Mesterséges intelligencia a beszédtartalom-felismerésben

Mihajlik Péter

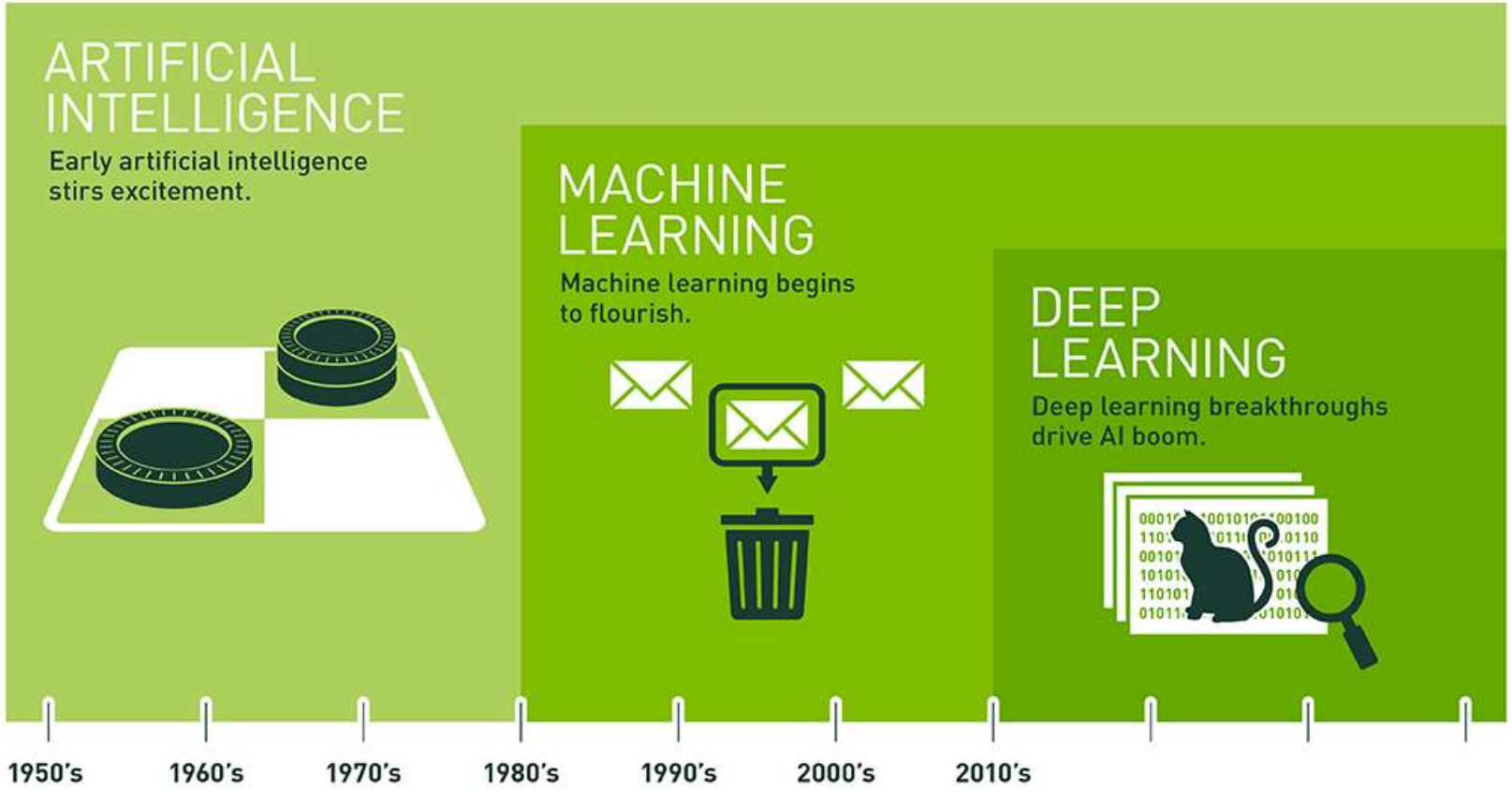
Beszédfelismerés Labor

SmartLab
Speech Recognition
https://www.tmit.bme.hu/lsr

Beszédtartalom-felismerés

- *Csak hangból*
- **<u>Szövegtartalom (Speech2Text)</u>**
- <u>Mikor ki beszél (Diarization)</u>
- Kora, neme
- Emocionális állapot
- …

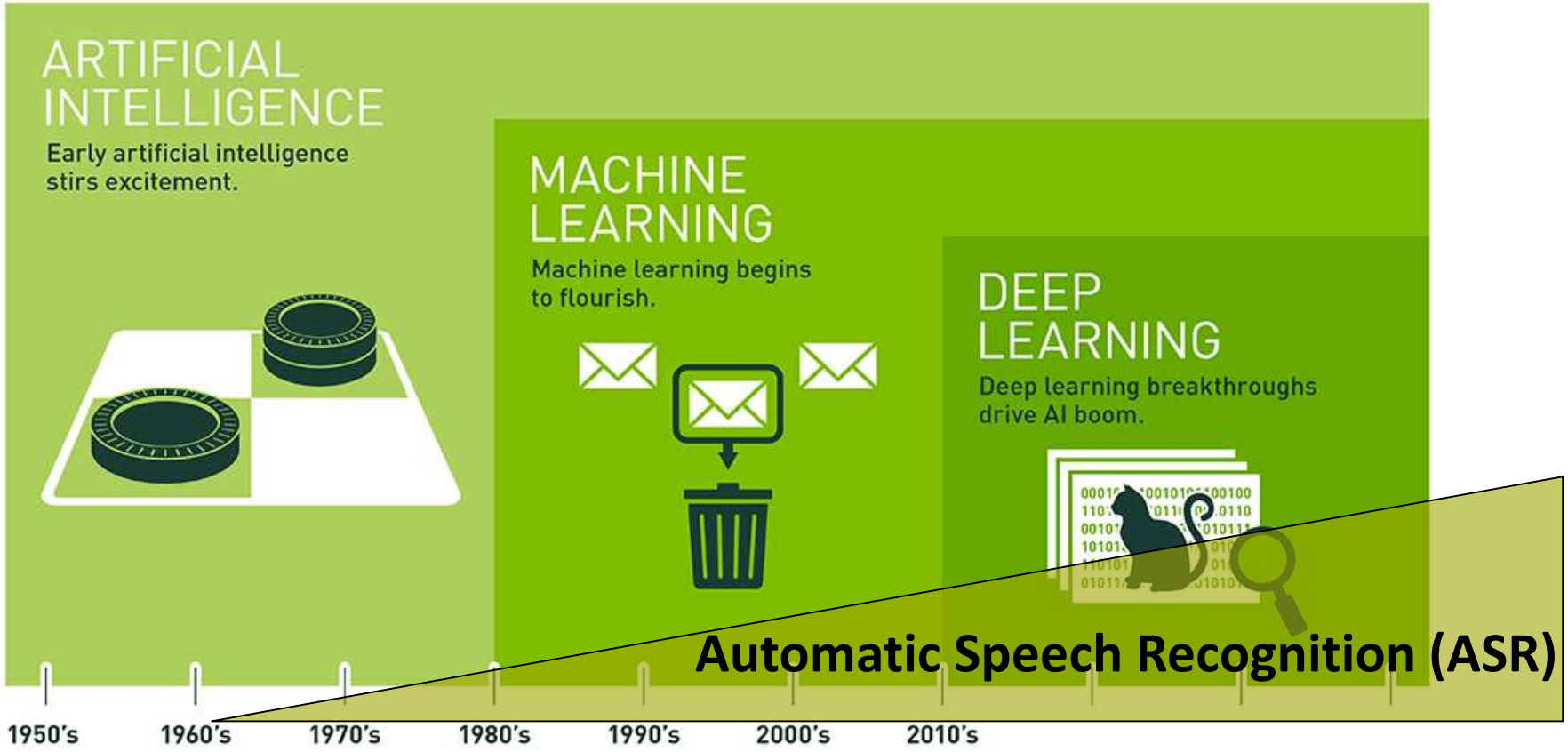
Mesterséges intelligencia



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Mesterséges intelligencia…

… és beszédfelismerés



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Mesterséges intelligencia a beszédfelismerésben

1990-2010

SmartLab
Speech Recognition

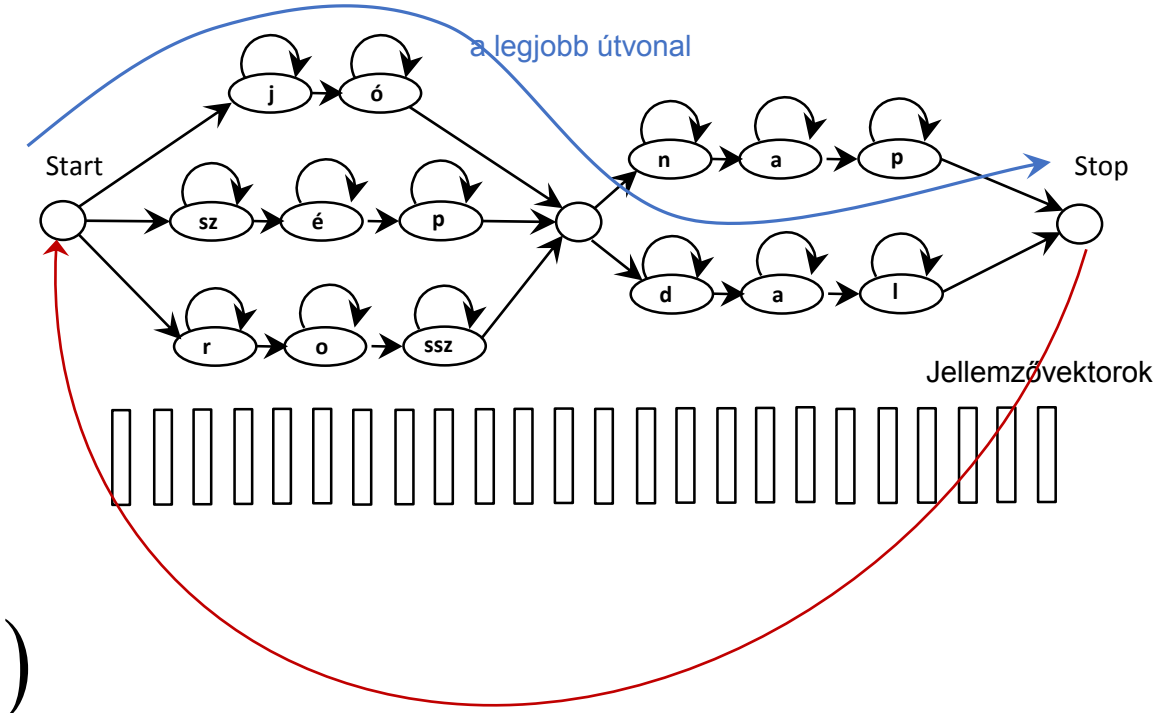
mihajlik@tmit.bme.hu

Rejtett Markov-modellek (HMM)

- Valódi ML (Machine Learning) technika
- Adat/statisztika vezérelt

"Every time I fire a linguist, the performance of the speech recognizer goes up".

