BME TMIT

MŰEGYETEM 1782

BME-VIK

Introduction to Automatic Speech Recognition (ASR)

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

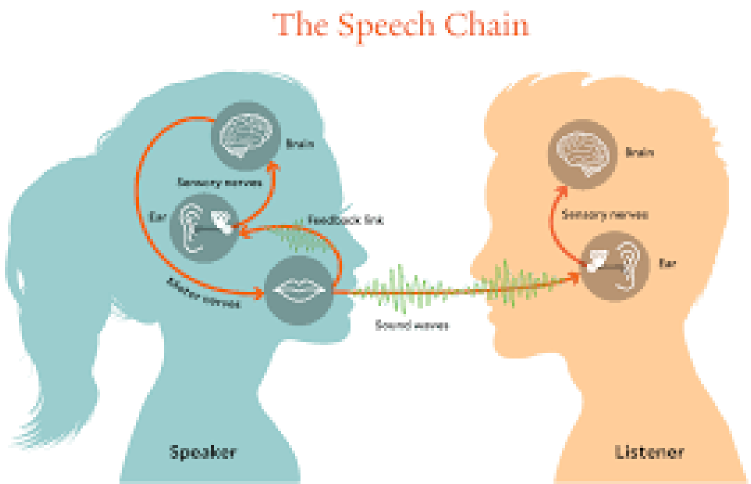
BME-TMIT
Speech Recognition Group
SmartLabs

BME VILLAMOSMÉRNÖKI ÉS INFORMATIKAI KAR

Speech technology?

Speech is complex…

- **<u>Speech-to-text (ASR)</u>**
- Text-to-speech (TTS)
- Who speaks when (diarization)
- Emotional state from speech
- Speaker recognition
- Speaker verification
- Dialogue managers…



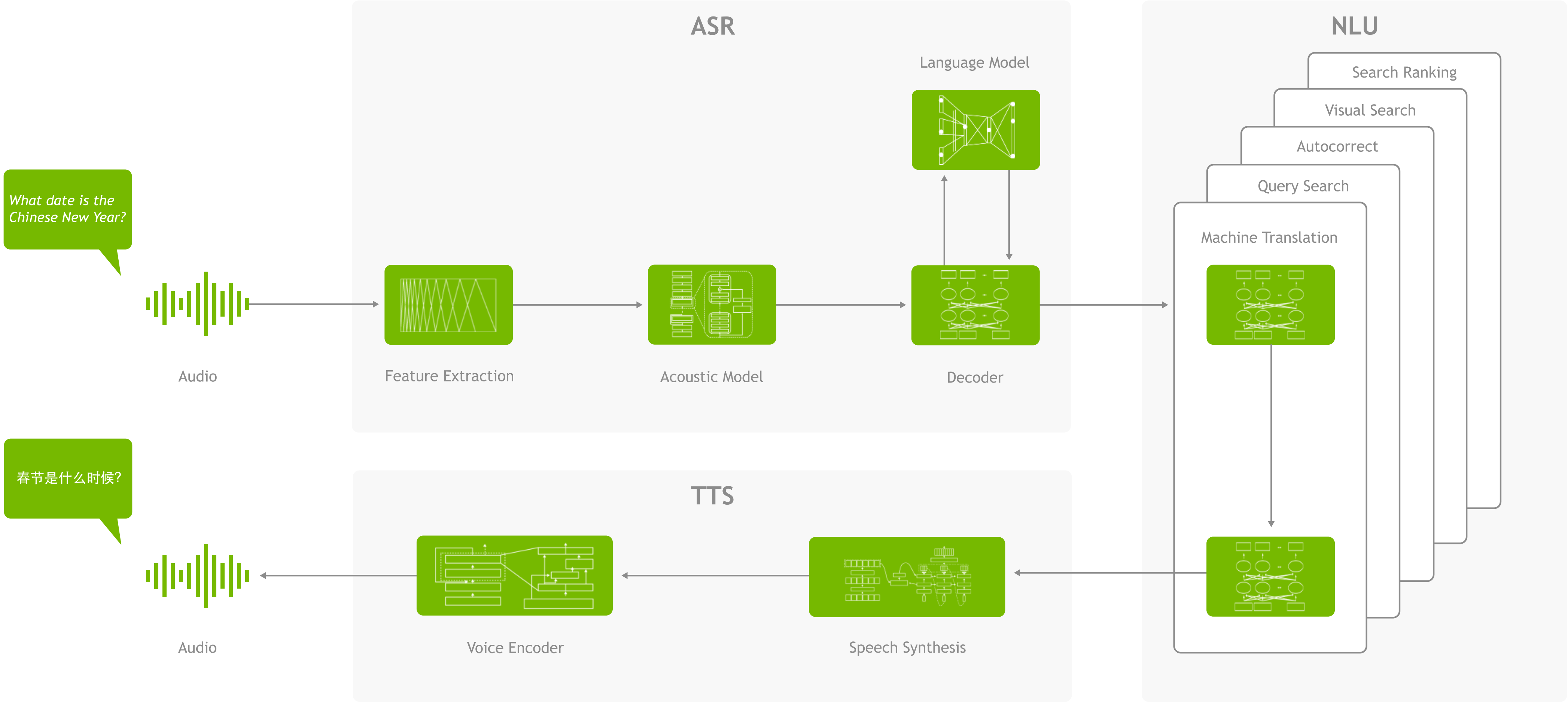
mihajlik.peter@vik.bme.hu

Speech technology and Artificial Intelligence

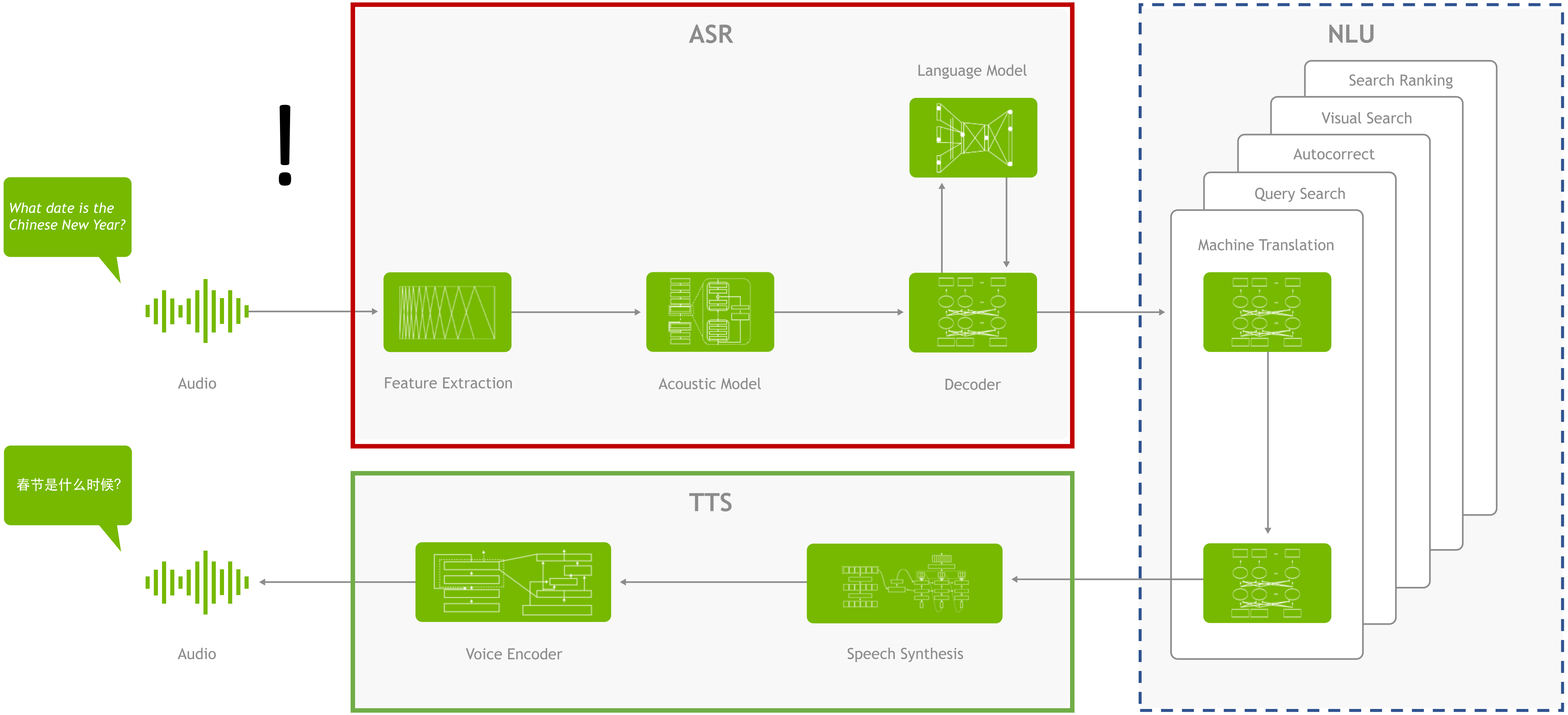
„Conversational AI”

mihajlik@tmit.bme.hu

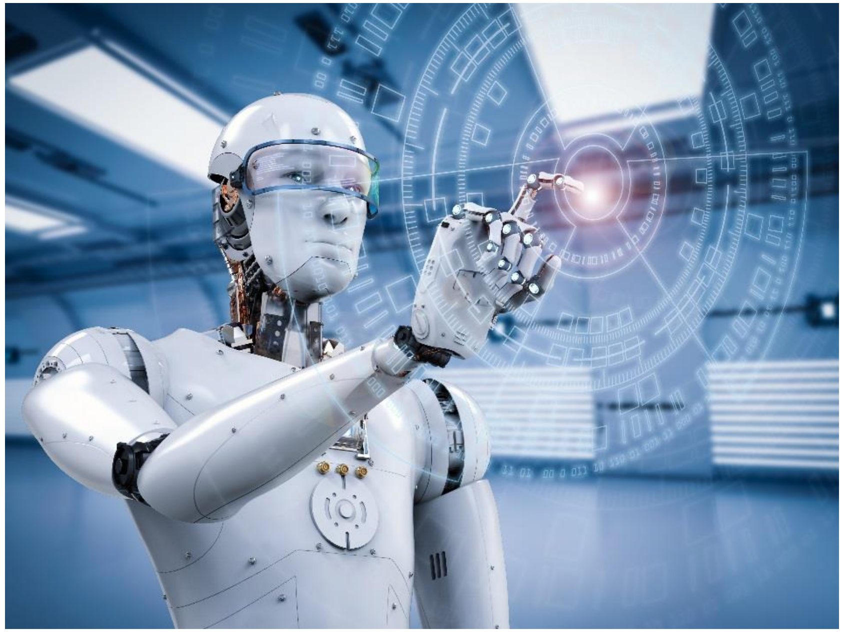
„Conversational AI”

