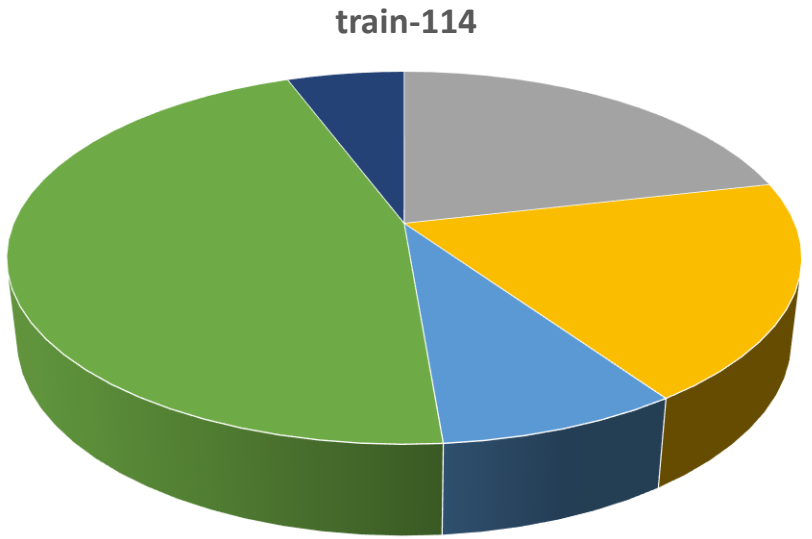
- Repeat
- Readsent

quasi-**spont**aneous

- Interview
- Opinion
- Summhist + Summplant
- Discourse

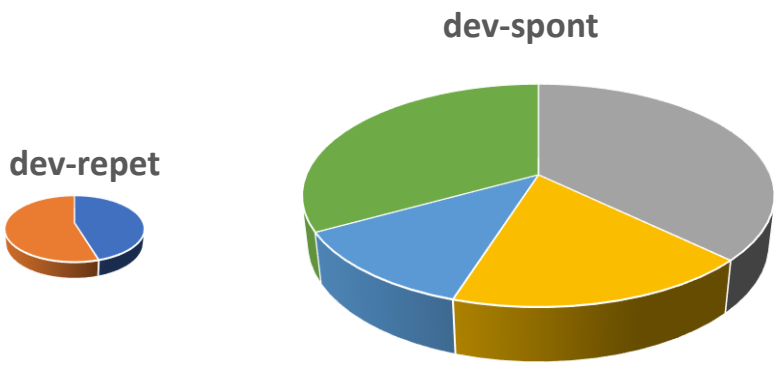
independent text

- Readtext

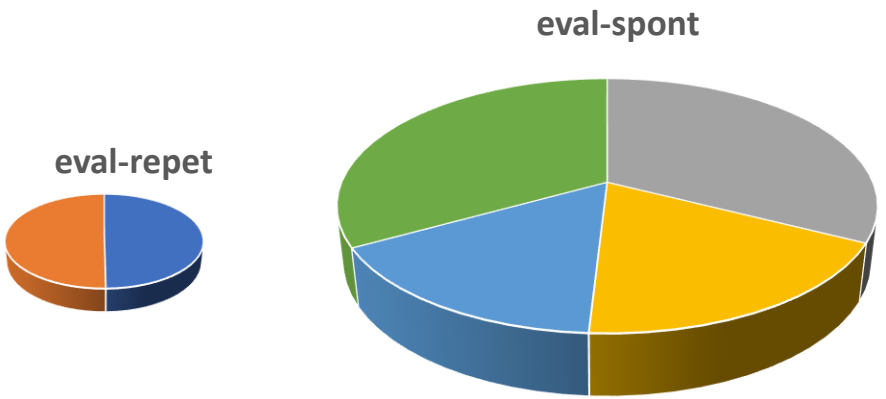


training

HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS



development



evaluation

mihajlik.peter@nytud.hu

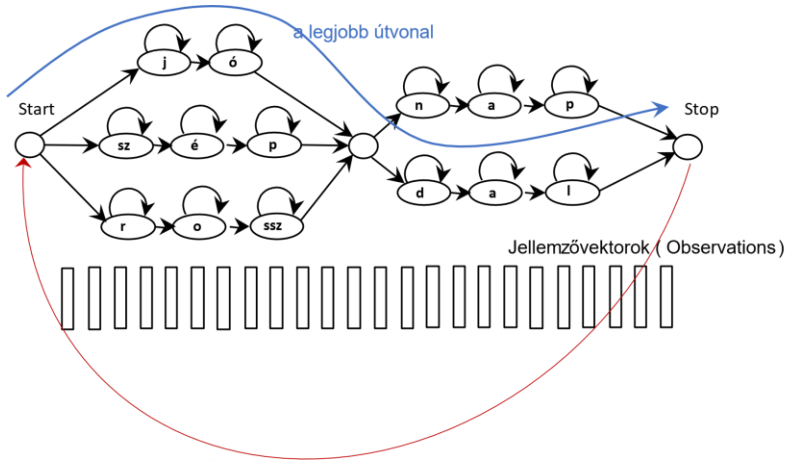
Classic ASR approach

HMM-DNN hybrid, KALDI

mihajlik.peter@nytud.hu

HMM-DNN hybrid setup

- "Kaldi chain" model
- LF-MMI training