~ 1975: IBM, Fred Jelinek



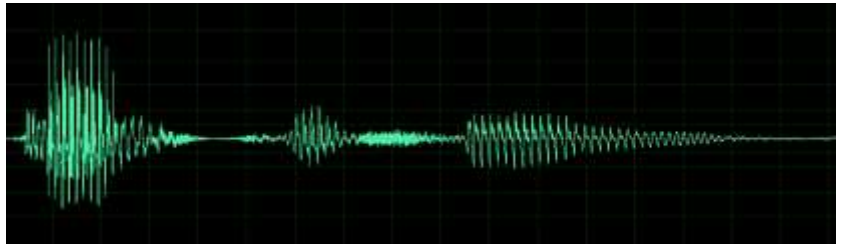
LM AM

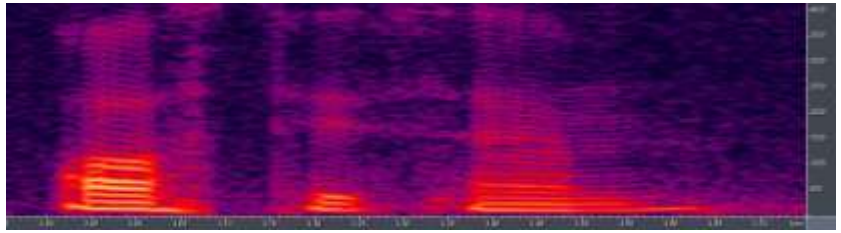
$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W)$$

A hangjel megfelelő reprezentálása

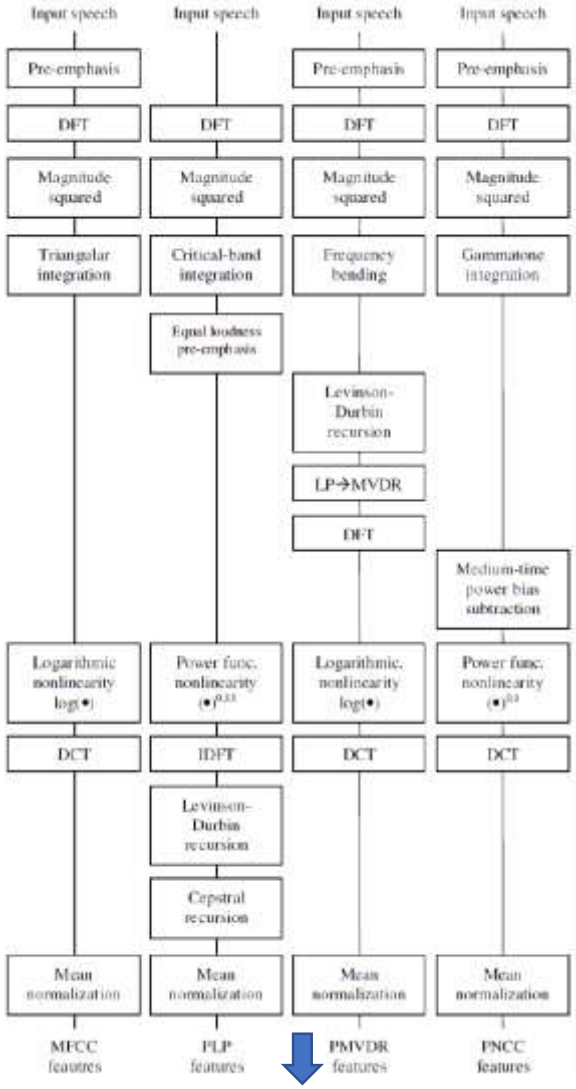
- Mérnöki intuíciókon alapuló jelfeldolgozás

„ a z t h i sz e m ”





MFCC / LPC / PLP …

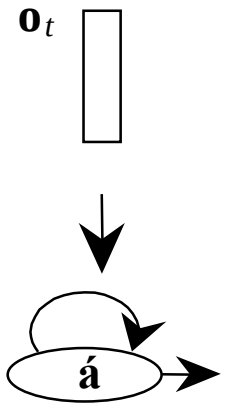


…

…

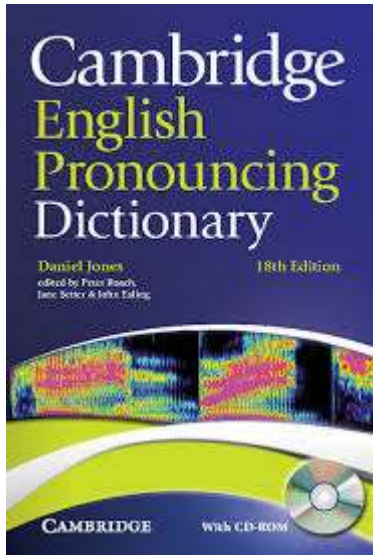
Hangmodellek

- Akusztikai hasonlóság mérése – beszédadatok egyszerű statisztikája alapján



Kiejtési szótár

- Fonéma alapon…



hagyja h a ggy a
hangya h a ny gy a

- ❑ Szabály alapon?
  - ❑ t + sz = c
- ❑ Kivételek?
  - ❑ Churchill = cs ö r cs i ll
- ❑ Kiejtési változatok?Gyakoriságok?
  - ❑ miért = m é r t, m i é r, stb…
- ❑ Ambiguitások?
  - ❑ Lachegyi vs. Lachema, malacsült, meggyógyít…

Koartikulációmodellezés

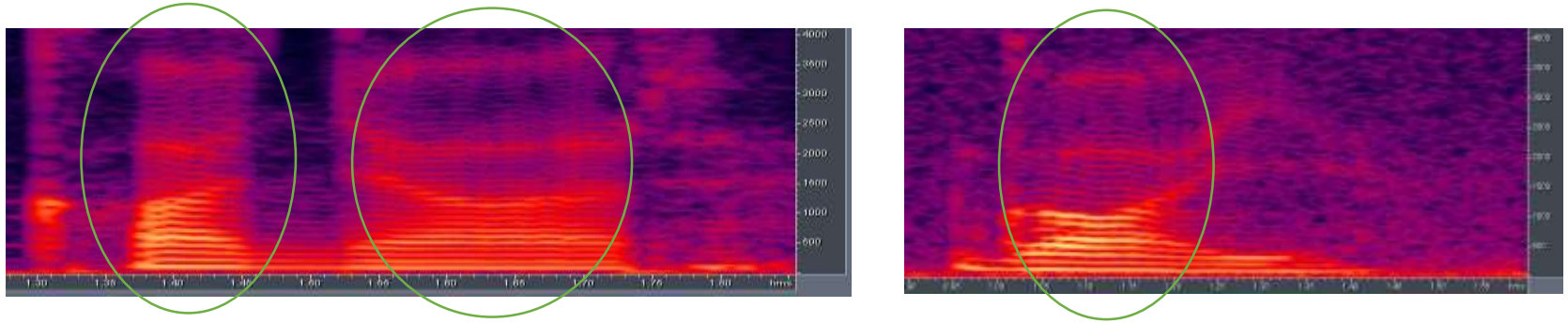
- Hasonulás, egybeolvadás, stb. – szóhatáron is!

értsd [te|ezt] →

é r dzs d sil t e
é r cs t e
é r dzs d [sil] e sz t …

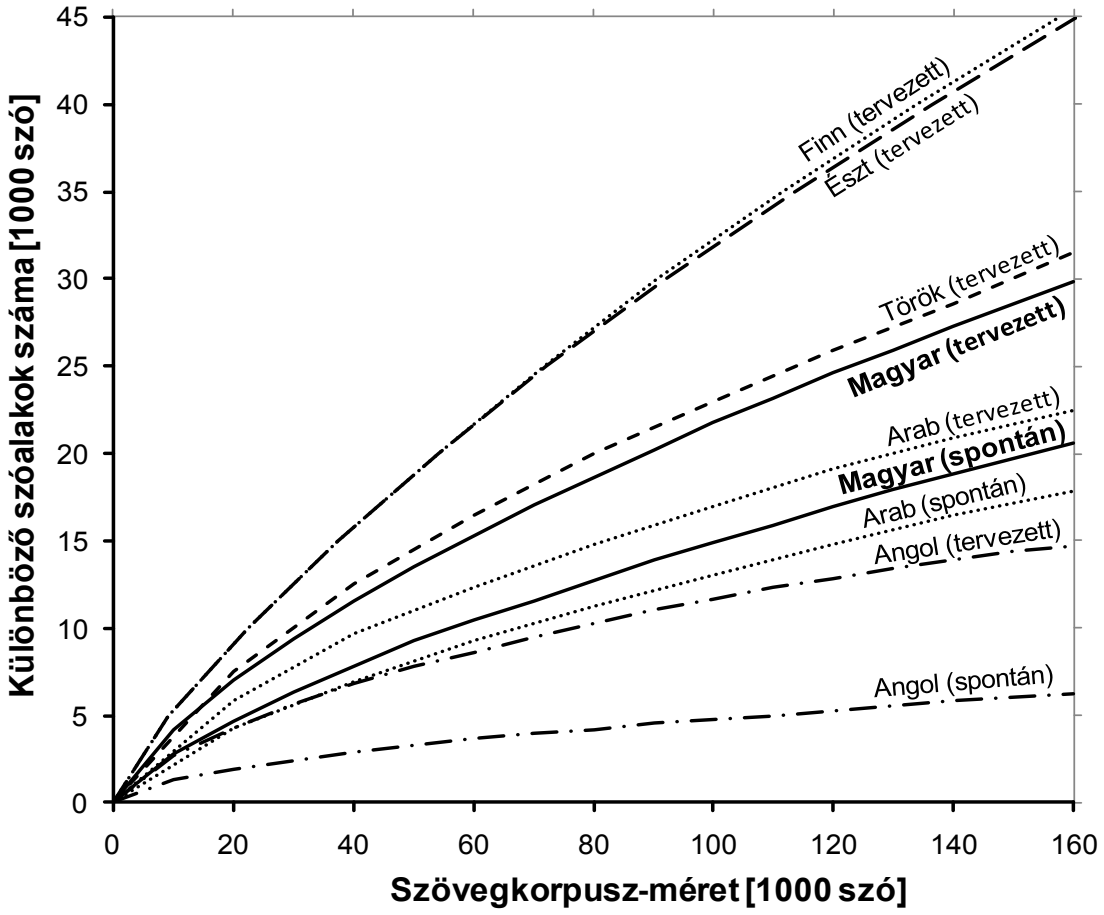
- Írásban nem jelölt változások

[k] [a] [ny] [a] [r] [g] [a] [lly]



