SpeechTex

The Speech Technology Expert

A folyamatos beszéd gépi felismerése – a kezdetektől (BME-TTT 90-es évek) napjainkig

DR. MIHAJLIK PÉTER

BME-TTT ´90 „Beszédfelismerők”

Gordos Géza

Tanszékvezető

Tatai Péter

Telecom. Signal Proc.
Labor vezető

Lükő Bálint

Szarvas Máté

Doktoranduszok

Fegyó Tibor

Mihajlik Péter

Az első, folyamatos beszédfelismerési alkalmazások (´90)

- 1993 IBM Personal Dictation System, the first dictation system for the personal computer
- 1995 PHILIPS developed SpeechNote, a dictation and transcription software.
- 1997 Dragon Systems released NaturallySpeaking 1.0 as their first continuous dictation product.

Jellemzőik:

- Angol nyelv
- 20-30 ezres szótárméret
- Szűkített témakör (pl. radiológia, jog)
- Nagy költségvetésű kutatóhelyi háttér, az 50-es 60-as évektől

A folyamatos beszédfelismerés kezdetei

~ 1975: IBM, Fred Jelinek: **HMM**

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

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

`80-as évek vége: **NN**

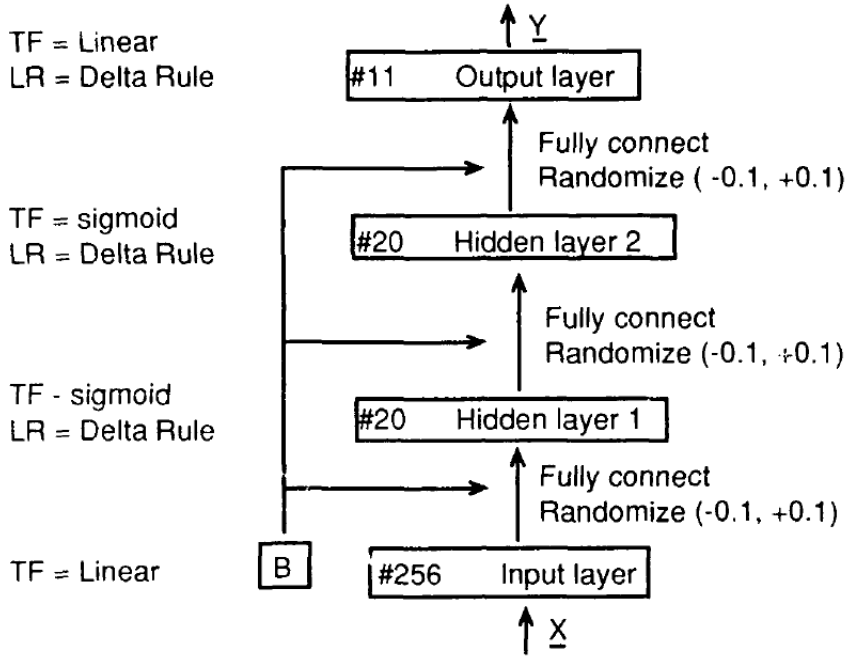


Figure 3. Back-propagation neural network topology

Mi kell(ett) a folyamatos magyar nyelvű beszéd felismeréséhez?