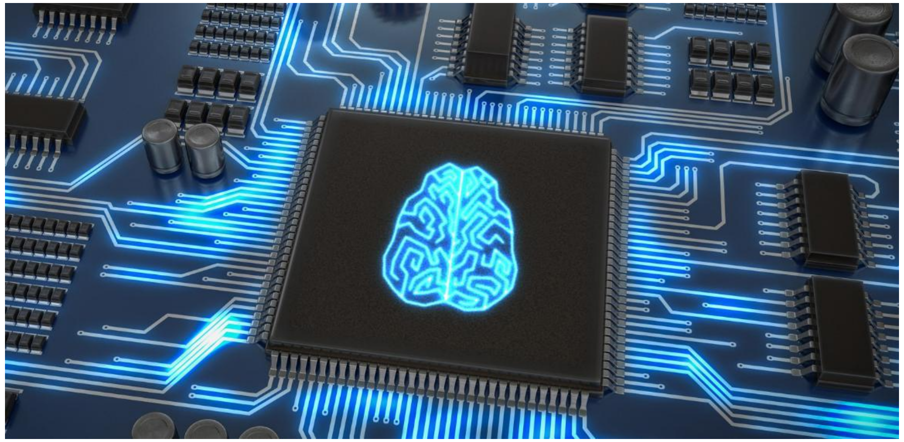


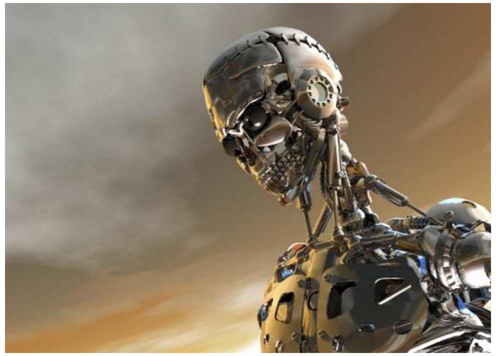
„Conversational AI" – key components



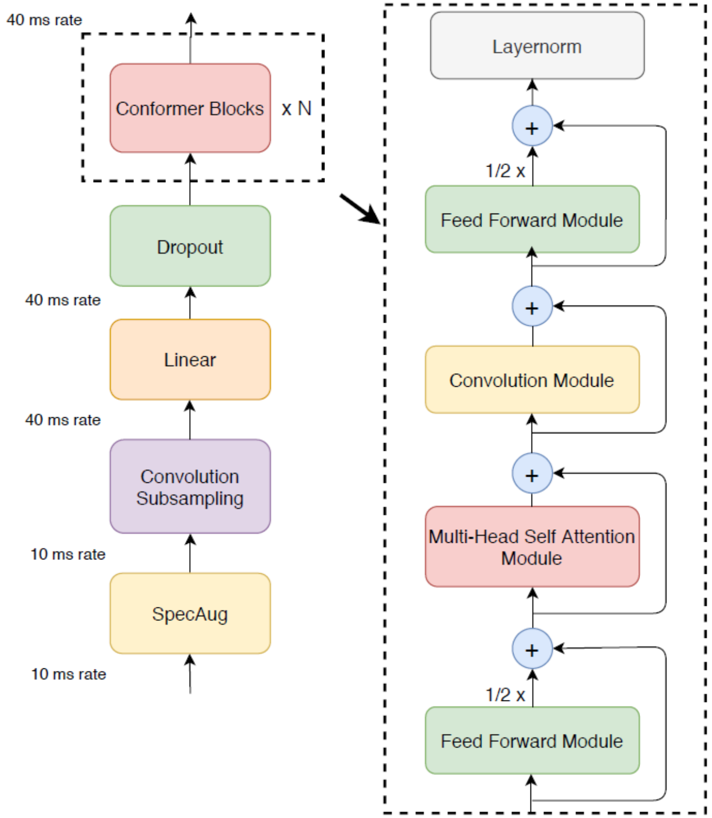
AI? – the „myth”:







AI – in reality:



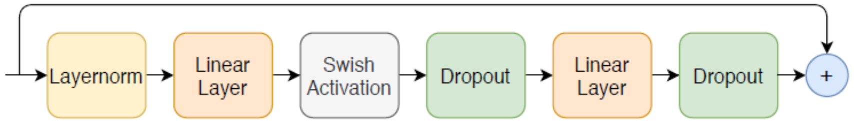


Figure 4: ***Feed forward module.*** *The first linear layer uses an expansion factor of 4 and the second linear layer projects it back to the model dimension. We use swish activation and a pre-norm residual units in feed forward module.*

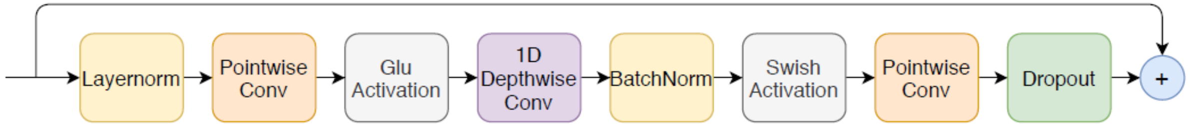


Figure 2: ***Convolution module.*** *The convolution module contains a pointwise convolution with an expansion factor of 2 projecting the number of channels with a GLU activation layer, followed by a 1-D Depthwise convolution. The 1-D depthwise conv is followed by a Batchnorm and then a swish activation layer.*

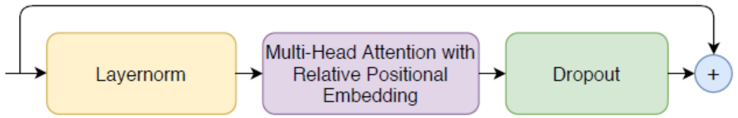


Figure 3: ***Multi-Headed self-attention module.*** *We use multi-headed self-attention with relative positional embedding in a pre-norm residual unit.*

Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

„**Conformer**: Convolution-augmented Transformer for Speech Recognition", by Anmol Gulati et al, in Proc. Interspeech-2020

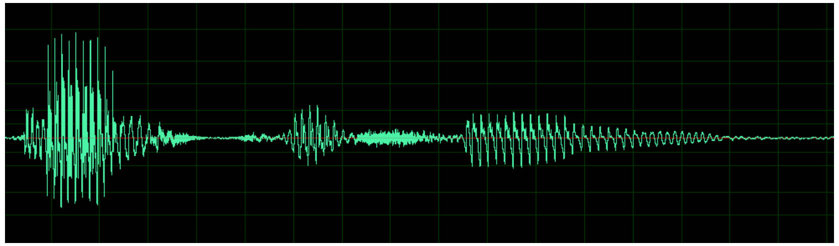
The evolution of ASR technology

A co-evolution with AI

mihajlik@tmit.bme.hu

Automatic Speech Recognition

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



„I think …”

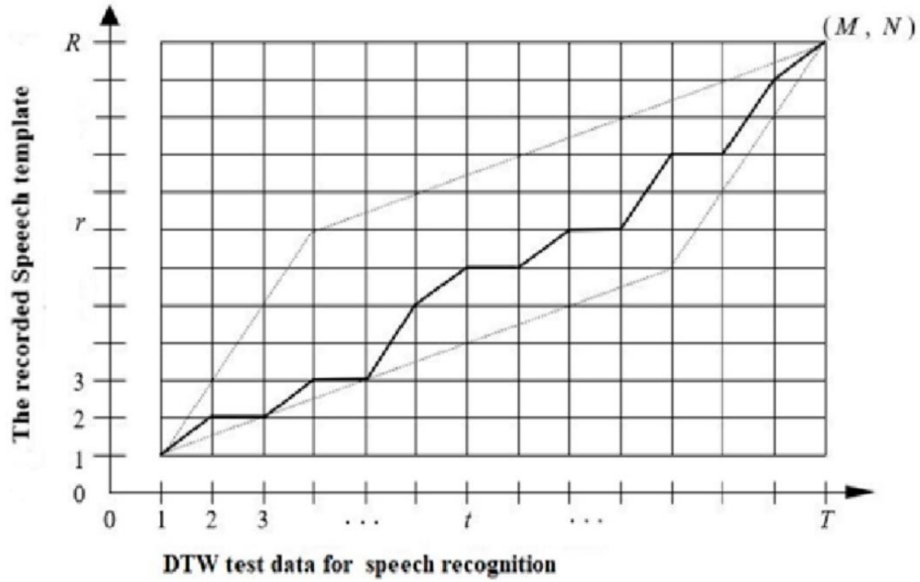
The beginning: electronic filters, rules-based algorithms

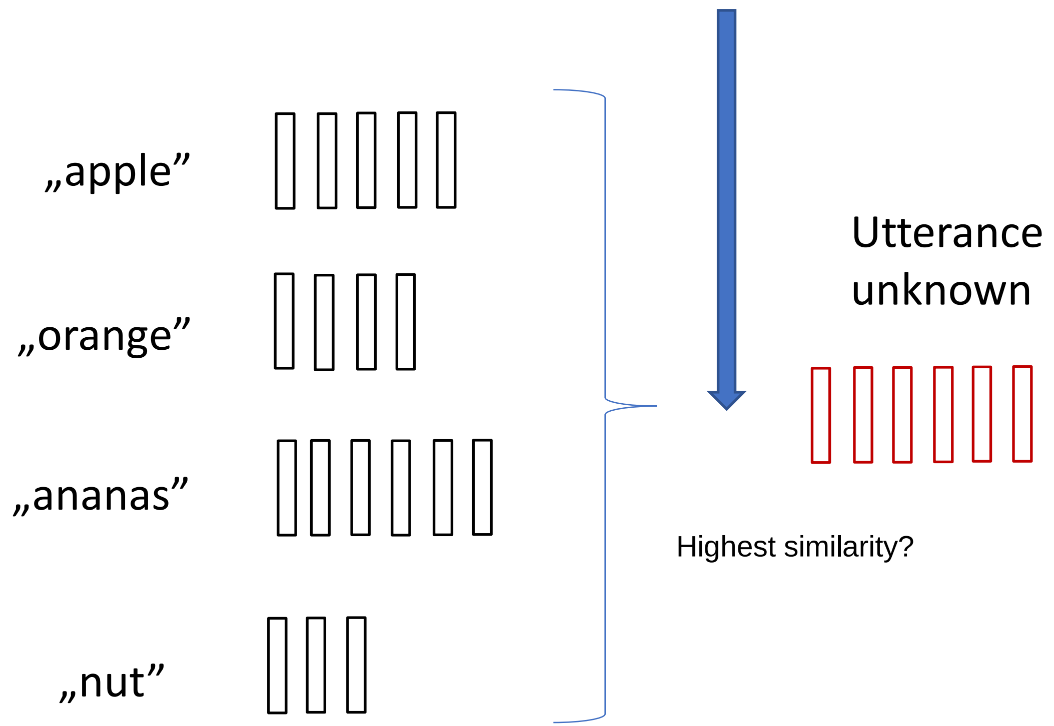
- 1950-52 Bell Laboratories:
  - Audrey (Automatic Digit Recognizer)
  - Numbers 1-9

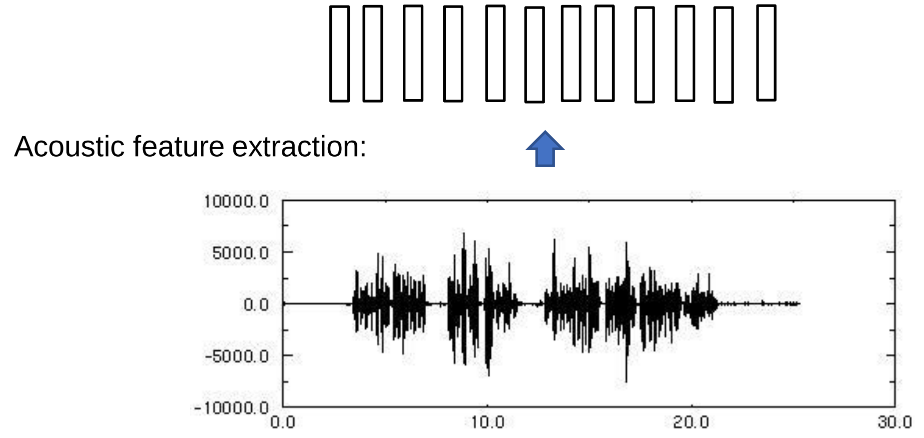
- 1961 IBM
  - Shoebox
  - Numbers 0-9,
  - 6 basic arithmetic operations