Morfológia kezelése

- Darabolás szótőre, toldalékokra…



láthattuk kóstolhattuk milyen jól tudnak főzni ha akarnak

lát hat tuk kóstol hat tuk mily en jól tud nak főz ni ha akar nak

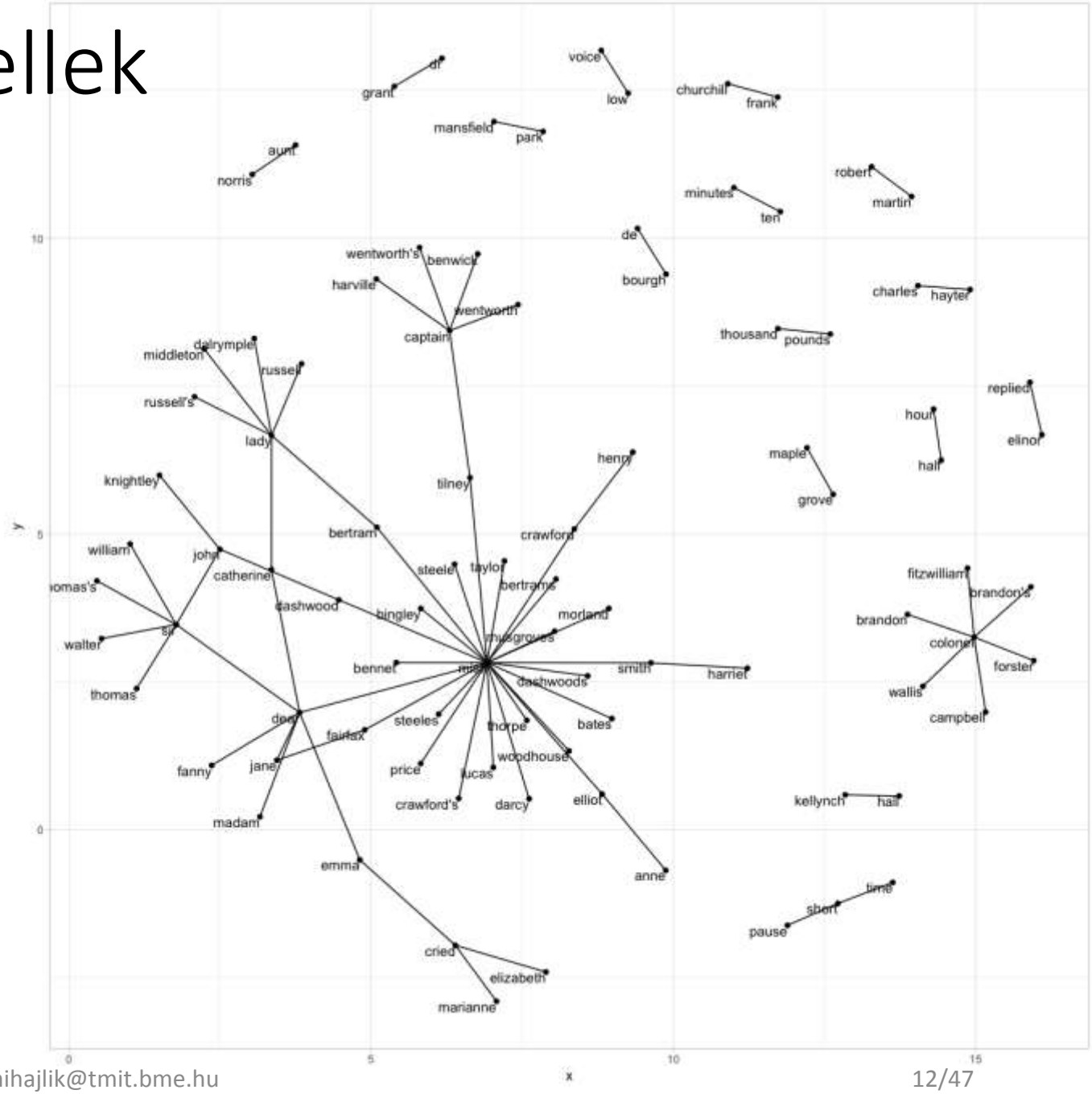
/we could see and taste how well they can cook if they want/

Módszer: szabály és/vagy statisztikai alapon

„Szókapcsolati” modellek

- N-gram modellek
- Szöveg-statisztika alapján
- Valószínűségi kapcsolat a szavak között!

- Language Models



Klasszikus ASR megközelítés

- Fonéma alapú
- Nyelvi tudás több szinten, nyelvi szakértő szükséges
- Jól elkülönülő explicit nyelvi szintek
- On-line, gyors
- Rugalmasan bővíthető
- Standard technológia (szabályok + statisztika, gépi tanulás)

- Gépi pontosság << Emberi pontosság

„A Deep Learning betörés”

2011 -

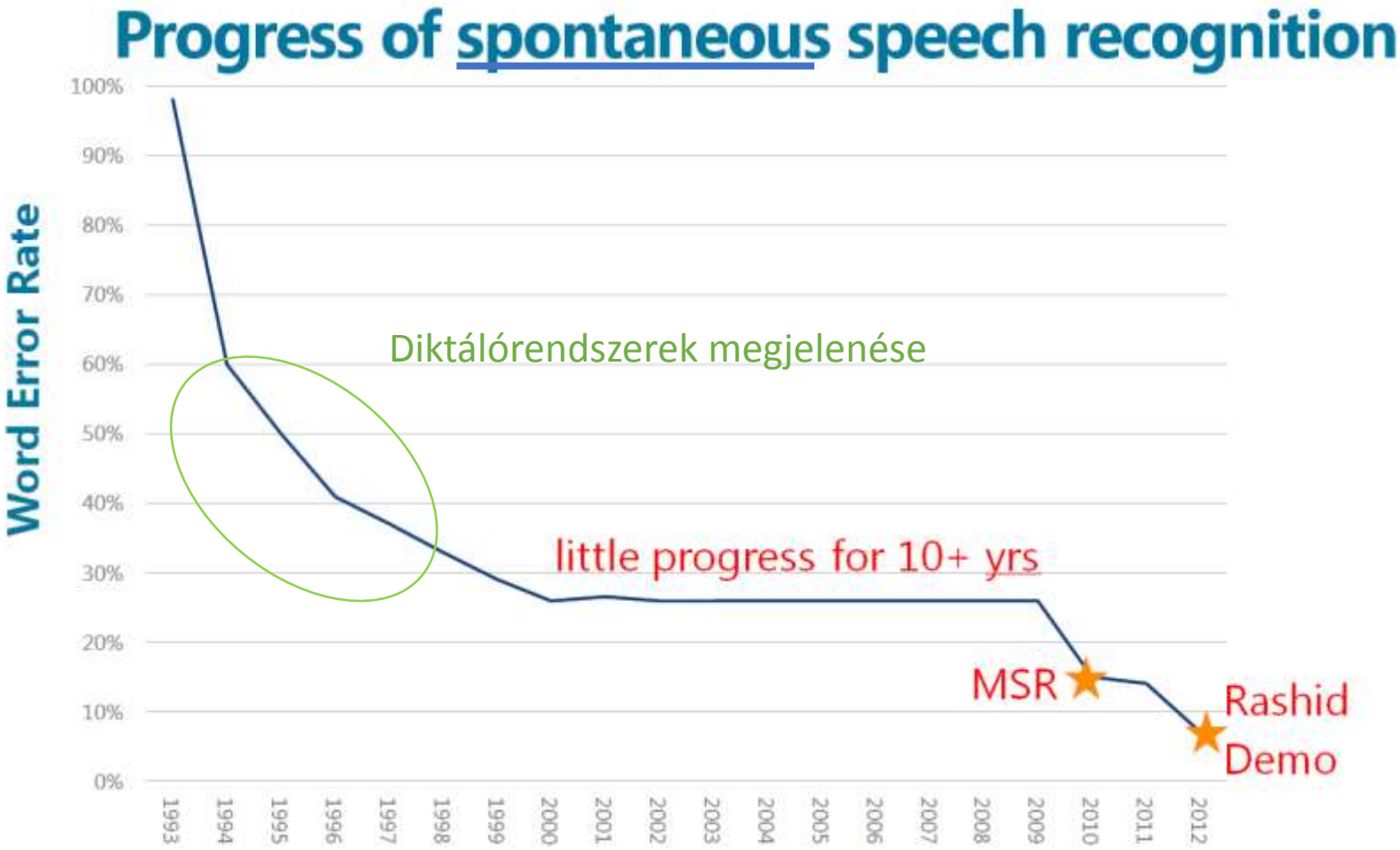
mihajlik@tmit.bme.hu

A Microsoft és a Rosetti-kő…

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.



15

Deep Neural Networks

1980-as évek

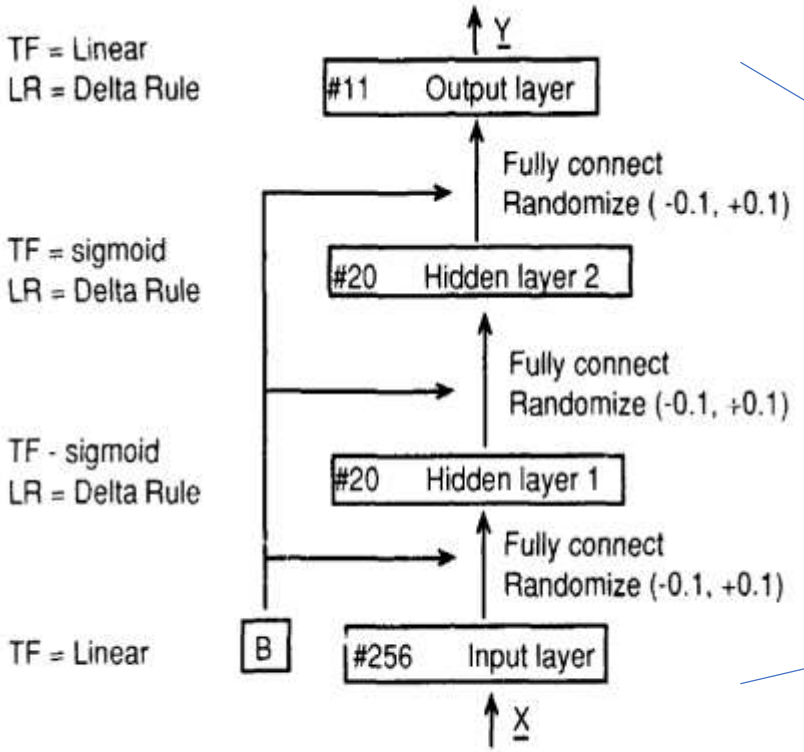
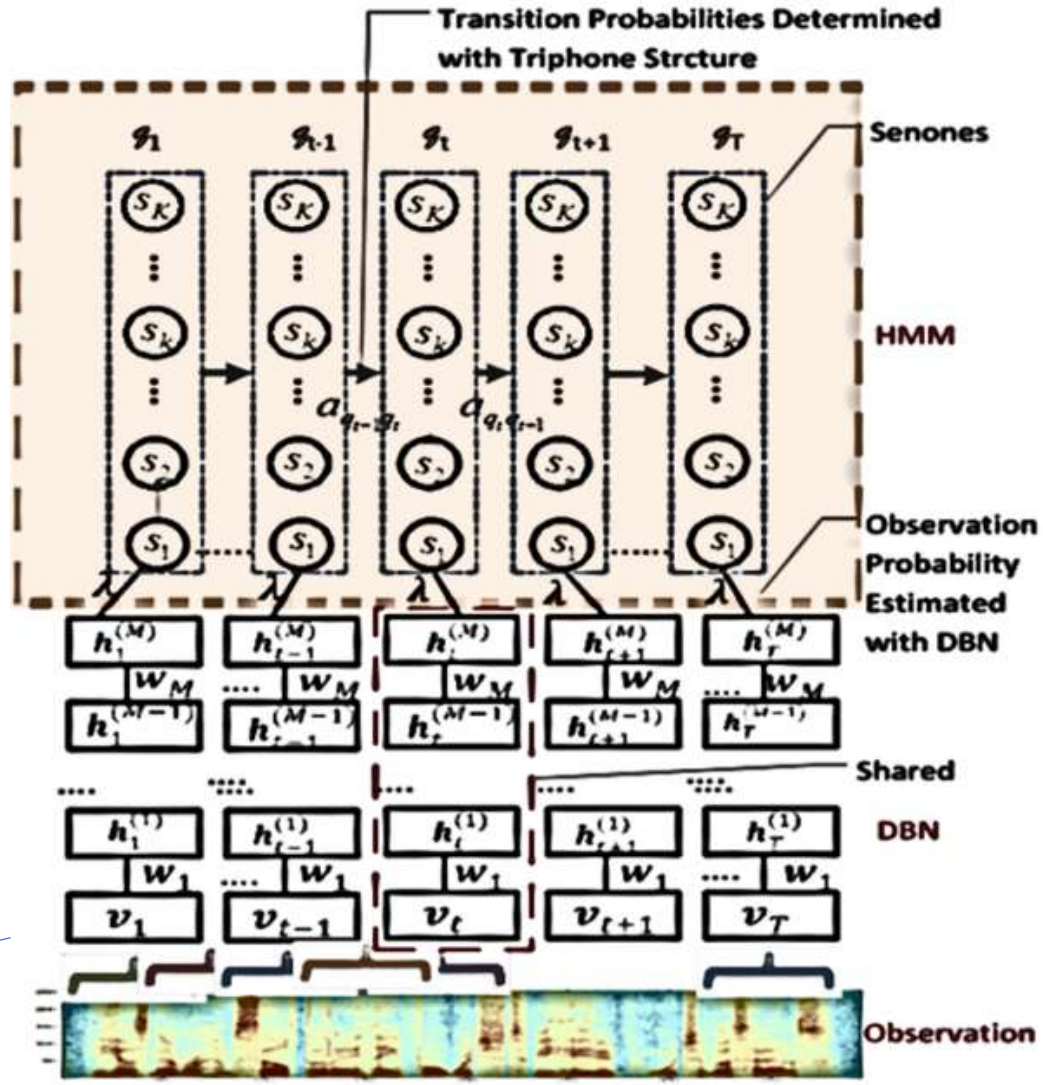