- WSJ s5 recipe – without i-vectors
- ±10% speed perturbation
- Word 3-gram LM
- One-pass decoding



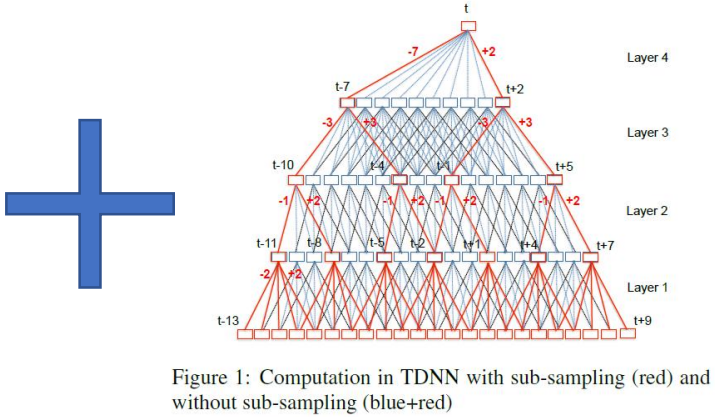


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

Kaldi based WER[%] results on BEA-Base

Structure / num of param.	Unit	eval-repet	eval-spont
TDNN-F / 18M	phoneme	6.26	28.41
	character	**6.08**	28.28
CNN-TDNN-F / 16M	phoneme	6.33	28.81
	character	6.28	**28.15**

End-to-end deep learning approach – fully supervised

Time-channel separable convolution / Conformer , starting from scratch vs. supervised pre-training