Template based, isolated-word recognition

- From 1970
- Dynamic Time Warping

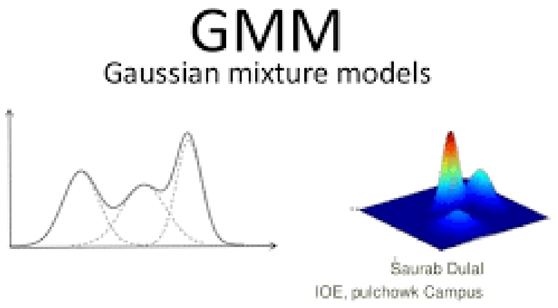


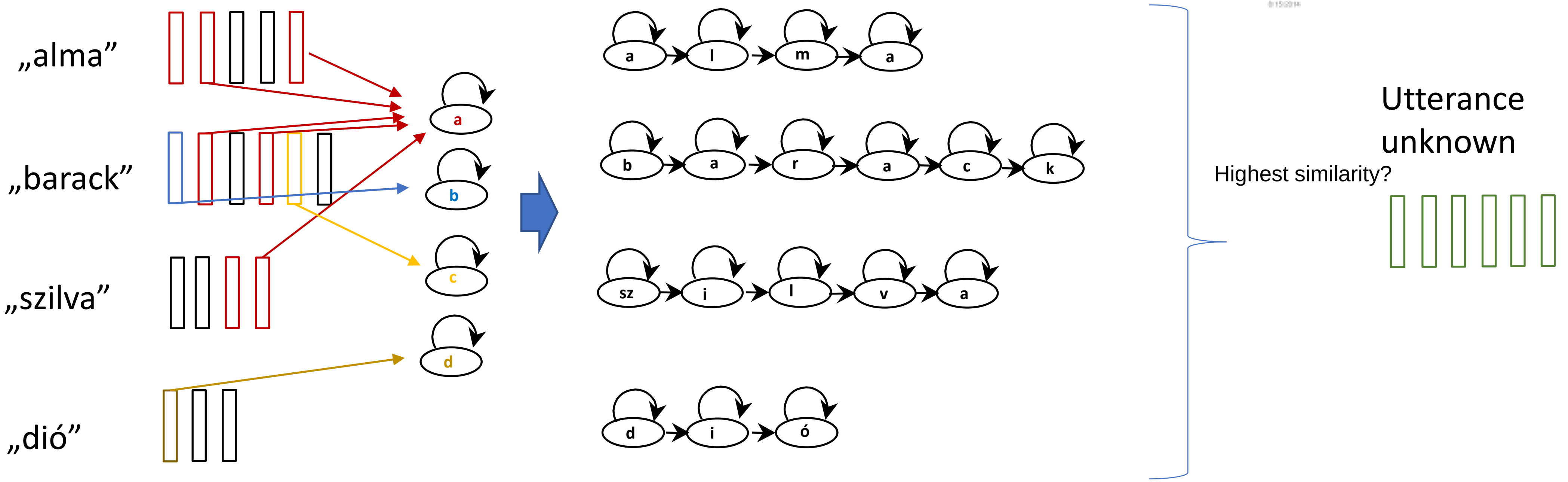




More data, phoneme-based ASR

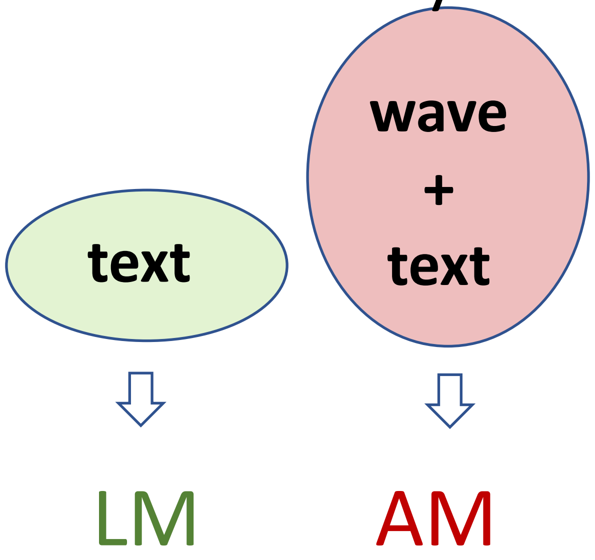
- Hidden Markov-modell (HMM), from 1975…
- Similarity measure: by GMM



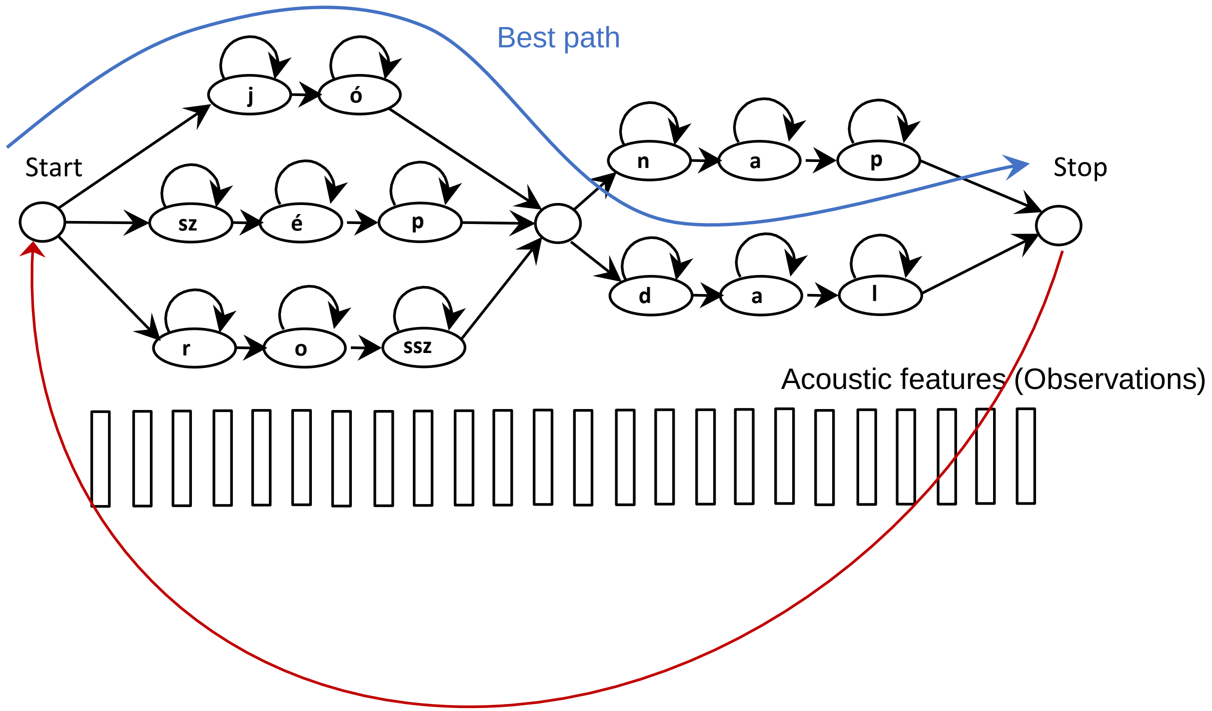


Adding text data and Language Model (LM)

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

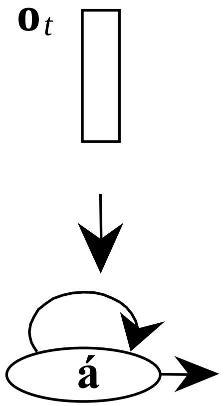


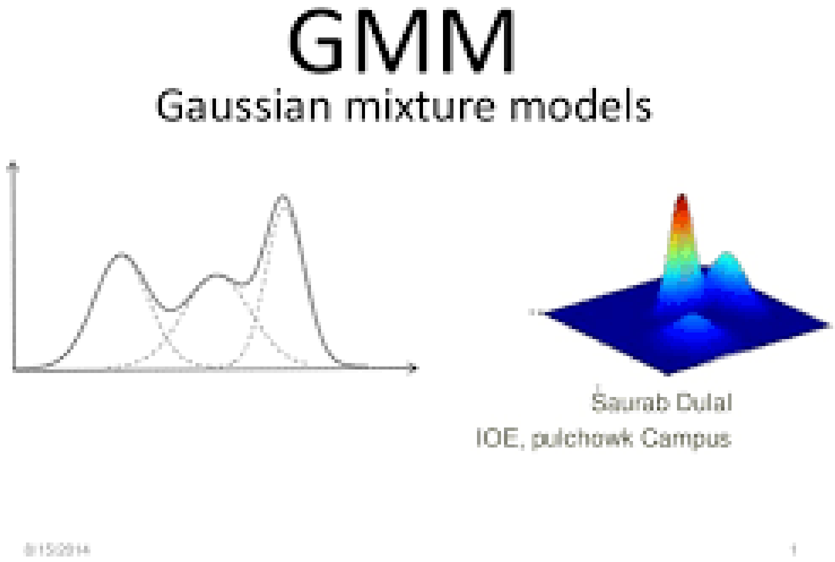
$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W)$$