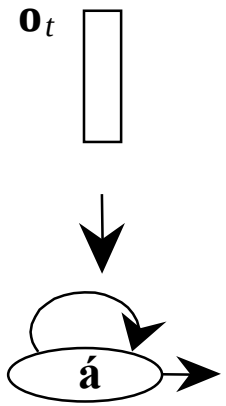
Figure 3. Back-propagation neural network topology

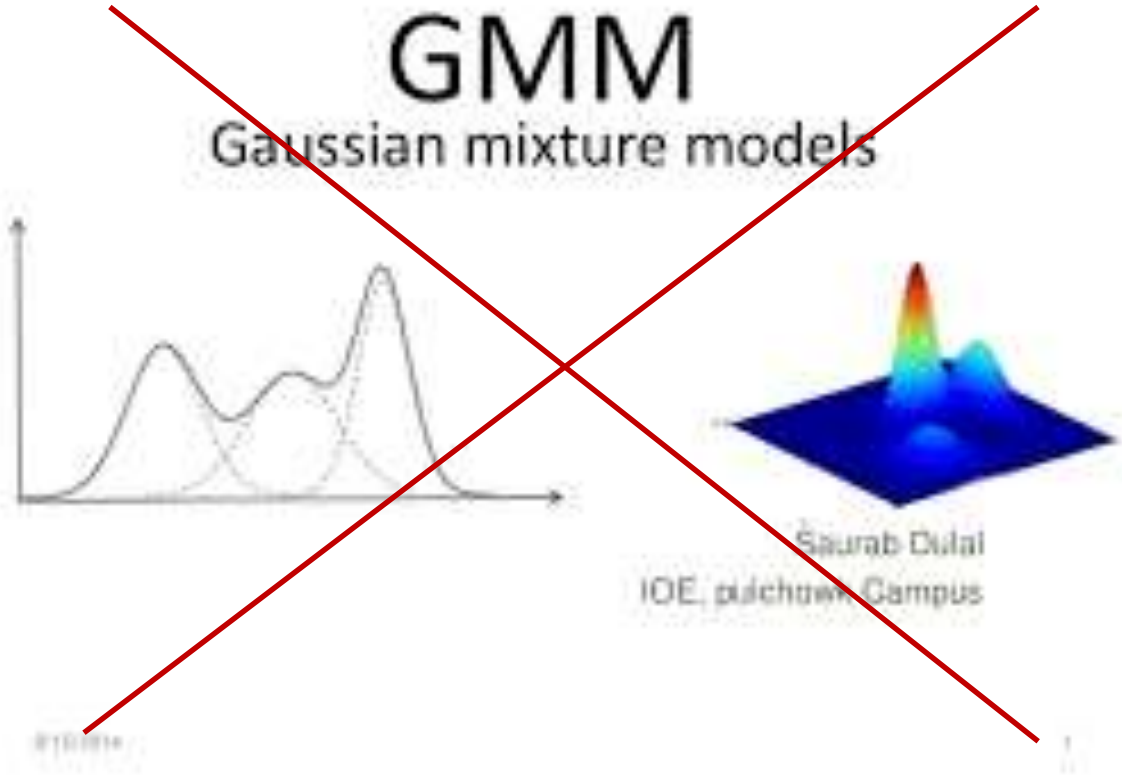


2010-es évek

Deep hangmodellek

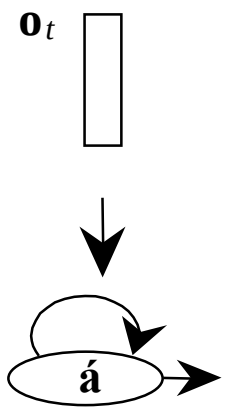
- Akusztikai hasonlóság mérése – beszédadatok mély tanulása alapján

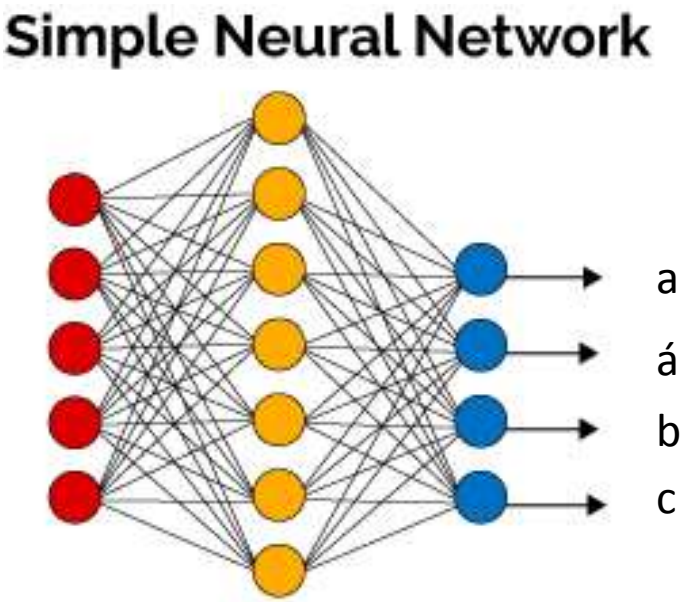


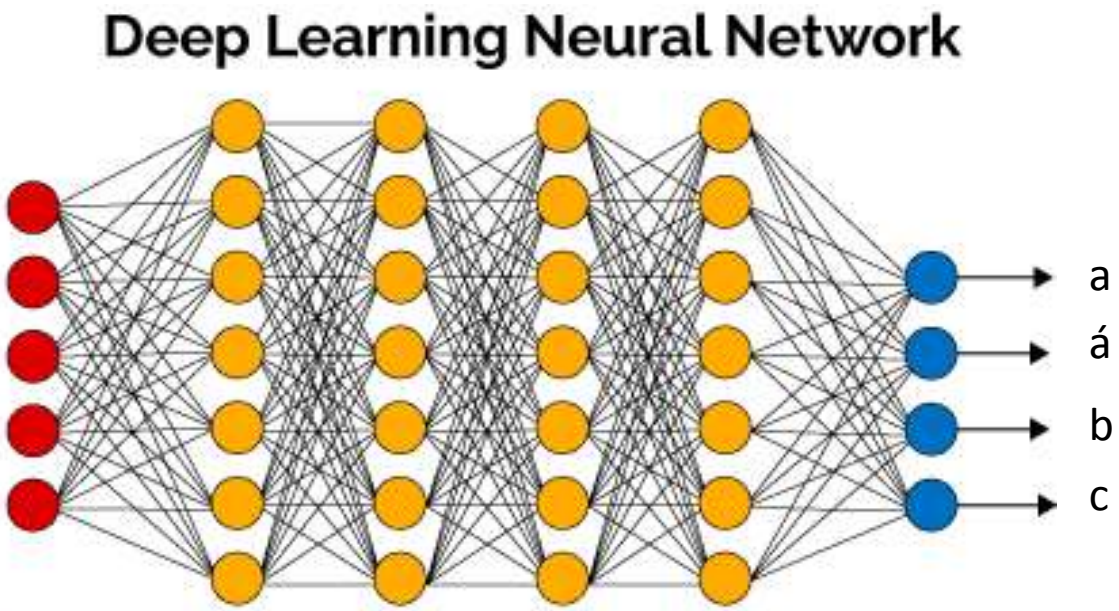


Deep hangmodellek

- Mélyebb tanulás – magasabb absztrakció



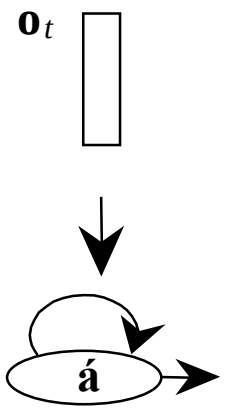


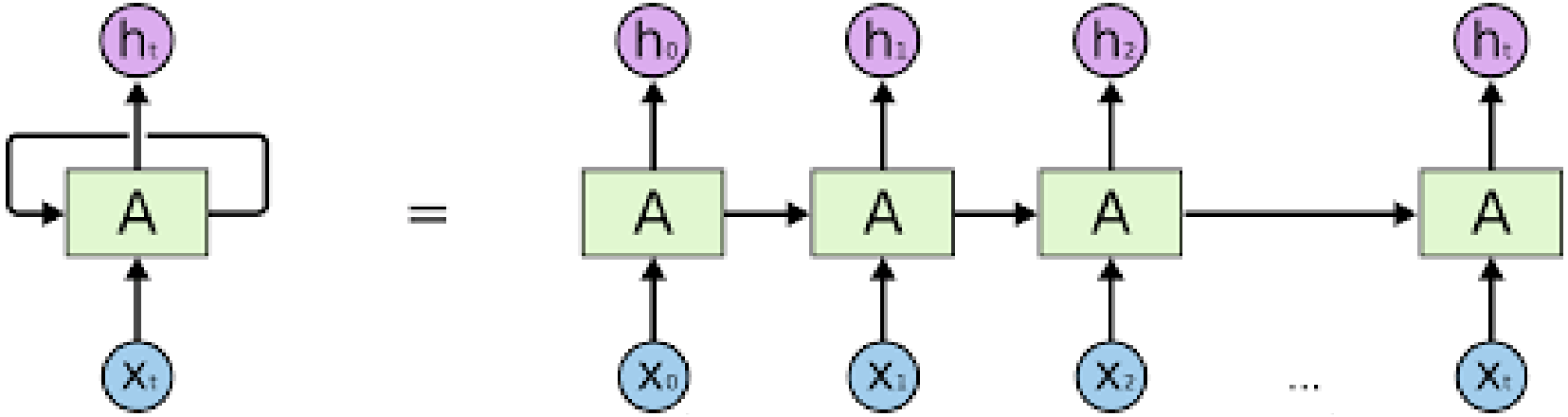


Input Layer Hidden Layer Output Layer

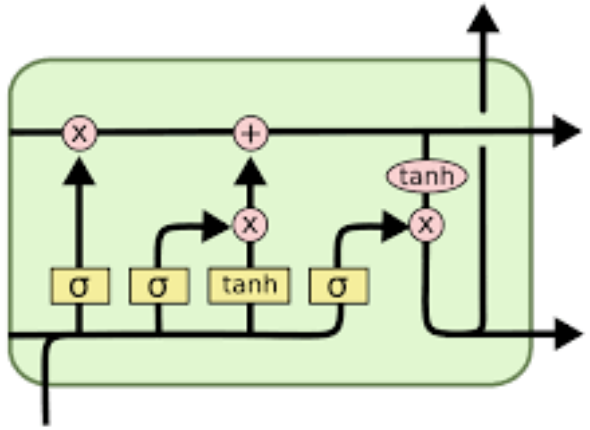
Deep hangmodellek (2)