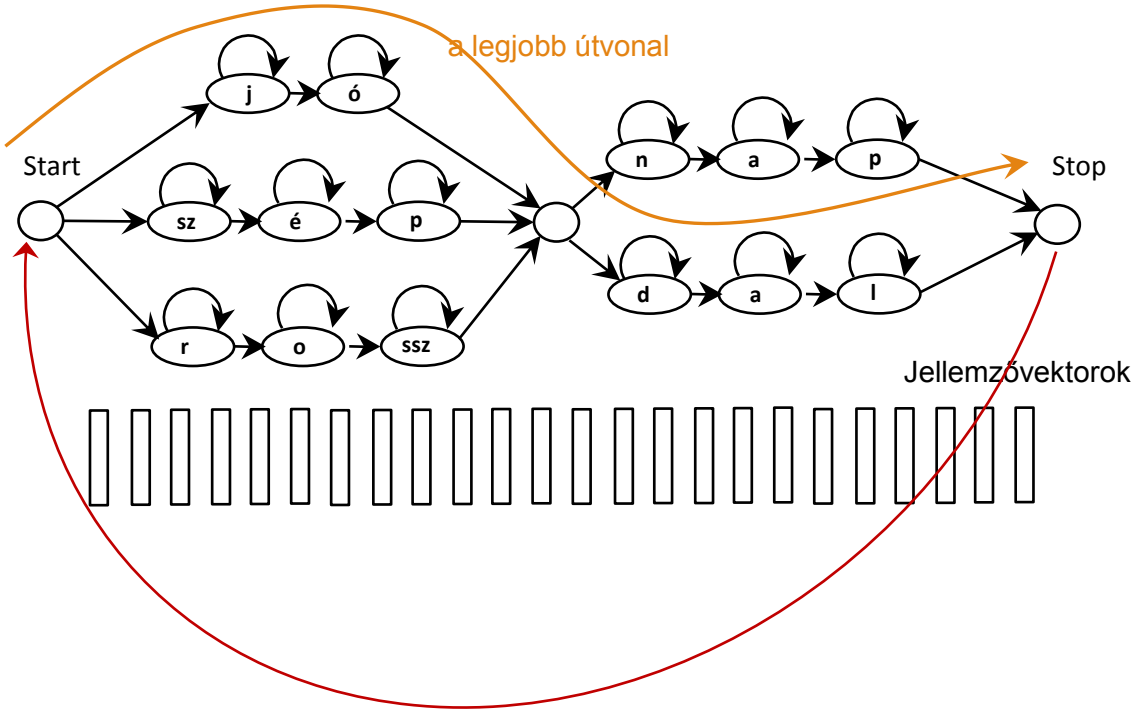
- Technológia…
- A hangjel megfelelő reprezentálása (időtartománybali jel nem alkalmas)
- A beszéd dinamikájának modellezése
- Kiejtési szótárak
- A koartikuláció (beszédhangok egymástól függése) modellezése
- Szóhatárokon fonológiai változások (egybeolvadás, hasonulás) modellezése
- Ragozás miatti hatalmas szóalakszám csökkentése
- Szavak visszaállítása morfémaszerű nyelvi egységekből
- Folyamatos beszédet (és leiratot) tartalmazó adatbázisok
- Sok-sok adat és erőforrás…

Folyamatos beszédfelismerési technológia

Rejtett-Markov modell

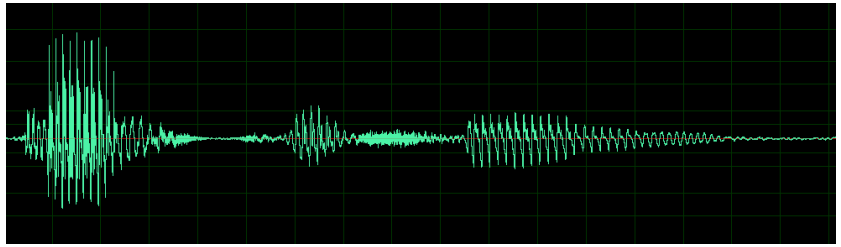
- Implementáció
  - Szarvas Máté: ´96 diplomamunka
  - Cambridge Entropic Research:

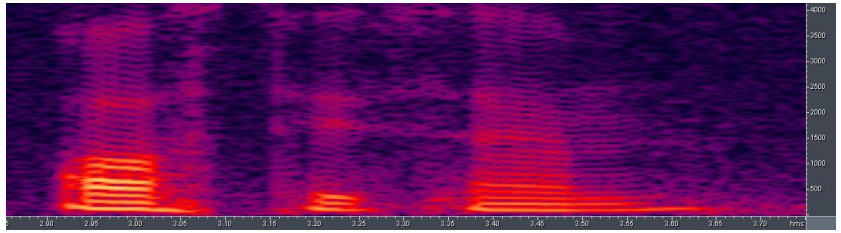
HTK – Hidden Markov-Model ToolKit



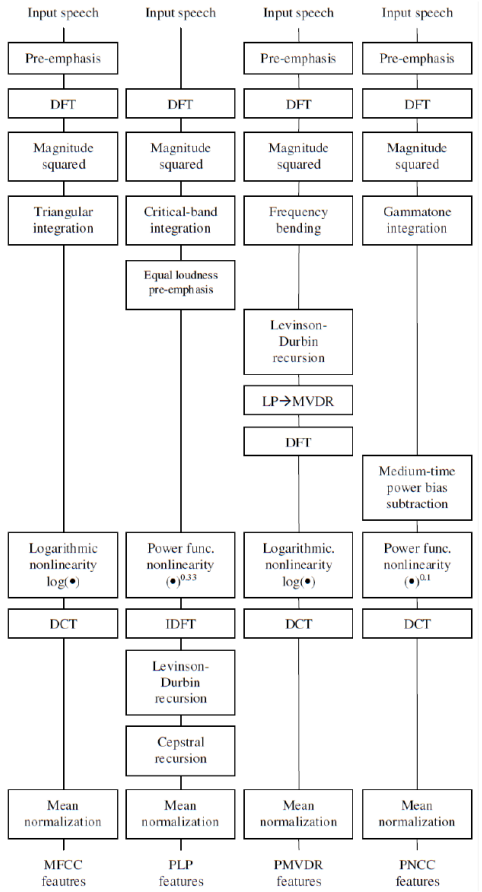
A hangjel megfelelő reprezentálása

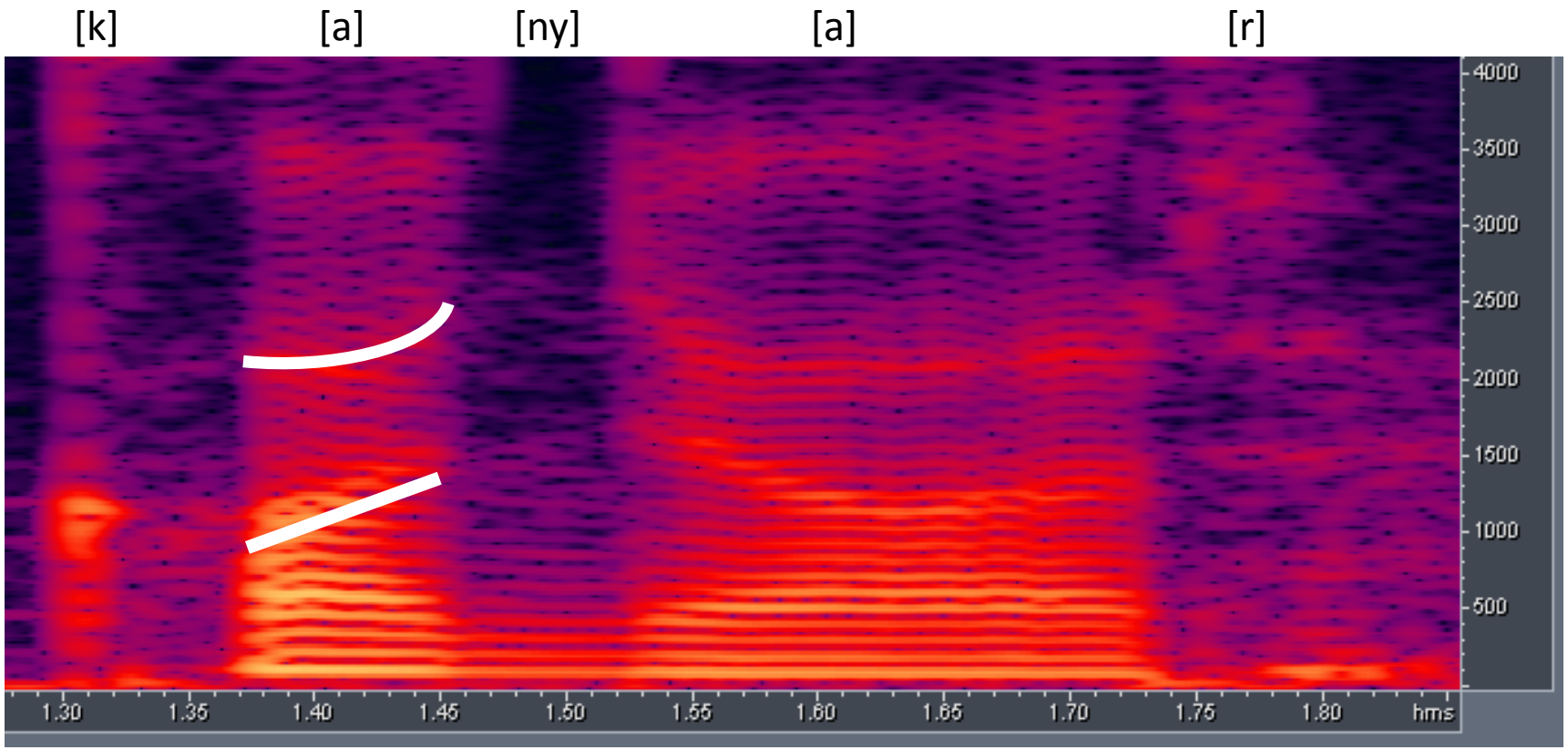
„ a z t h i sz e m ”





MFCC / LPC / PLP …





+ Δ?
+ ΔΔ?
+ ΔΔΔ?