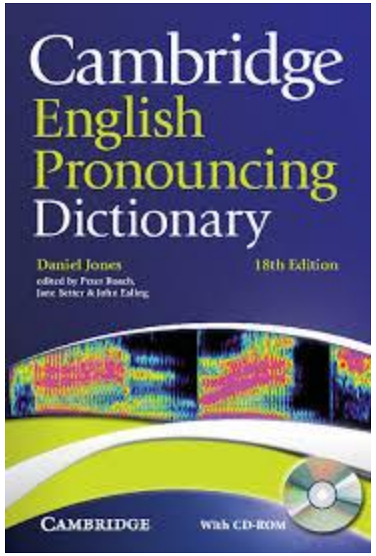
Acoustic modeling

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





Phonetic pronunciation dictionary

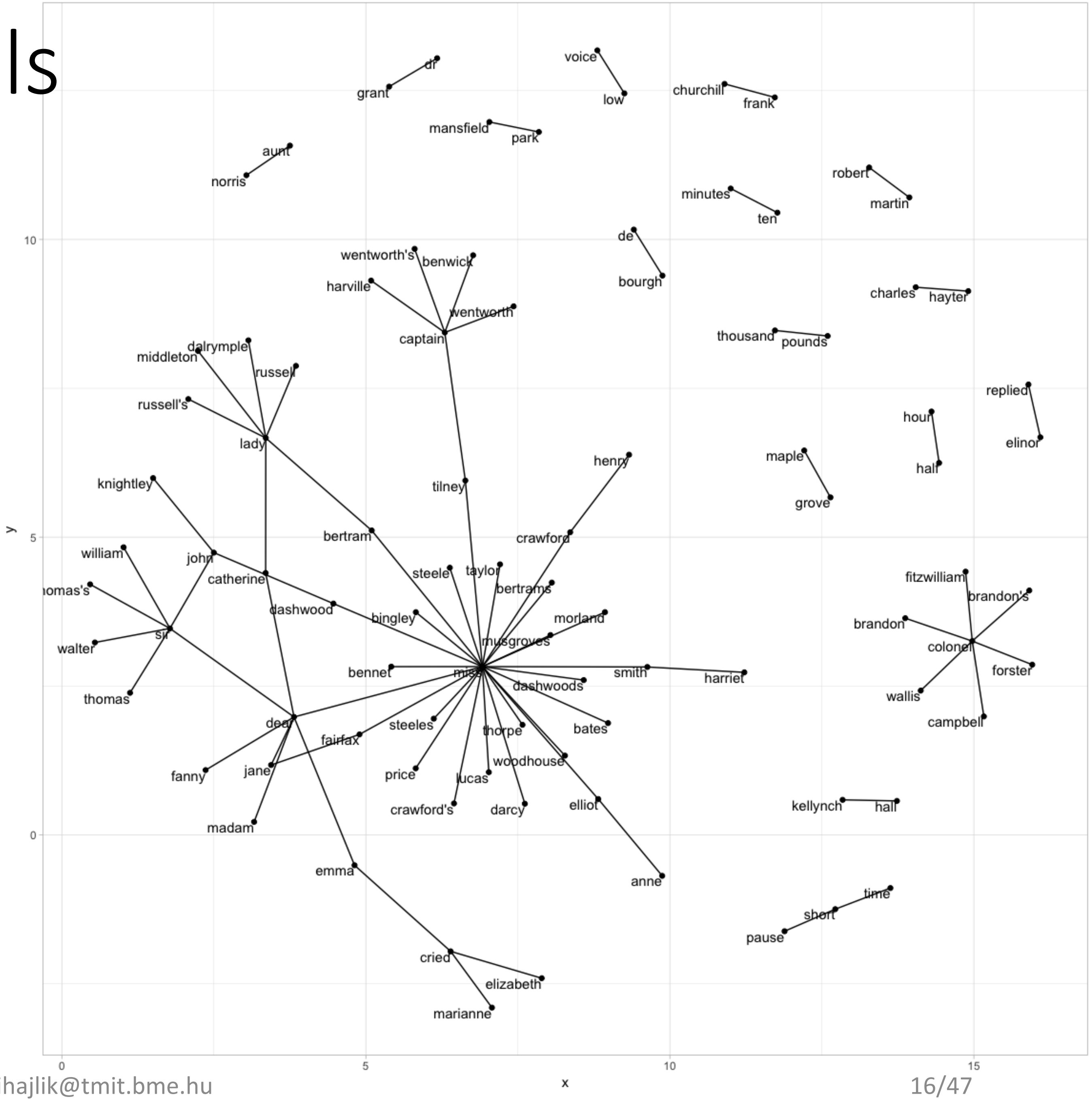



```
ABOARD   AH0 B AO1 R D
ABODE   AH0 B OW1 D
ABOHALIMA   AE0 B AH0 HH AH0 L IY1 M AH0
ABOLISH   AH0 B AA1 L IH0 SH
ABOLISHED   AH0 B AA1 L IH0 SH T
ABOLISHES   AH0 B AA1 L IH0 SH IH0 Z
ABOLISHING   AH0 B AA1 L IH0 SH IH0 NG
ABOLITION   AE2 B AH0 L IH1 SH AH0 N
ABOLITIONISM   AE2 B AH0 L IH1 SH AH0 N IH2 Z AH0 M
ABOLITIONIST   AE2 B AH0 L IH1 SH AH0 N AH0 S T
ABOLITIONISTS   AE2 B AH0 L IH1 SH AH0 N AH0 S T S
```

Word language models

- N-gram models
- Based on text stat
- Assigns probability to word sequences

- Language Models



Classic ASR

- Phoneme based
- Linguistic knowledge extensively used
- Expert linguists needed
- Separate levels of language modelled explicitely
- On-line, fast
- Flexible

- ASR Accuracy << Human accuracy

„The Deep Learning revolution”

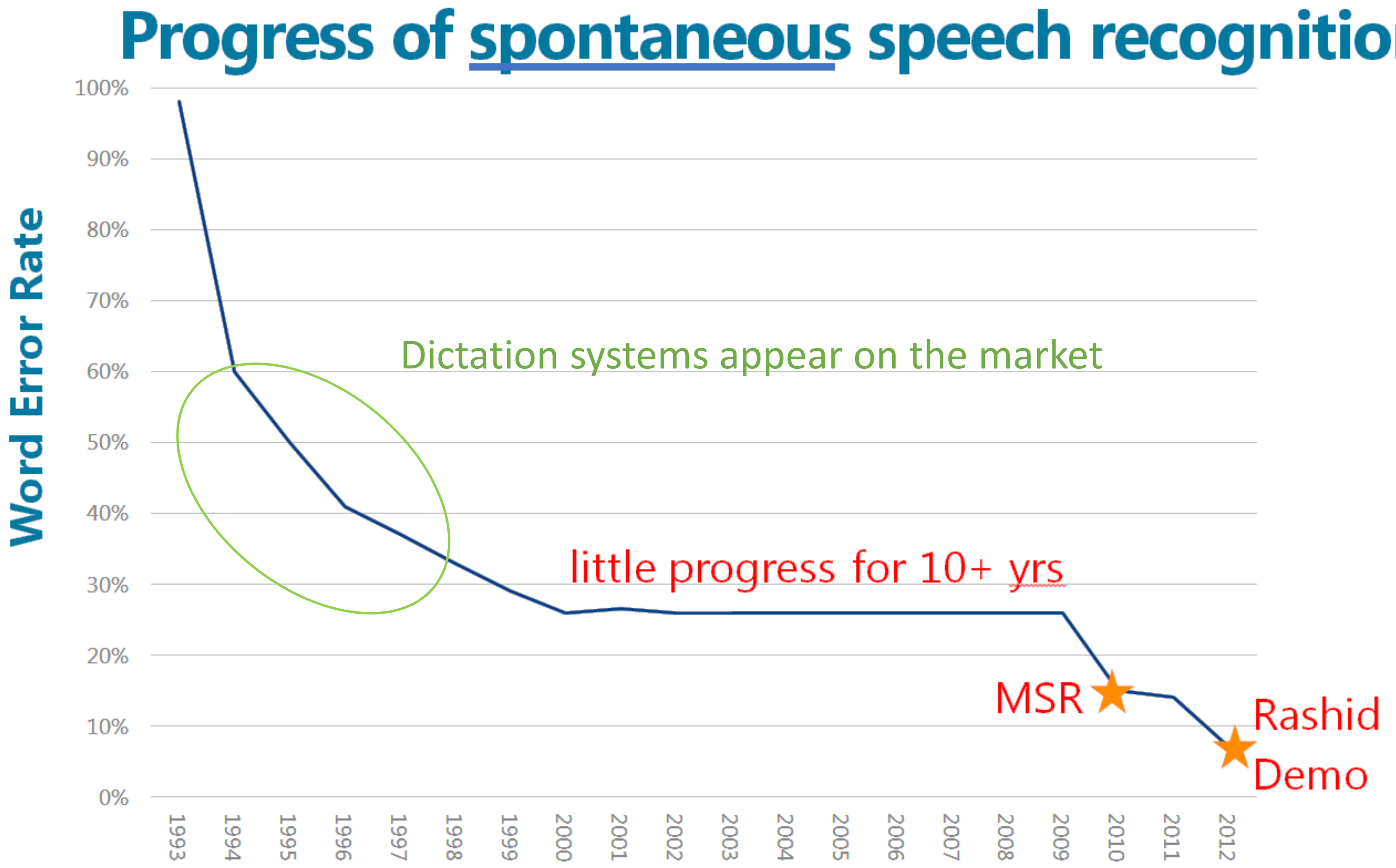
2011 -

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.



19

Deep Neural Networks

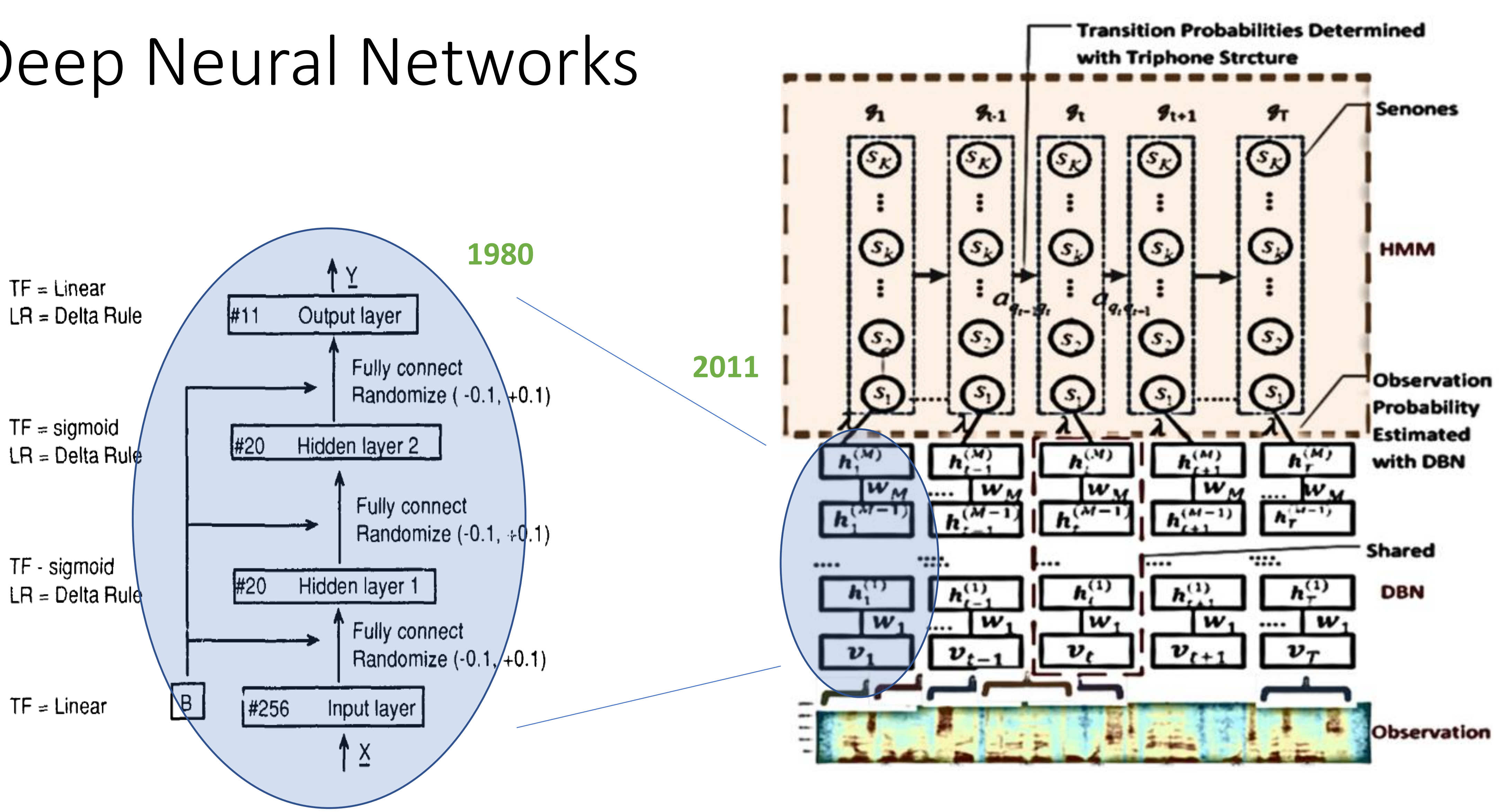
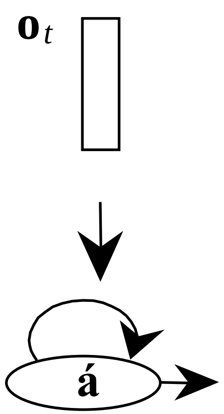
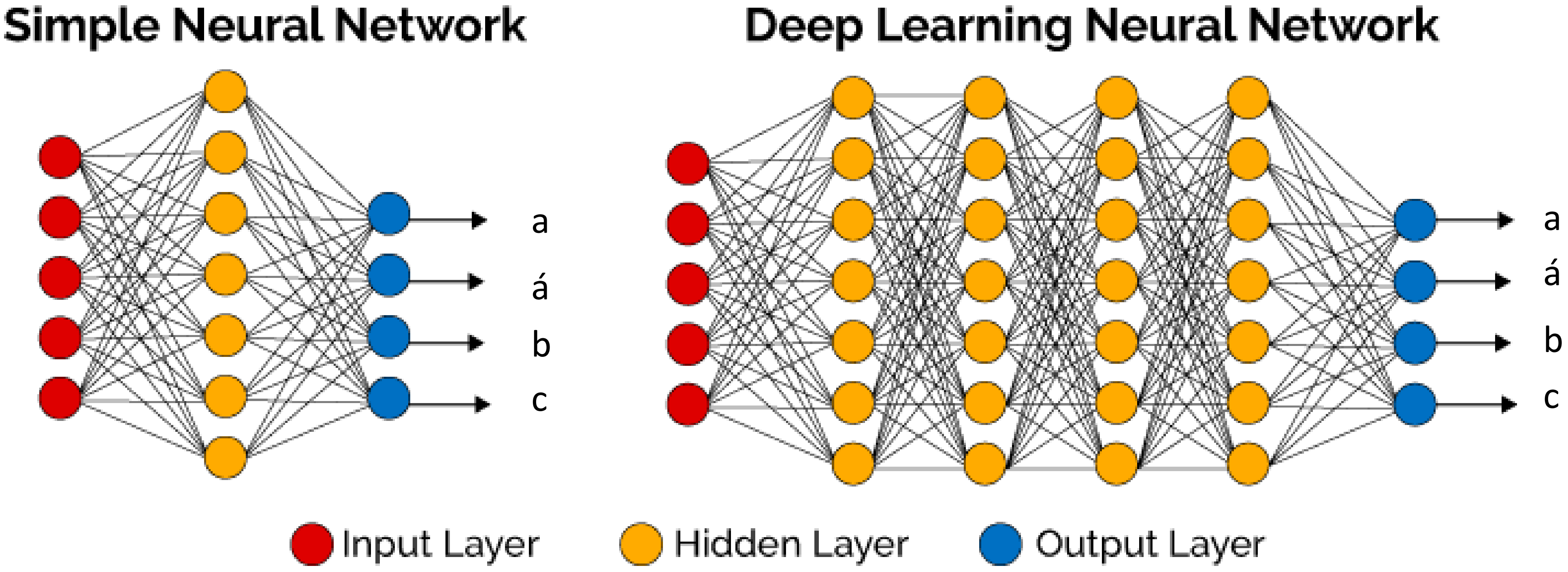