- Rekurrens modell – nem felejtjük el, mi történt az előző lépésben



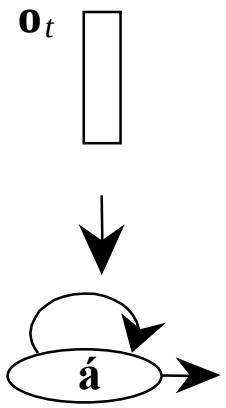


LSTM (Long Short-Term Memory)



Deep hangmodellek (3)

- Időkésleltetéses mély neuronhálózatok: alacsony szintű feldolgozás, nagy időátfogás



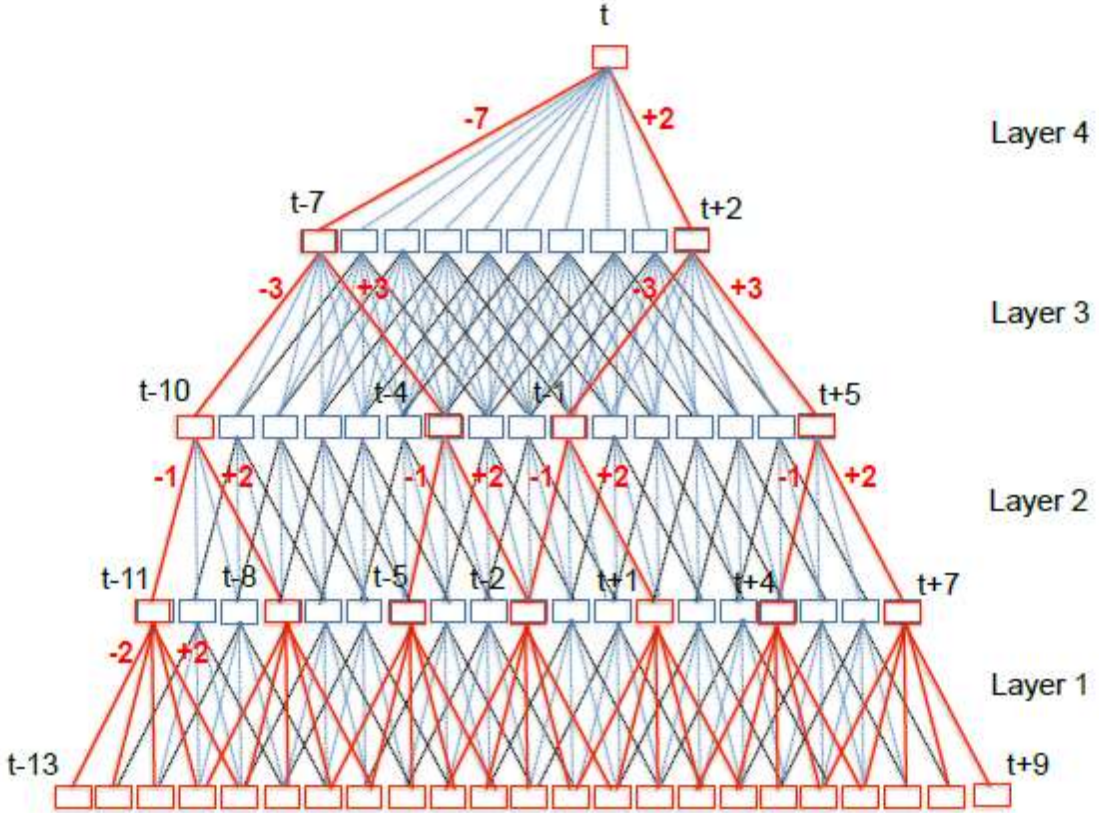
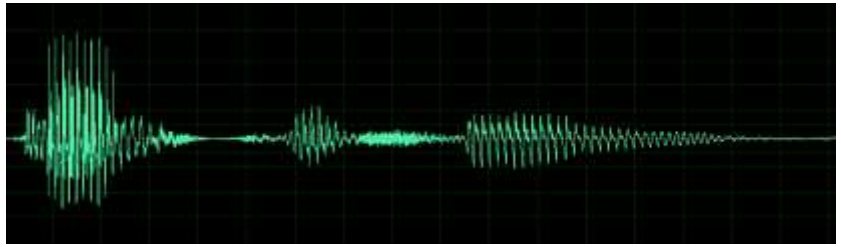


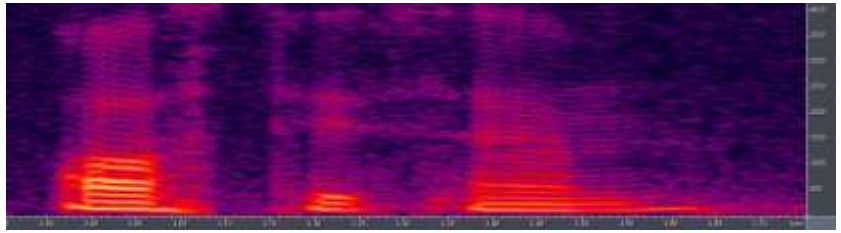
Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

Deep jelfeldolgozás

- ~~Mérnöki intuíciókon alapuló jelfeldolgozás~~

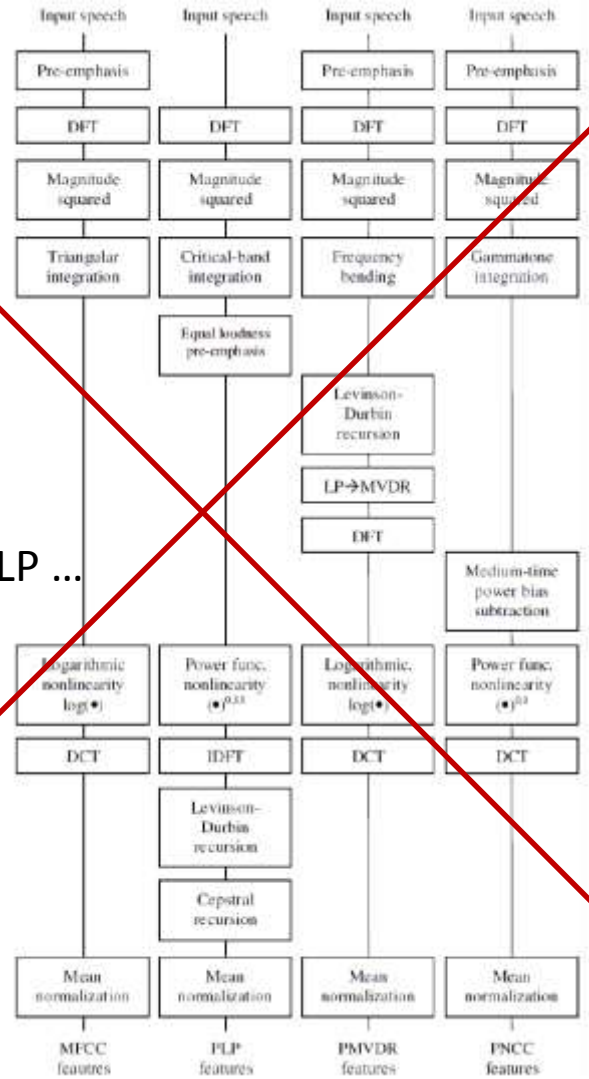
„ a z t h i sz e m ”





Spektrogram – idő/frekvencia kép

MFCC / LPC / PLP …



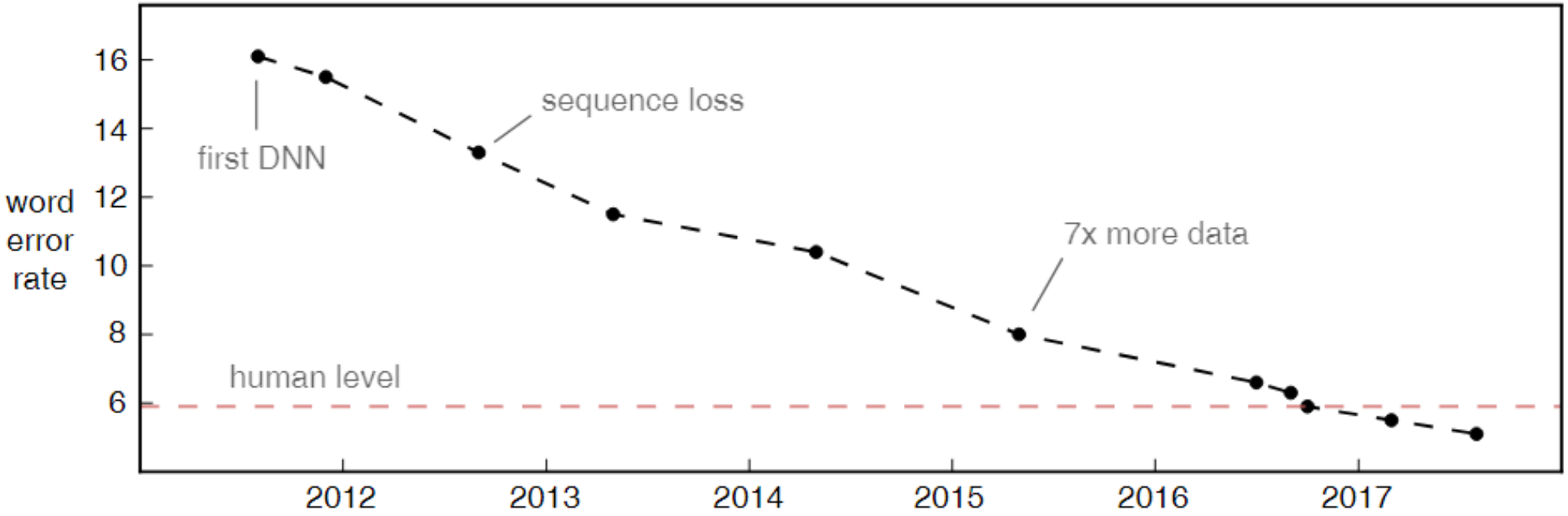
… …

Deep „nyelvtan”

- Kiejtés
- Koartikuláció
- Morfológia
- Szókapcsolatok

Minden adatvezérelt, gyakorlatilag nyelvi szabályoktól mentes, triviális vagy gépi tanulás/mélyneuronháló alapú

A Deep Learning hatása a hibára



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.