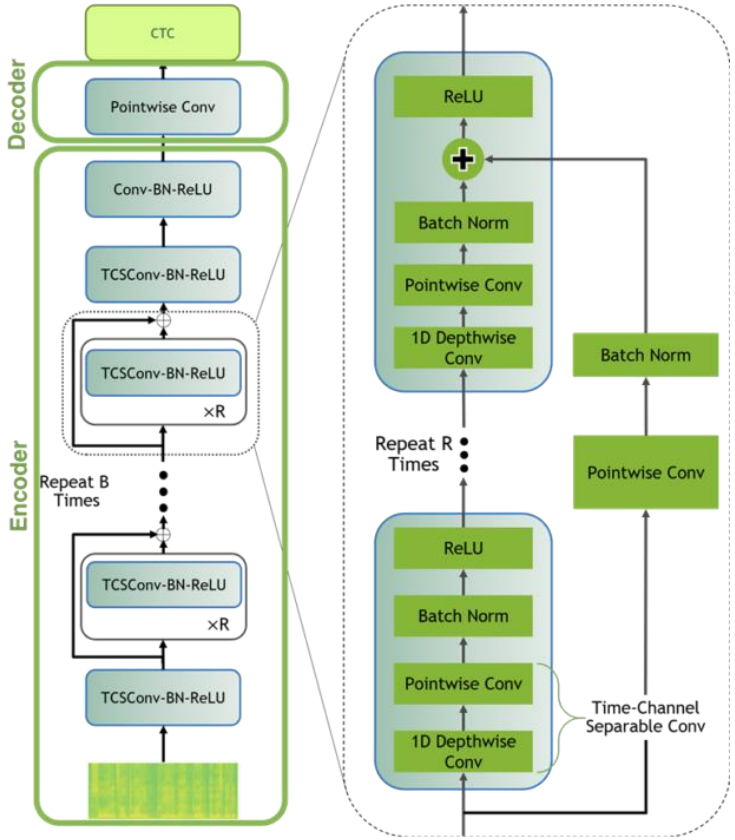
mihajlik.peter@nytud.hu

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
QuartzNet 15x3 / 12.7M	–	11.56	26.70
	3-gram	6.86	26.83
Conformer-Small / 13M	–	12.73	25.31
	6-gram	7.98	22.78
Conformer-Large / 121M	–	10.98	24.93
	6-gram	**5.65**	**21.01**

QuartzNet



Supervised pretraining* + fine-tuning

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

Pretraining language	Pretraining data size [hours]	LM	eval-repet	eval-spont
English	3k	–	10.63	**24.87**
		3-gram	**5.83**	25.23
English » Italian	3k + 160	–	11.91	25.24
		3-gram	6.39	25.84
English » German	3k + 700	–	13.12	25.67
		3-gram	6.20	26.09

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

English	~10k	–	11.22	21.39
		6-gram	4.96	17.77
		–	5.2	17.24
		6-gram	**3.66**	**16.25**

* Pretrained models were downloaded from https://catalog.ngc.nvidia.com/models

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

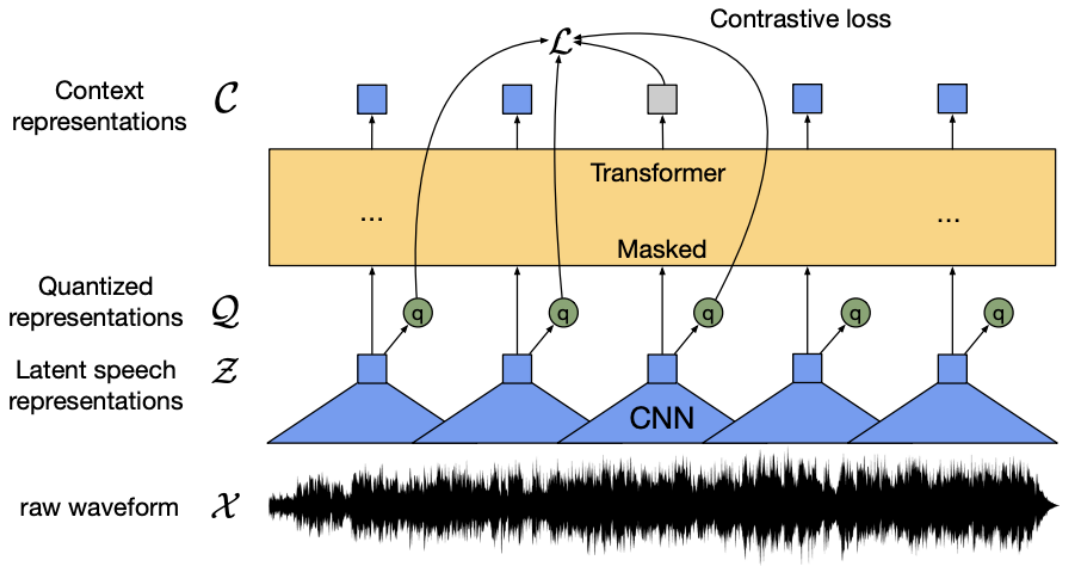
Transcription-free SSL pre-training + wav2vec2 encoder + attentional decoder

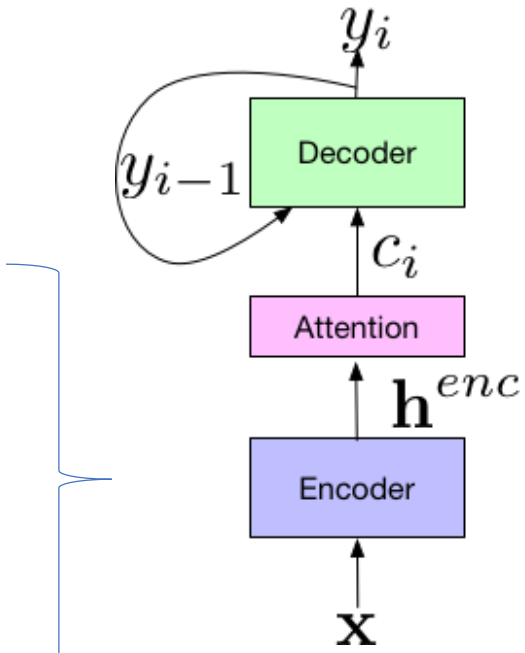
mihajlik.peter@nytud.hu

Self-Supervised Pretraining based Transfer Learning

- SSL pretraining
- SSL + supervised pretraining
- All pretrained models are dowloaded from HuggingFace