Figure 3. Back-propagation neural network topology

Deep learning acoustic models

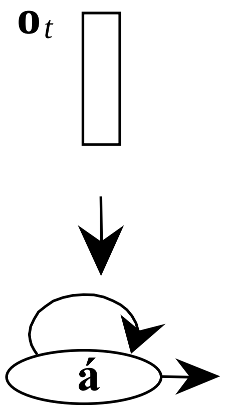
- Deeper structures – higher abstraction

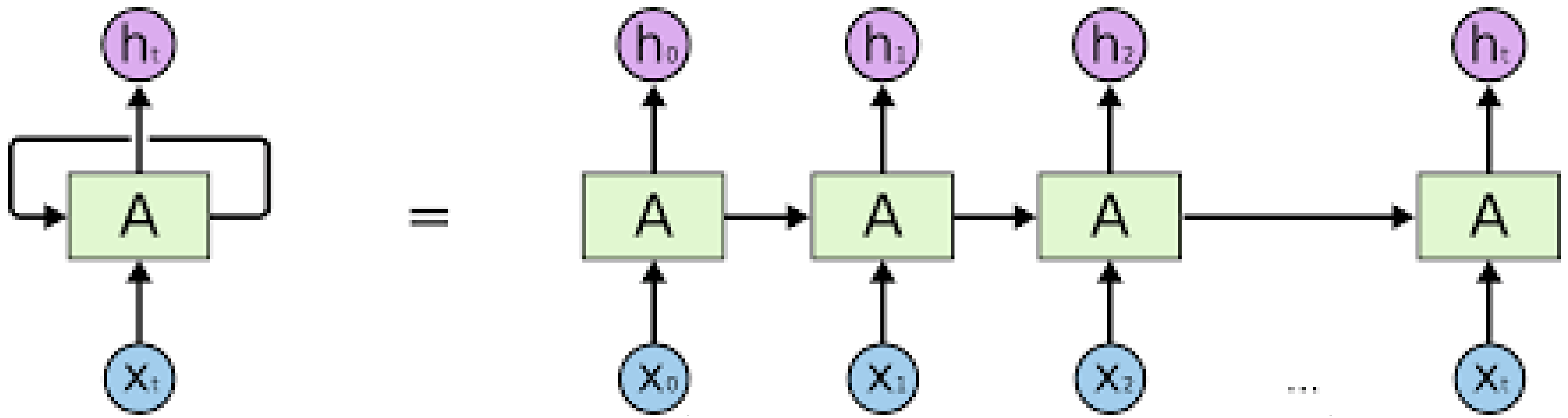




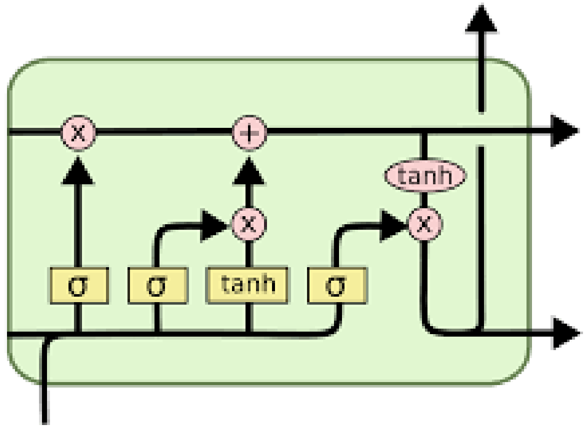
Deep learning acoustic models (2)

- Recurrent structure – „we don't forget what has happened before"



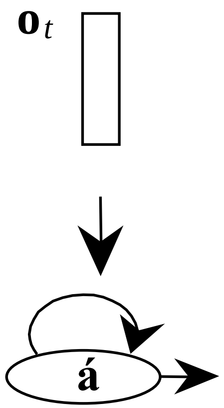


LSTM (Long Short-Term Memory)



Deep learning acoustic models (3)

- Do we really need to remember everything from the past?



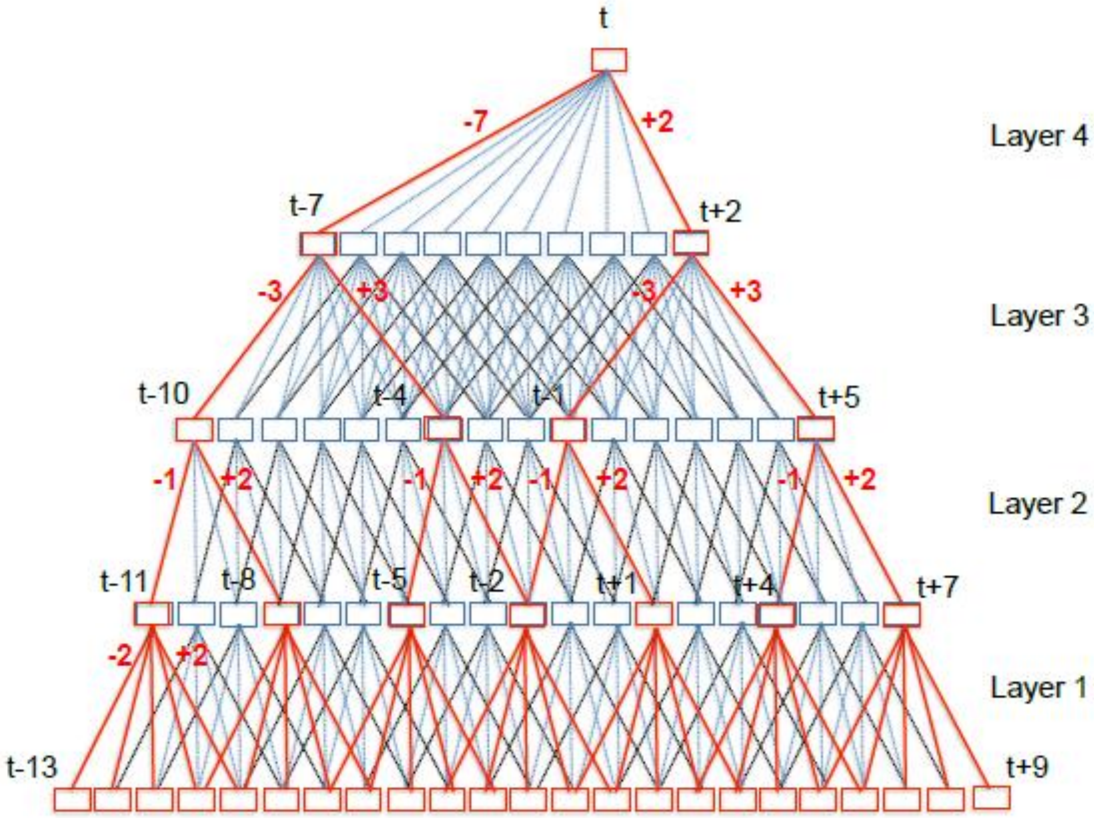
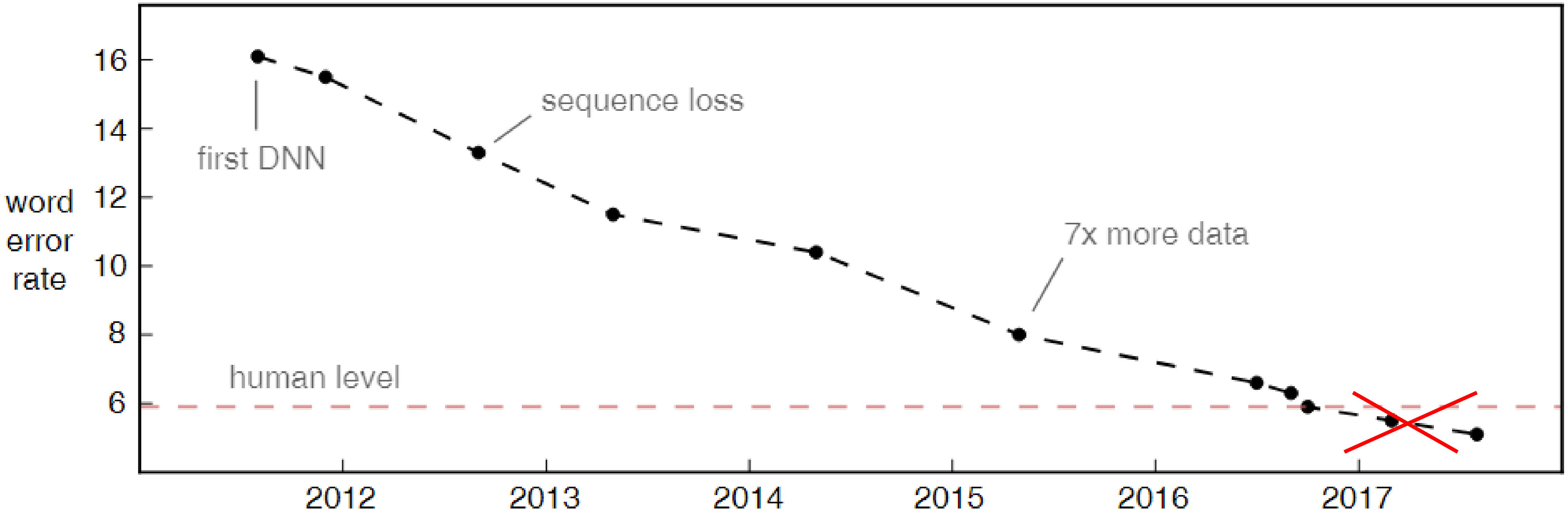