„Neoklasszikus” beszédfelismerés

- Deep Learning + klasszikus (letisztított) beszédfelismerés
- Explicit nyelvi szintek
- Minimális, egyszerű kiejtési kivételszabályok
- Fonéma/graféma modellek
- Nagyon gyors, rugalmas (100x-1000x gyorsabb is lehet a valós időnél)
- Sok adat, sok GPU…

- Gépi pontosság <= Emberi pontosság

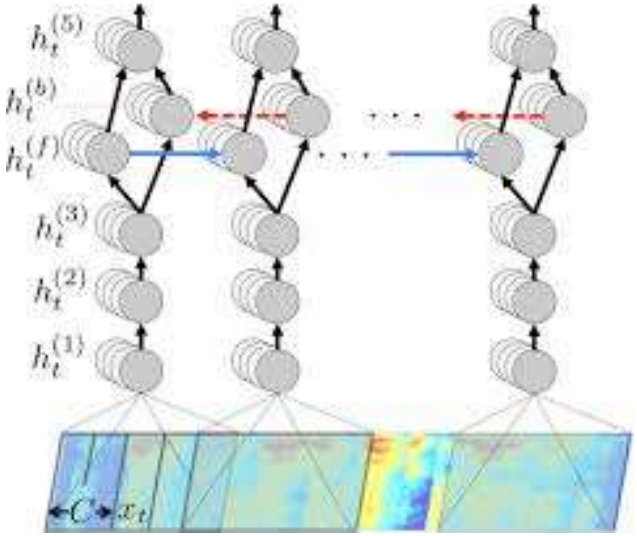
Ha már „deep”, legyen „kövér”

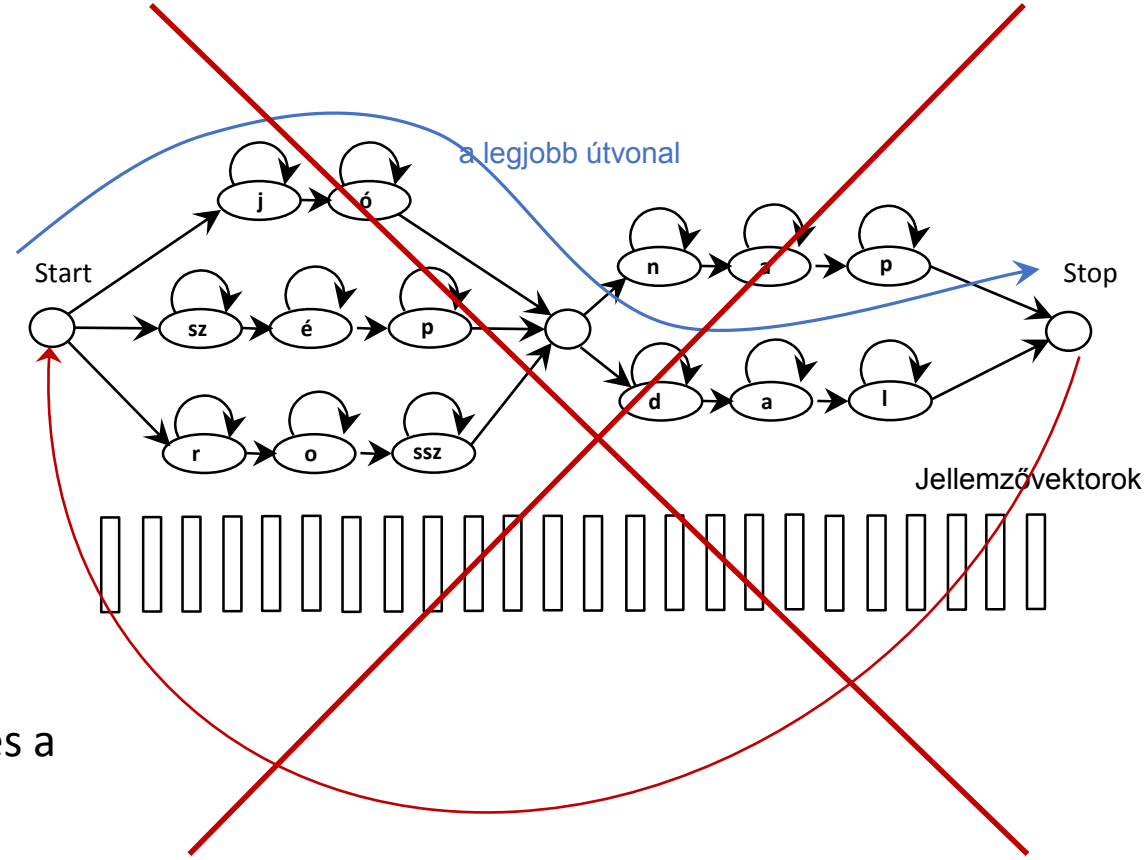
2012 – **End-to-end** speech recognition

mihajlik@tmit.bme.hu

End-to-end speech recognition

- Szakítunk a rejtett Markov-modellekkel





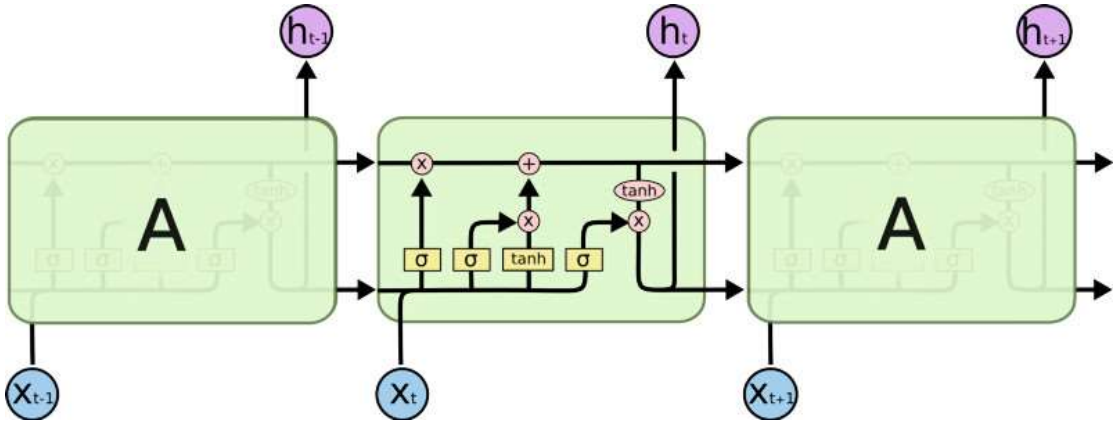
Csak a hanghullámra (spektrogramra) és a kimeneti karakterekre épít.

Tisztán neuronhálós megközelítés…

Alapötlet: sequence 2 sequence rekurrens hálóval

- LSTM

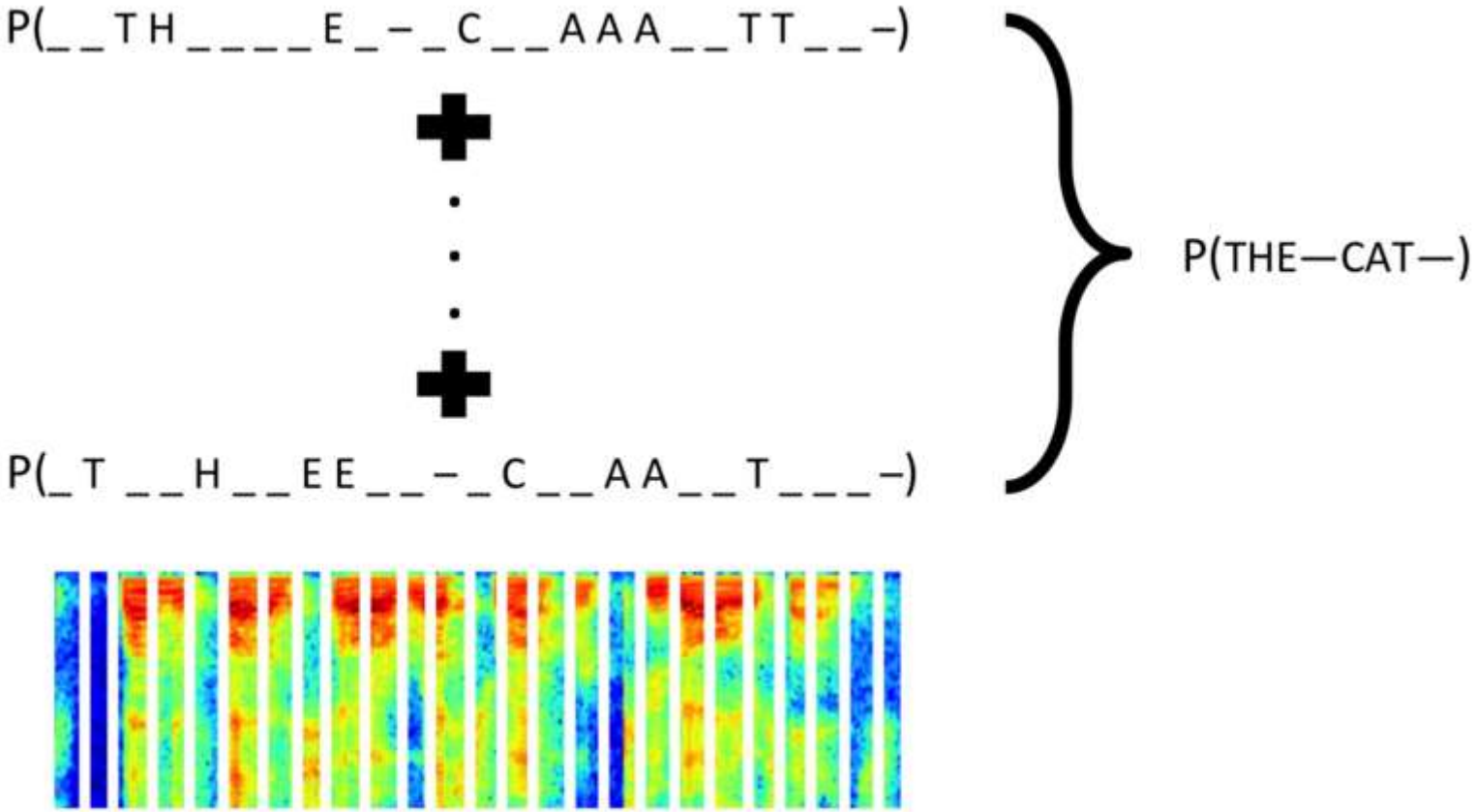
Szöveg (karakterek vagy szavak)



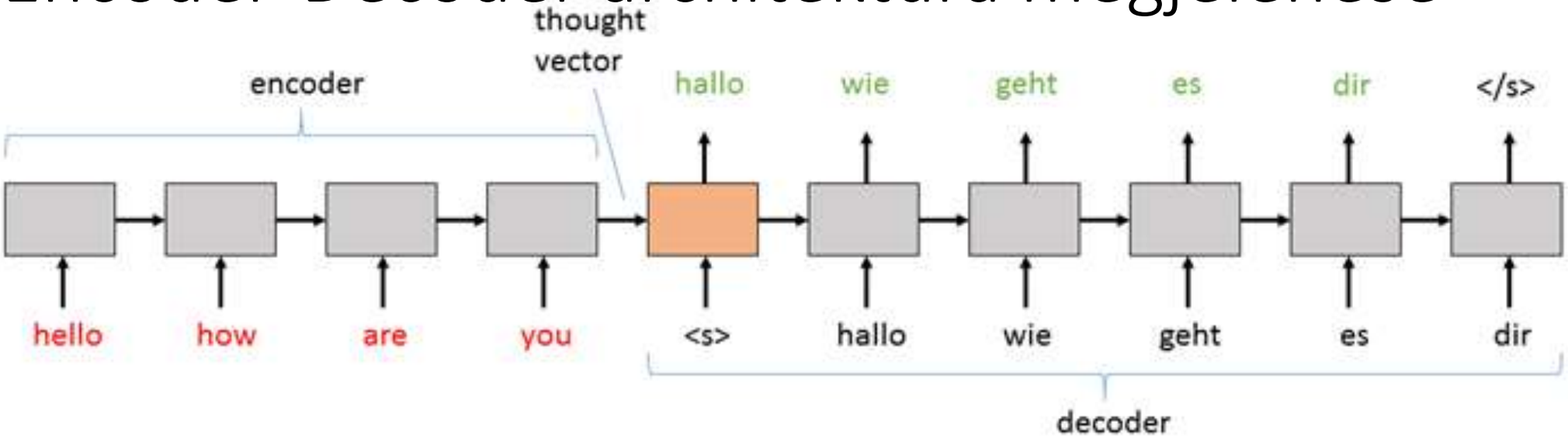
Hang (akusztikus jellemzővektorok)