- Units: BPE (600)
- Loss: CTC + NLL
- Decoder: GRU
- Encoder: wav2vec2.0-large, 320M
- LM: –





Wav2vec2 based Transfer Learning Results

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

Pretraining language	Pretraining data size [hours]	eval-repet	eval-spont
English (SSL)	60k	8.46	19.17
Italian (SSL)	4.5k	7.45	19.07
Multi » German	53k + 700	6.66	17.99
Multi » Turkish	53k + 20	7.50	18.06
Multi » Hungarian	53k + 33	5.60	17.00
Multi (SSL)	53k	5.81	16.62
Mega (SSL)	440k	6.16	15.61
Uralic (SSL)	40k	**4.24**	**11.55**

Multi = wav2vec2-large-xlsr-53
Mega = wav2vec2-xls-r-300m

Wav2vec2 based Transfer Learning Results

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

Pretraining language	Pretraining data size [hours]	eval-repet	eval-spont
English (SSL)	60k	2.59	5.94
Italian (SSL)	4.5k	2.50	6.44
Multi » German	53k + 700	2.25	5.55
Multi » Turkish	53k + 20	2.46	5.53
Multi » Hungarian	53k + 33	2.63	5.45
Multi (SSL)	53k	2.09	5.53
Mega (SSL)	440k	2.39	5.11
Uralic (SSL)	40k	**1.67**	**3.68**

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

Whisper in zero-shot and fine-tuning setups

mihajlik.peter@nytud.hu

Whisper results: zero-shot vs. fine-tuning

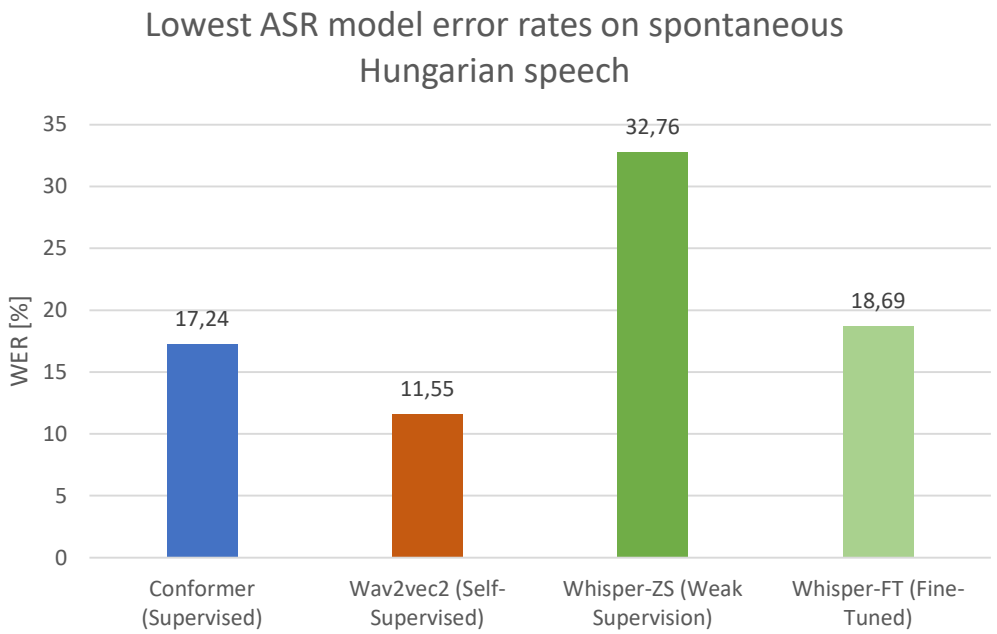
Whisper Medium/Large_v2 WER results of BEA-Base