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

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

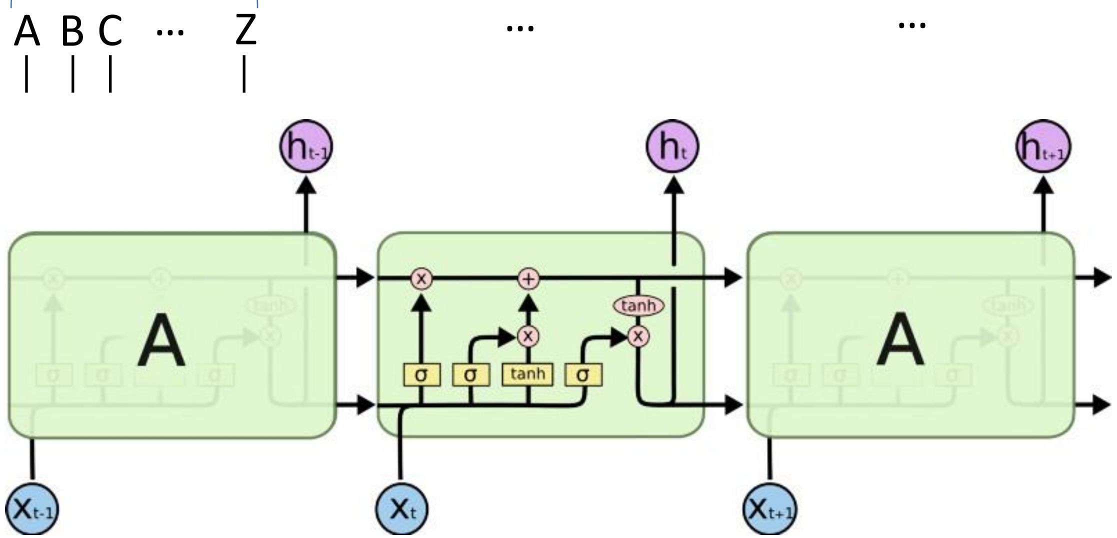
mihajlik.peter@vik.bme.hu

Basic idea: sequence 2 sequence modeling using recurrent nets

- LSTM

Highest probability?

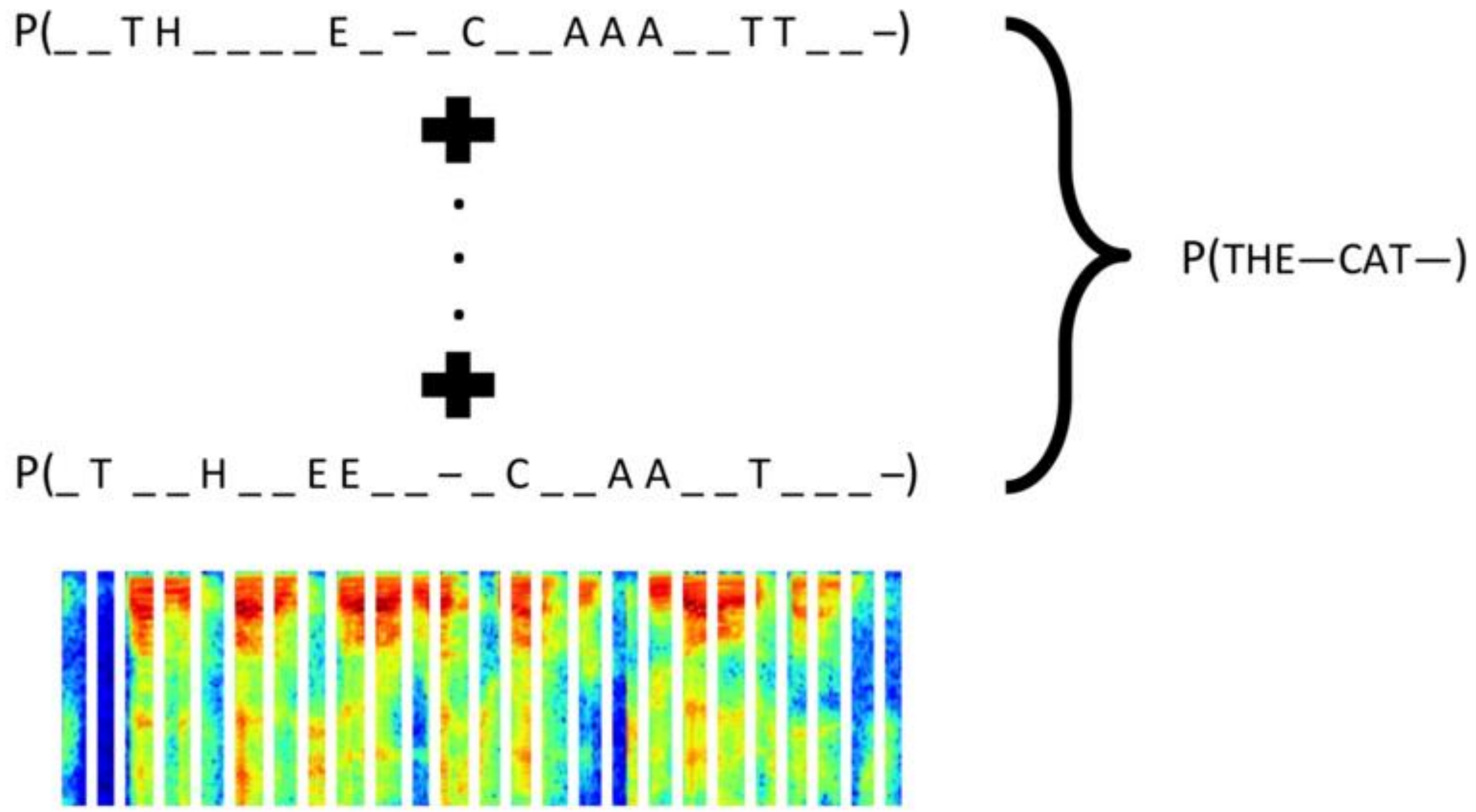
Text (chars, words, word fragments ...)



Acoustic feature vectors

The challenge: time alignment

„Connectionist Temporal Classification”

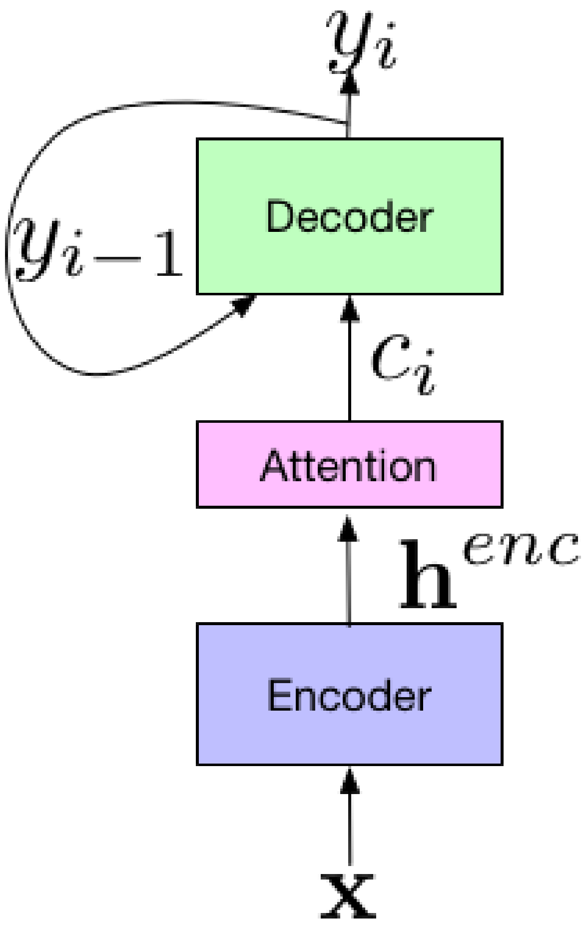


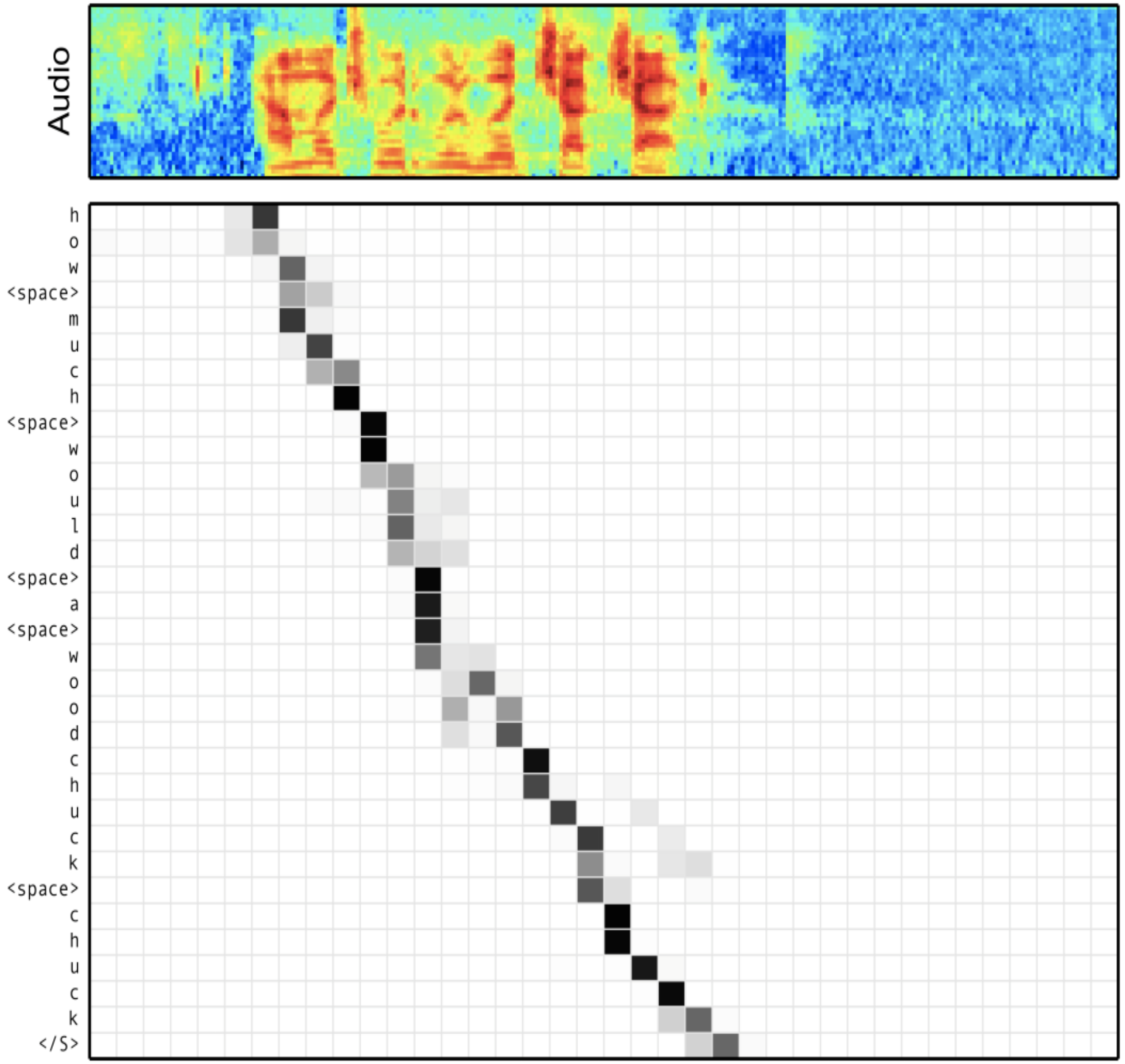
Listen – Attend – Spell (LAS) end-to-end (2016)