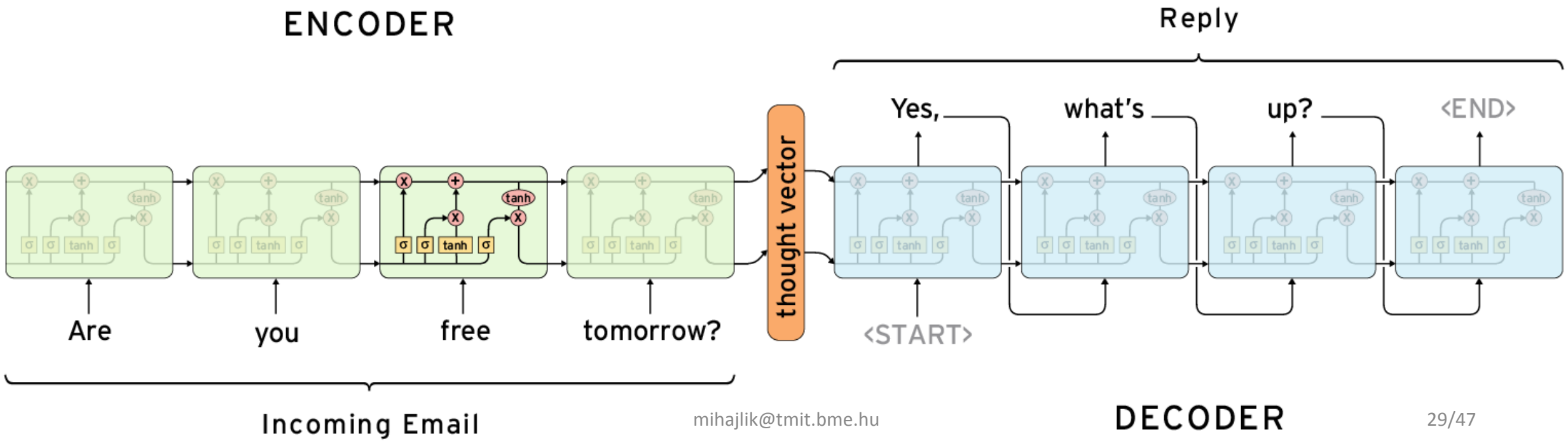
Neuronhálós időillesztés?

„Connectionist Temporal Classification”



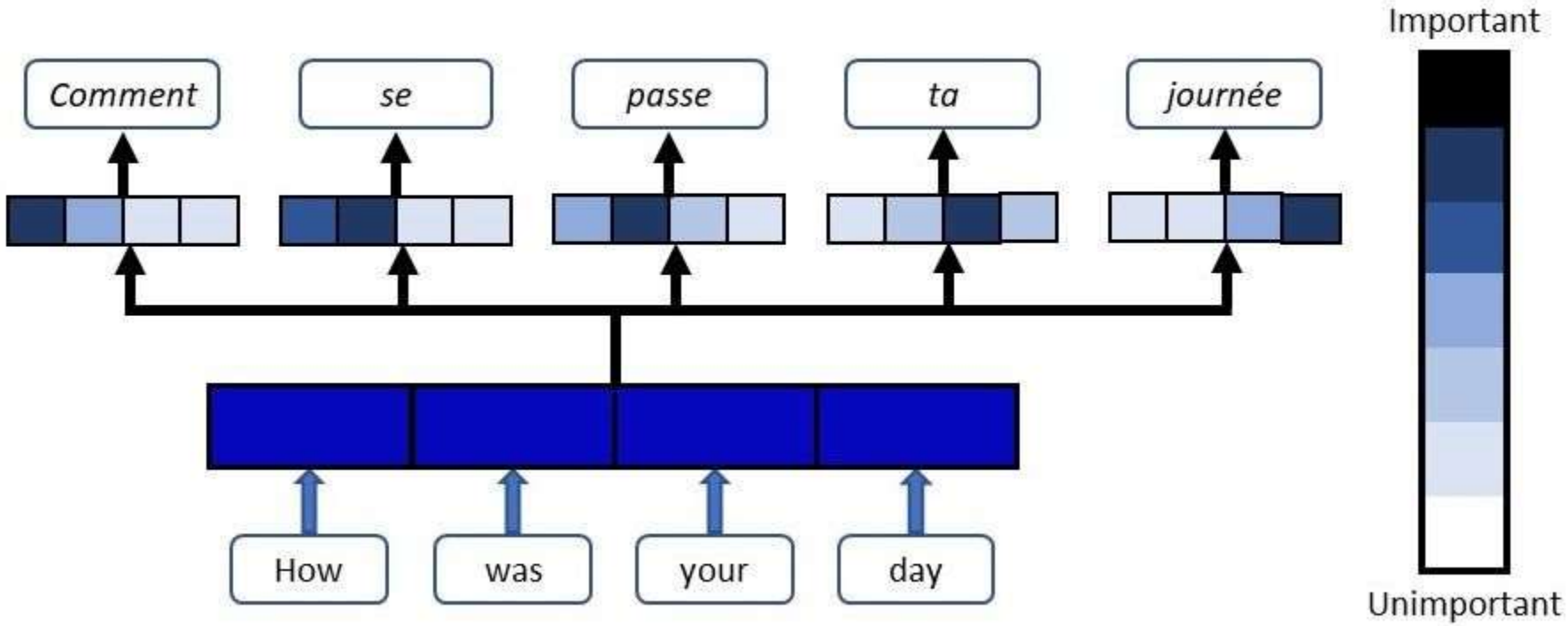
Encoder-Decoder architektúra megjelenése



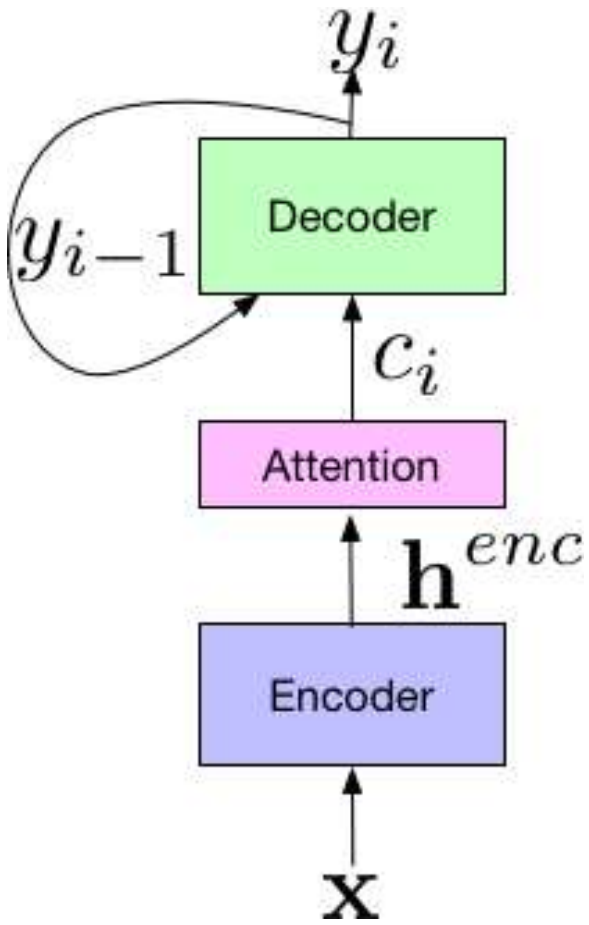


Megtanulunk odafigyelni (S2S)

- „Attention mechanism”



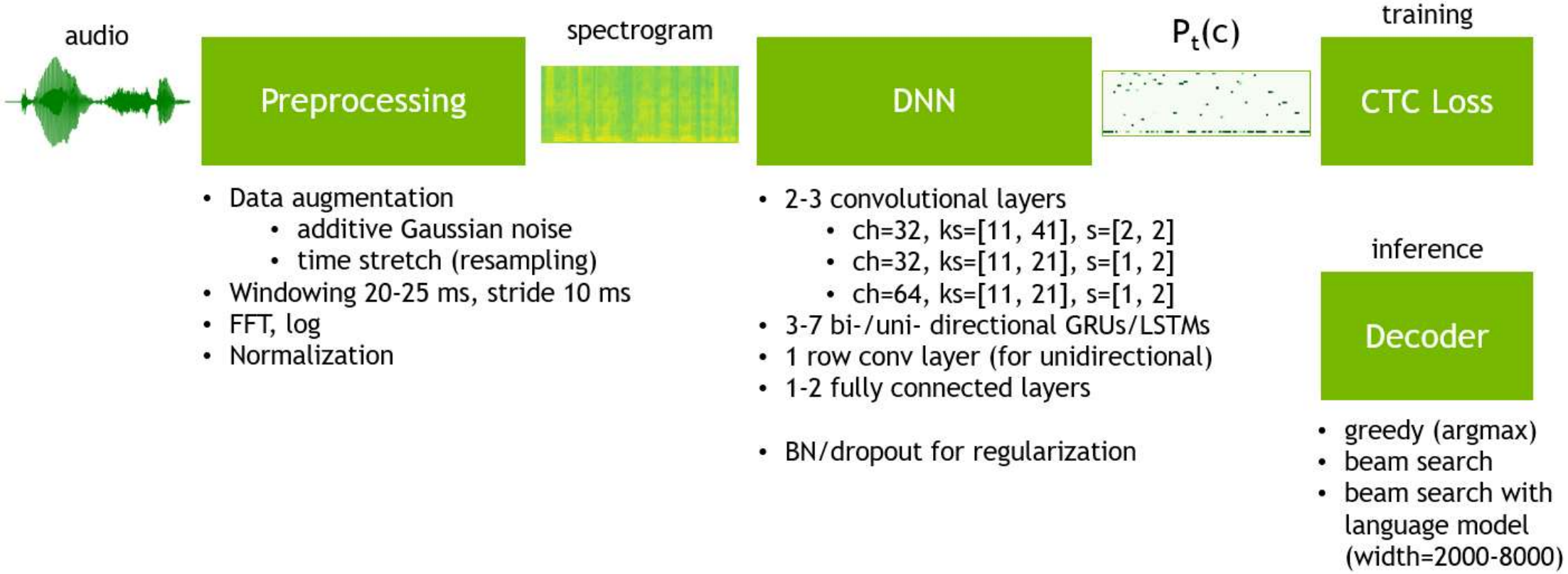
Listen – Attend – Spell (LAS) end-to-end



Google

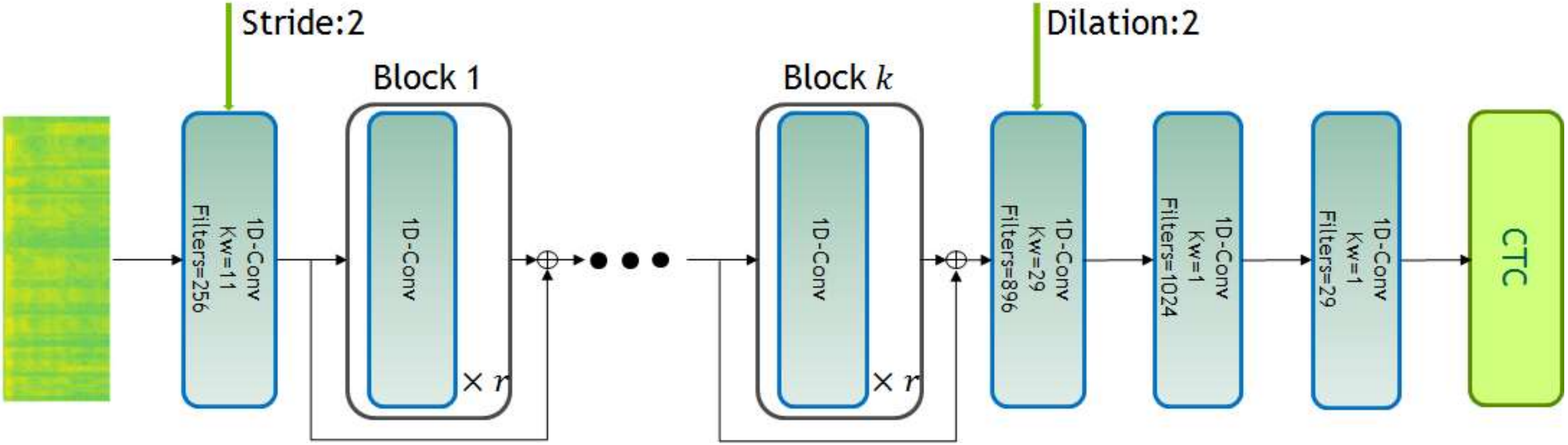
Konvolúciós end-to-end

- Képfelismerésből: a látókéreg neuronelrendezése alapján…
- Baidu – DeepSpeech2



Konvolúciós end-to-end (2)