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

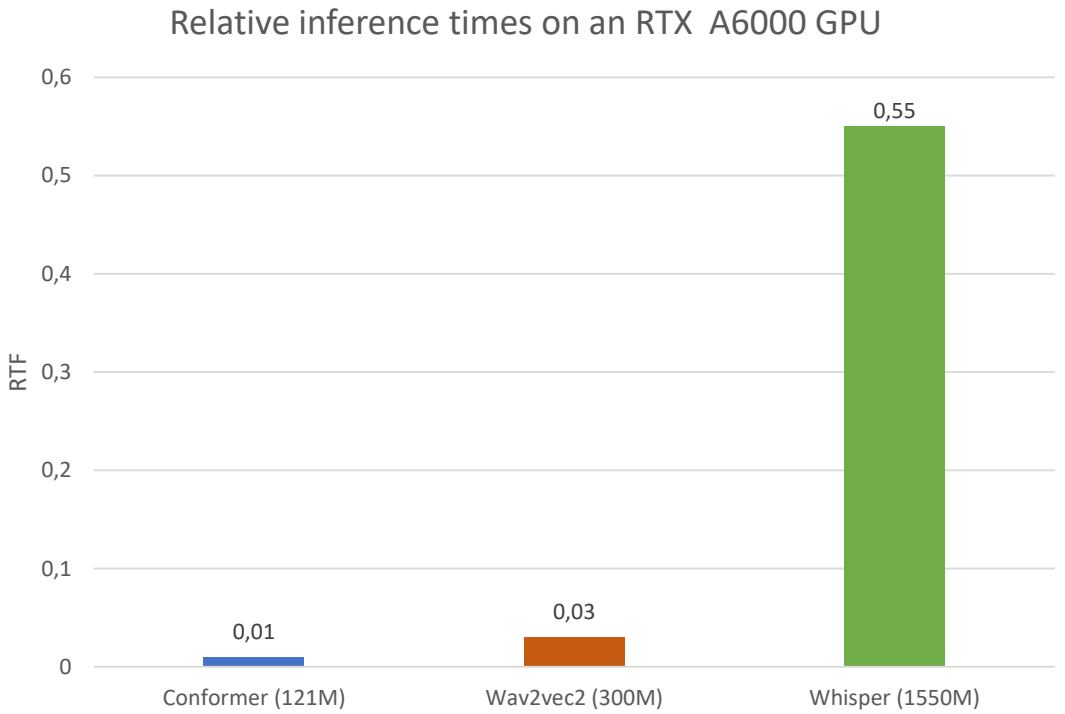
Final comparison – key takeaway

Best of Conformer vs. Wav2vec2 vs. Whisper

Best ASR results on spontaneous Hungarian vs. inference times



/Word Error Rate: the lower the better/



/Real-Time Factor: the lower the better/

Thank you for your attention.

Comment, questions…

WER(Word Error Rate) for Hungarian?

```
Format:
========================== WER in[%]
<eps> ; reference  ; on ; the ; first ;  line
  I   ;     S      ; =  ;  =  ;   S   ;   D
 and  ; hypothesis ; on ; the ; third ; <eps>


================================================================================ WER = 33.33%
a ;   fehér    ; ruhás ; is ; meglepődött ; kissé
= ;     S      ;   D   ; =  ;      =      ;   =
a ; fehérruhás ; <eps> ; is ; meglepődött ; kissé
================================================================================ WER = 57.14%
és ; felmutatott  ; egy ; szürke ; tollú ;    kis     ; madárra
=  ;      S       ;  =  ;   =    ;   S   ;     S      ;    D
és ; felmuthatott ; egy ; szürke ; tolló ; kismadárra ;  <eps>
================================================================================ WER = 71.43%
és ; nézzétek ; hogy ; lüktet ; <eps> ; a ; kicsi  ; torka
S  ;    =     ;  =   ;   S    ;   I   ; = ;   S    ;   S
s  ; nézzétek ; hogy ;  lük   ; tehát ; a ; kicsit ;  orka
================================================================================ WER = 100.00%
fel ;       is      ;  kelt ; nyomban ; és ; továbbindult ; <eps>
 S  ;       S       ;   D   ;    D    ; =  ;      S       ;   I
 a  ; liskátnyomban ; <eps> ;  <eps>  ; és ;    tovább    ; indult
================================================================
```