LM

Time alignment

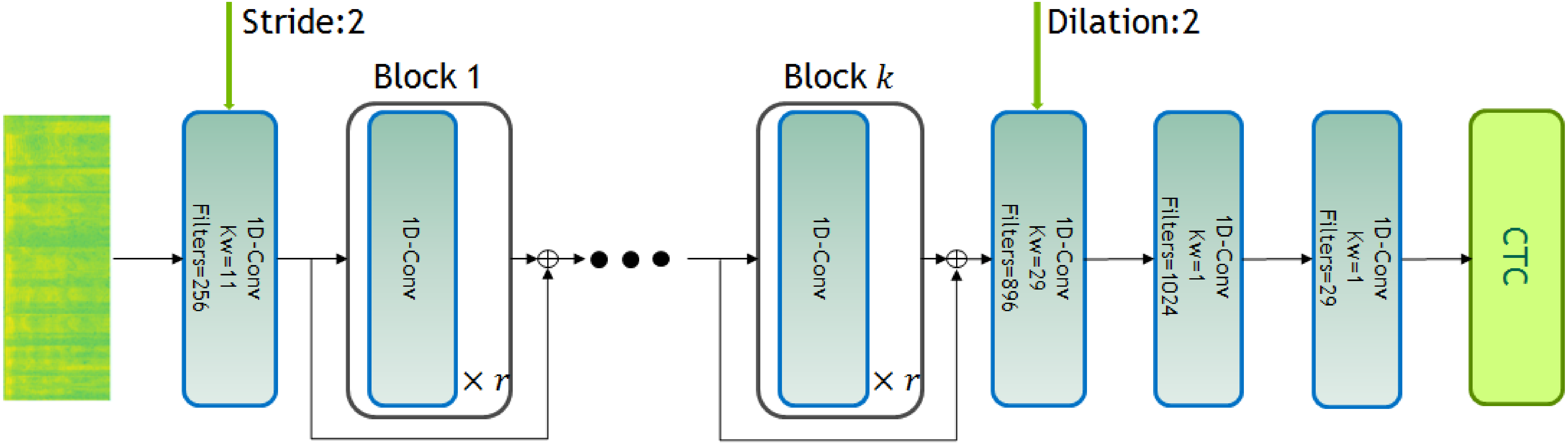
Acoustic
preprocessing





Convolutional end-to-end (2019)

- NVIDIA – Jasper (Just Another Speech Recognizer)

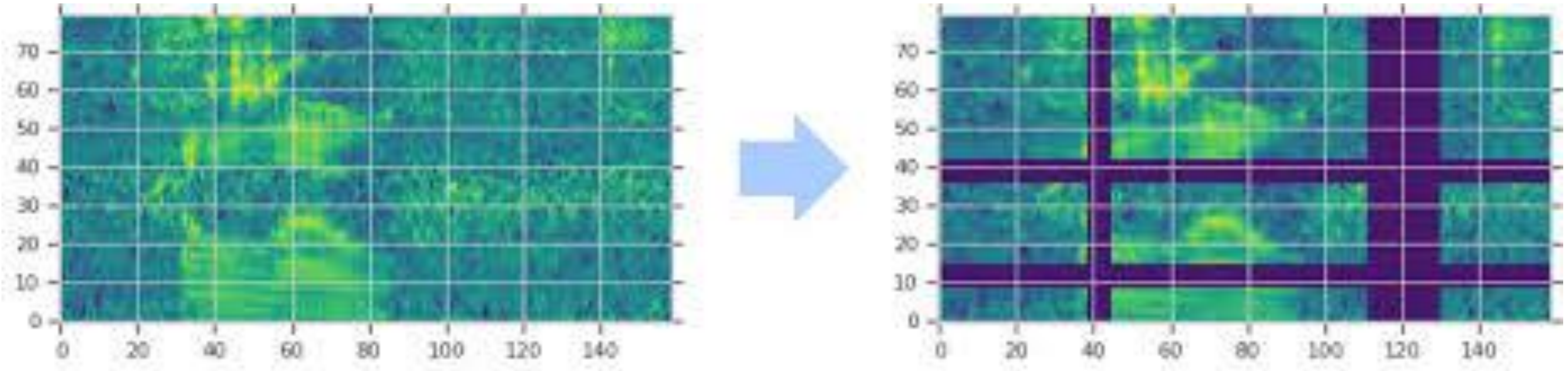


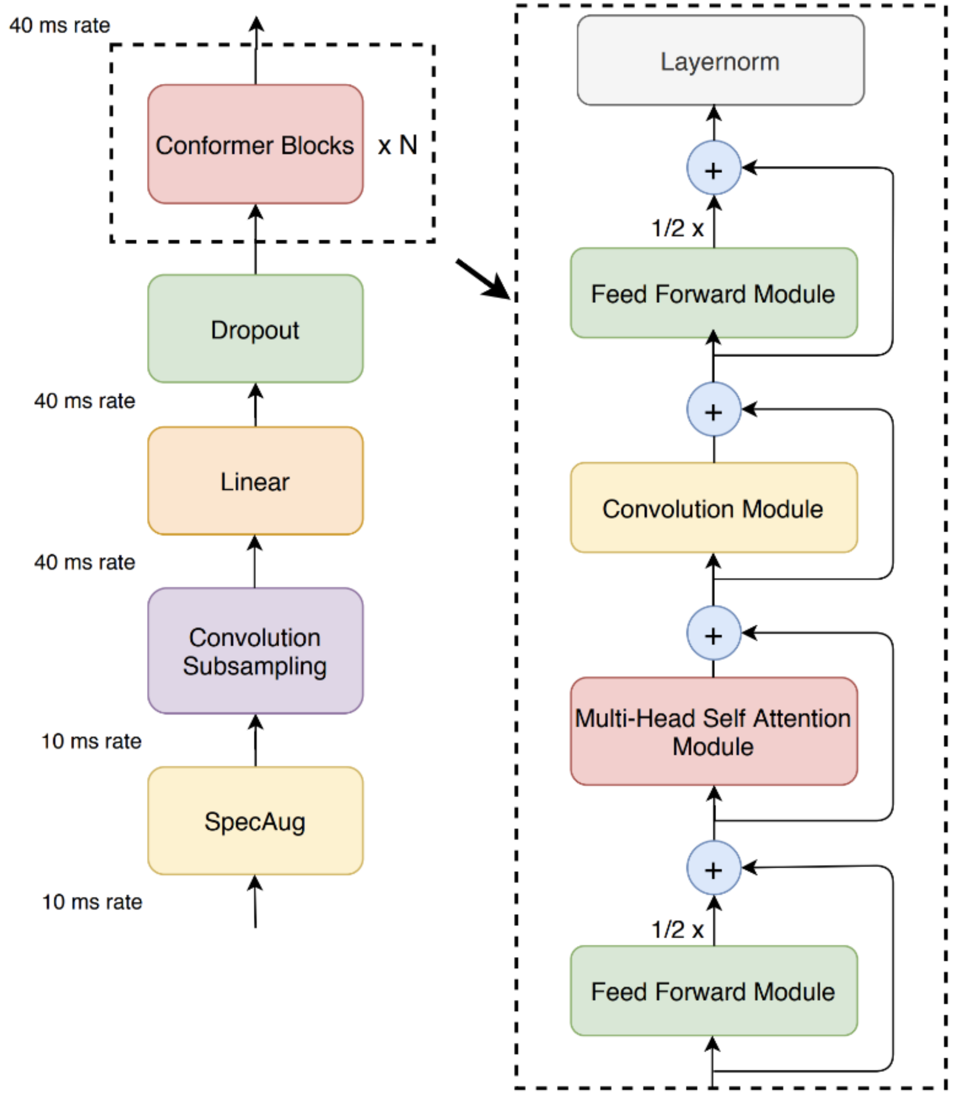
SmartLab
Speech Recognition
mihajlik.peter@vik.bme.hu

Data augmentation

- Speed perturbation
- Noise addition
- Room Impulse Response
- …
- <u>Spectral masking</u>!

SpecAugment (2019)





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

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)!