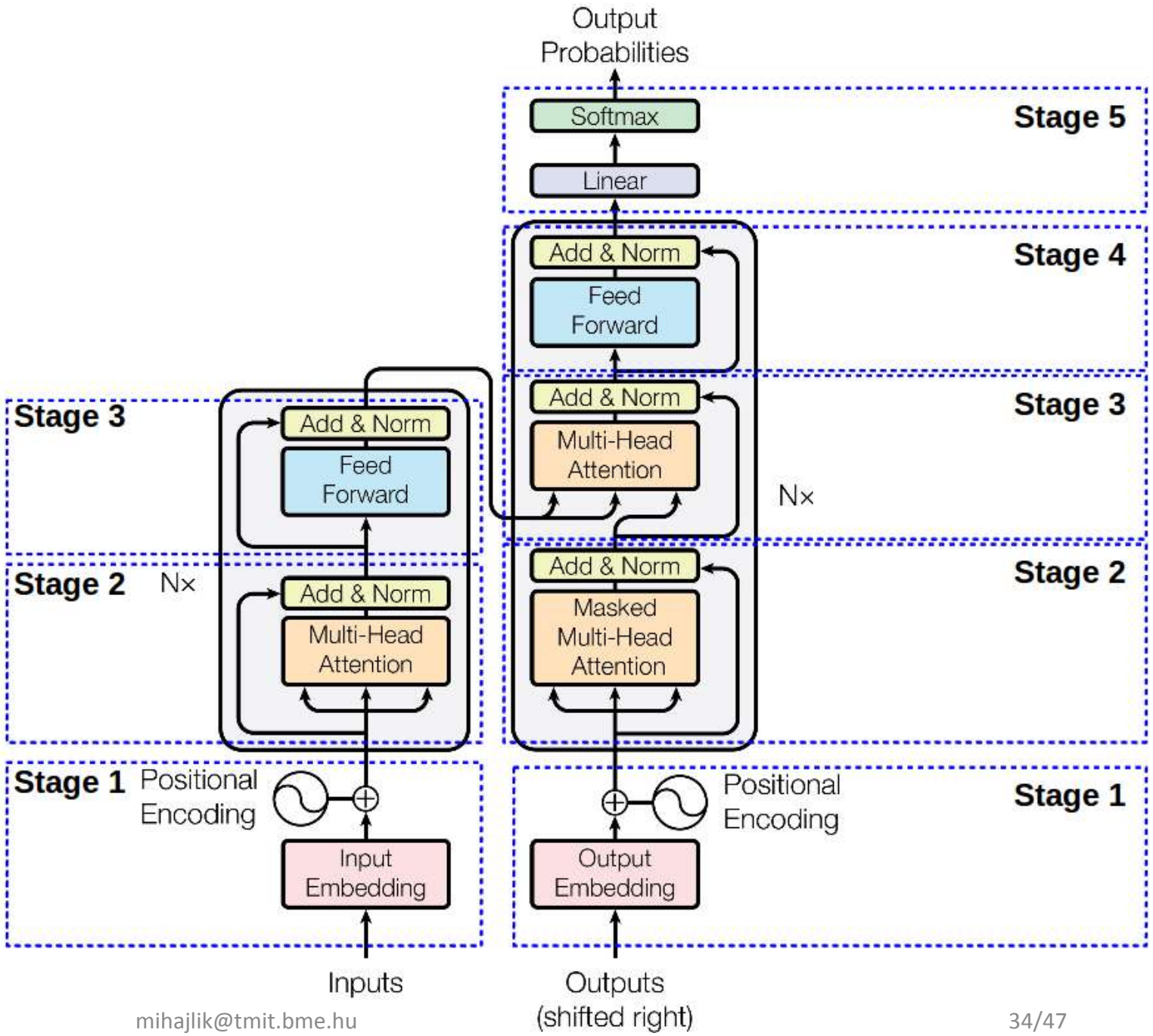
- NVIDIA – Jasper (Just Another Speech Recognizer)



Az új adu-ász:

a Transformer

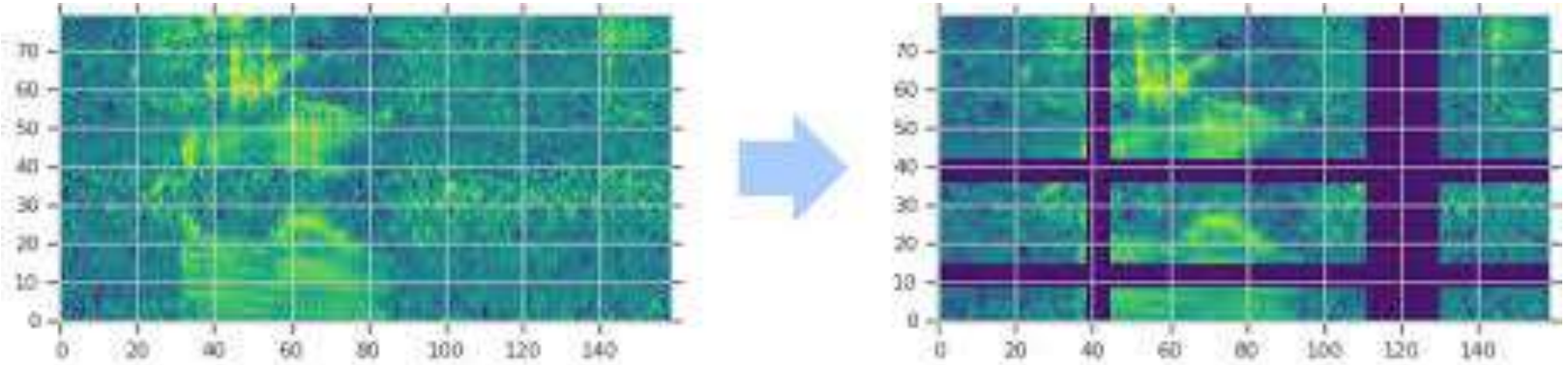
https://talktotransformer.com/



Az adatdúsítás (data augmentation) szerepe

- Sebesség variálás
- Hangerő változtatás
- Zaj hozzáadás
- Viszhangosítás
- …
- Spektrális maszkolás!

SpecAugment



End-to-end beszédfelismerés

- Nagy potenciál, jelenleg a leginkább kutatott terület
- Semmilyen explicit nyelvi tudás, csak mély NN
- Nagyon sok adat, nagyon sok GPU…
- Körülményes, online működés nehézkes
- Ipari bevezetése kezdeti szinten

- Tipikusan *mérsékelt* pontosságjavulás a neoklasszikus modellezéssel szemben

De mennyi az annyi?? (Word Error Rate)

- Amatőr felolvasás (LibriSpeech):

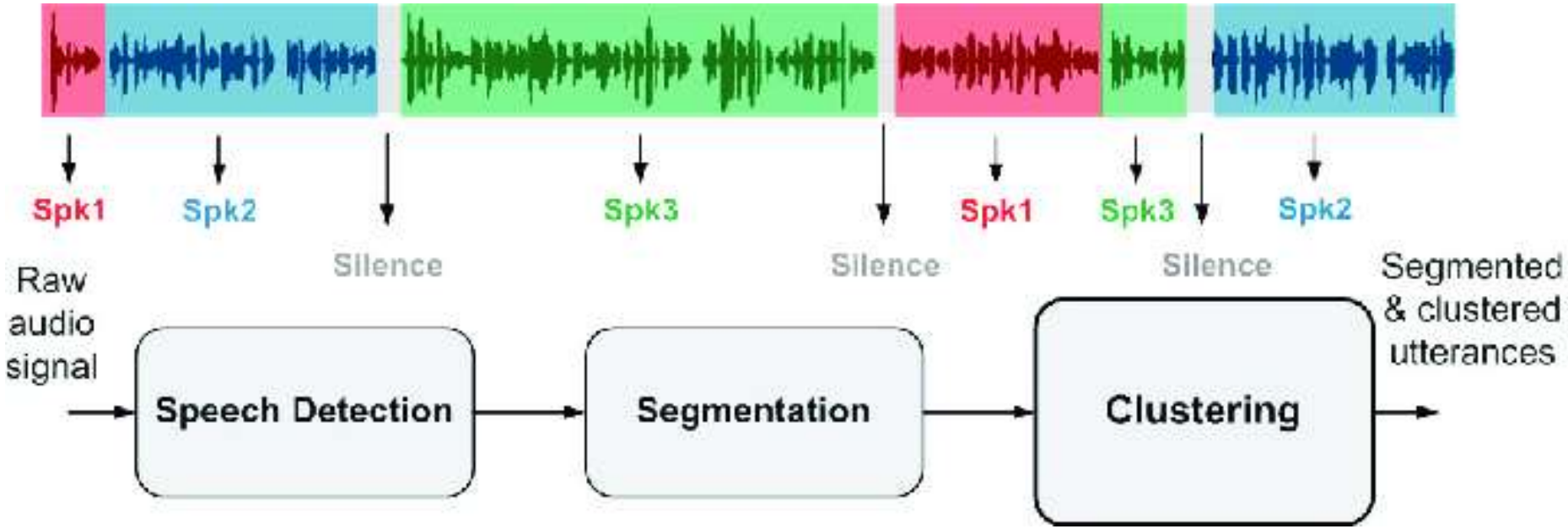
Tiszta	Egyéb	Megközelítés
5.8%	**12.7%**	**Emberi leiratozás**
2.3%	4.9%	Neoclassic + End2end (+ SpecAugment) /Transformer AM/
2.3%	5.0%	Neoclassic + End2end /Transformer LM/
2.5%	5.8%	End2end (+ SpecAugment)
3.7%	8.7%	Neoclassic
2.8%	9.3%	End2End

- Telefonos beszélgetés (SwitchBoard/CallHome):

5.0%	9.1%	Neoclassic
6.8%	14.1%	End2end (+ SpecAugment)

És a beszélődiarizáció?

- (Mikor ki beszél)
- Ugyanaz a deep learning technika, nehezebb feladat



- Legkisebb hiba: ~ 20% !

Konklúziók

- Óriási fejlődés a beszédtartalom-felismerésben az elmúlt 10 évben
- A hajtóerő a mesterséges intelligencia (deep learning)
- Kikristályosodott, „azonnal” bevethető „neoklasszikus technológia”
- Emberi pontosság már nem utópisztikus
- Erős konvergencia a NLP-vel (és minden más Deep Learning témával…)
- Még több adat, még több GPU, párhuzamosítás, jelentős komplexitás növekedés
- További javulás – kvantumszámítógépek?

Köszönöm a figyelmet!

Demó: ki lehet próbálni, hol tartunk.

mihajlik@tmit.bme.hu