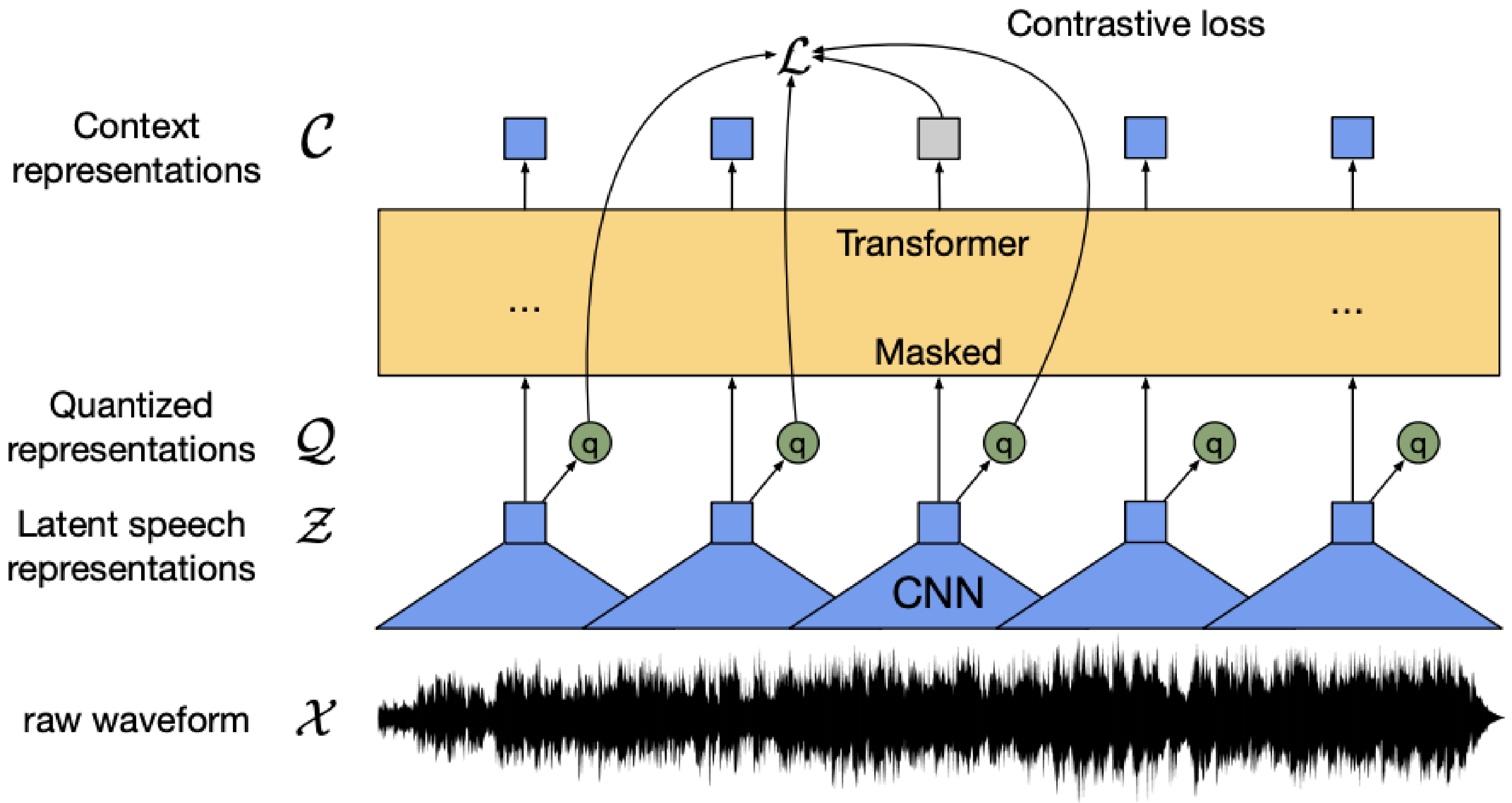
Unsupervised pre-training on pure acoustic data?

- Retsricted Boltzmann-machines (outdated)
- ?

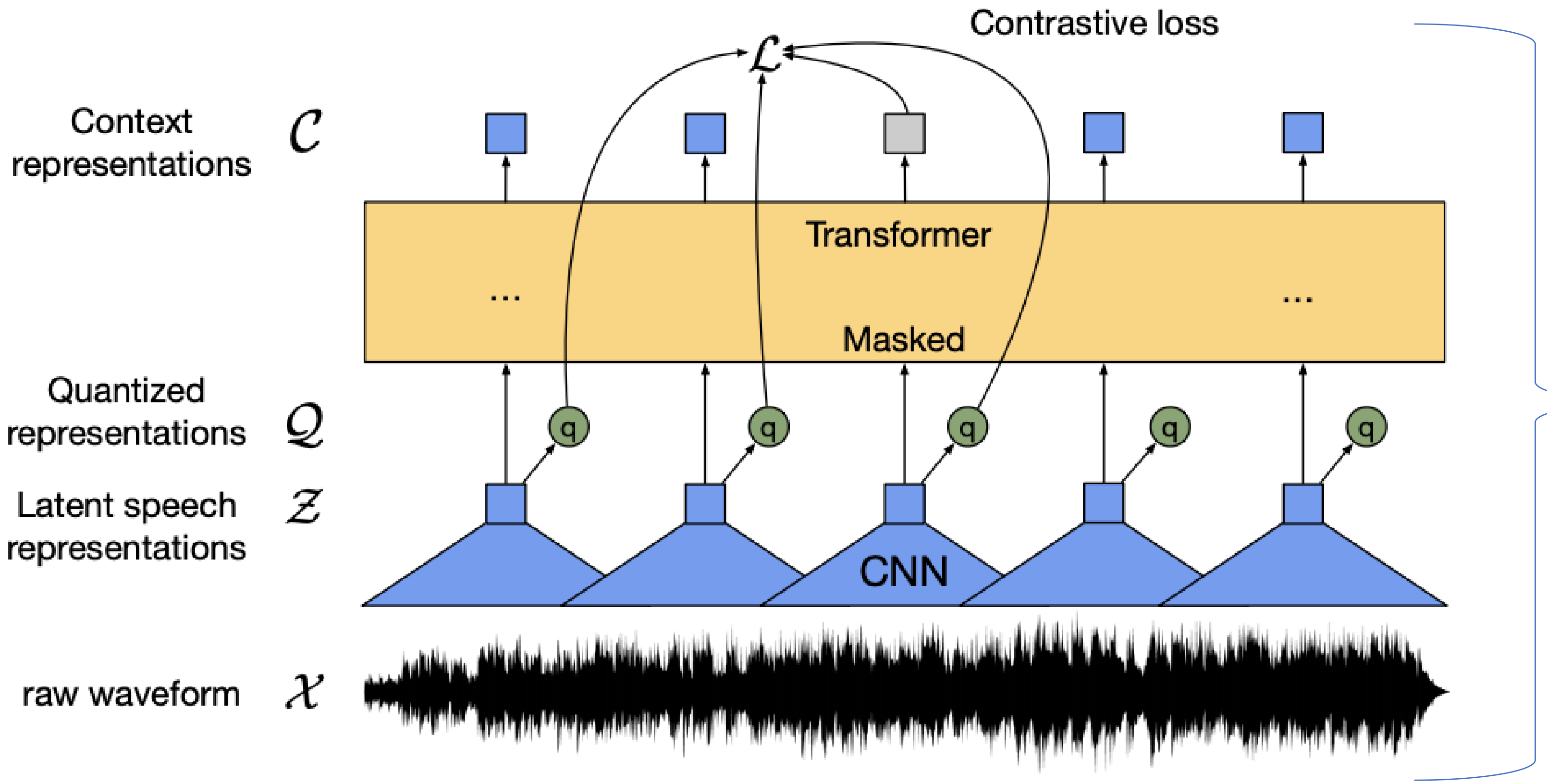
Self-supervised pre-training!

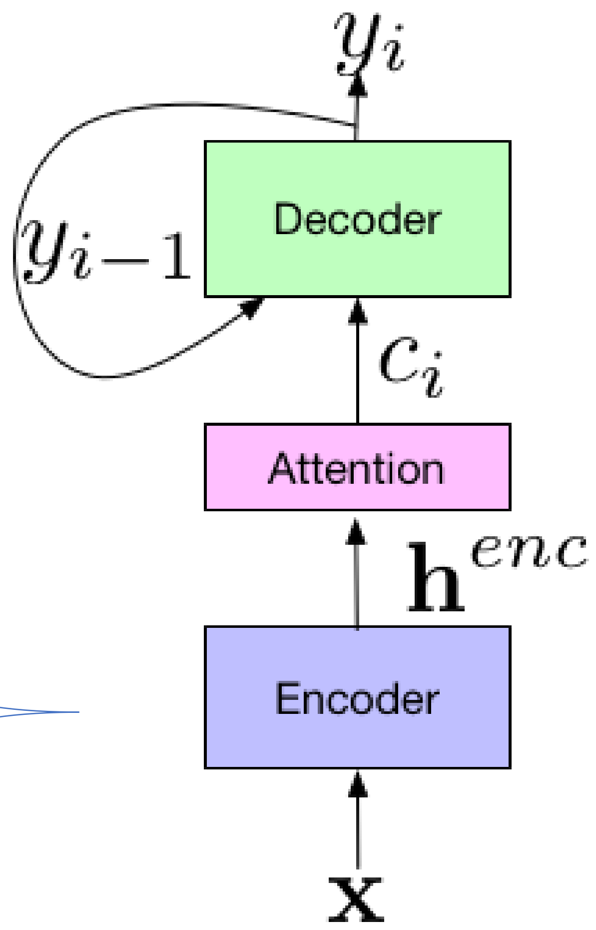
- Based on the very successful BERT training…

Wav2vec2.0


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

Wav2vec2.0 + fine tuning





In practice…

Lots of tools, more and more data

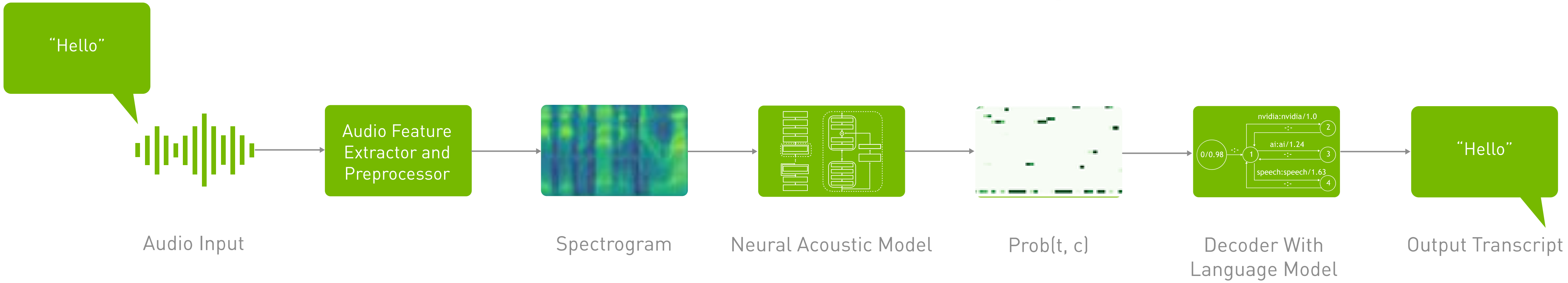
- Hybrid

KALDI

- End-to-end

SpeechBrain

FAIRSEQ



Languages: English, Deutsch, русский, français, italiano, Español ...

NVIDIA NEMO

Google

tensorflow/lingvo

moz://a DeepSpeech

Common Voice moz://a

Summary

Using a deep learning (ASR) toolkit can be hard in real life…

… but it is getting easier!

No ASR without deep neural nets.

Wav2vec self supervised pretraining + fine tuning seems unbeatable

Fully unsupervised techiques are coming!

Deployment needs simpler models.

If you are interested in ASR e-mail me.

mihajlik.peter@vik.bme.hu

Questions? Remarks?

Thank you.

Librispeech training process

0h	Prediction: h'otodtozaorodozortafzogoaronorhf rngoaoahnaoacazdntoazarmanazarazaglalanagad … Reference: i am so glad we met them so we drove along talking together we each assured the girl …
40h	Prediction: Reference: she can't help it and the funny thing is i don't believe that in her heart she is capable of …
240h	Prediction: m a t b a ts fots o an an ts sen Reference: thaddeus i i had a letter from jehiel to day you did and never told me why harriet what he …
700h	Prediction: form nt tis as the bots drown u pon the shaltof pea seem let mear twoise than they woulds ban … Reference: from that distance the boats drawn upon the sheltered beach seemed like mere toys then they would span…
5000h	Prediction: another truth which his abscare t me i wished to know if man constisfy you for broken vows with other… Reference: another truth which is obscure to me i wish to know if man can satisfy you for broken vows with other…
200Kh	Prediction: their upper jaw they move wonder if tom rockford will do anything with that invention of his wasting… Reference: their upper jaw they move wonder if tom rochford will do anything with that invention of his wasting …