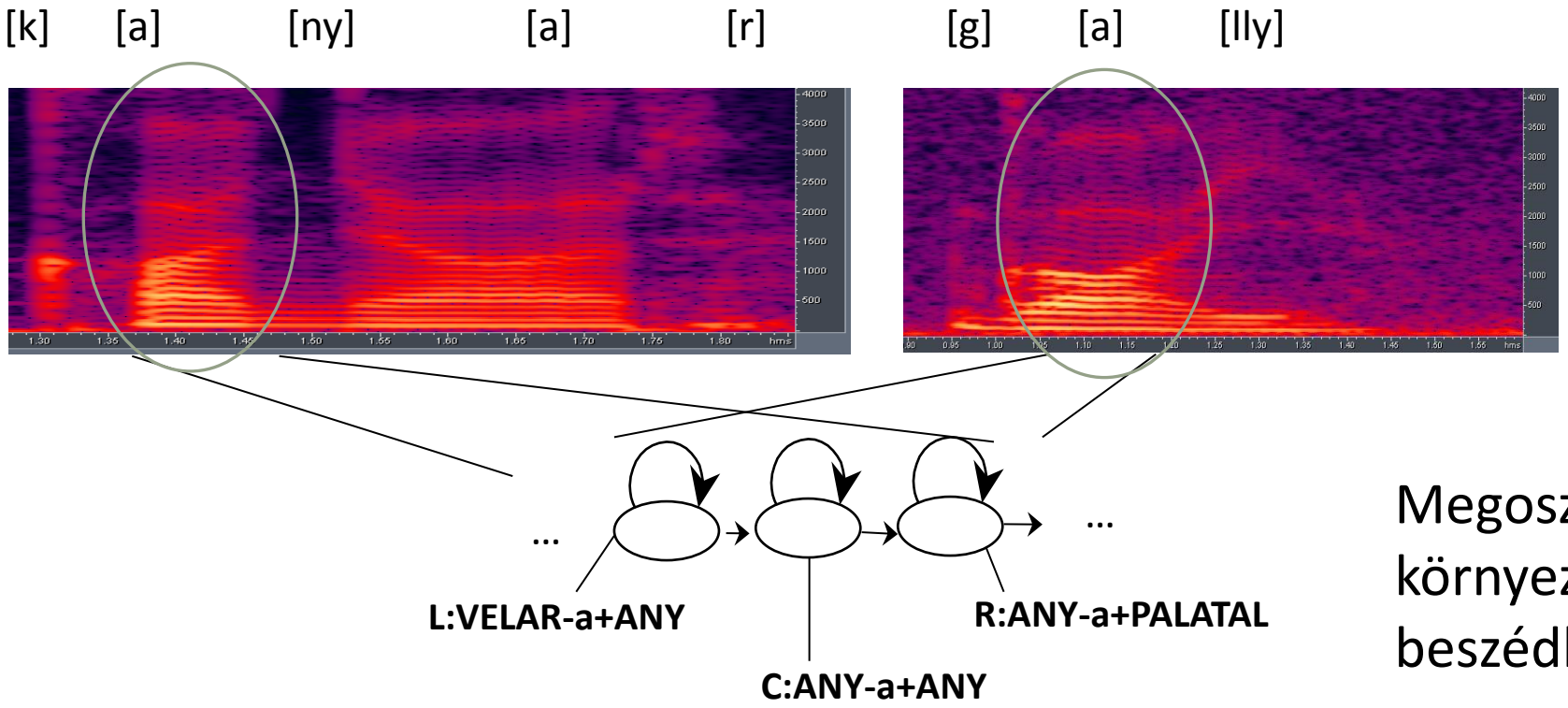
Beszéddinamika-modellezés

hagyja h a ggy a
hangya h a ny gy a

Kiejtési szótár

- 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…

Koartikuláció modellezés



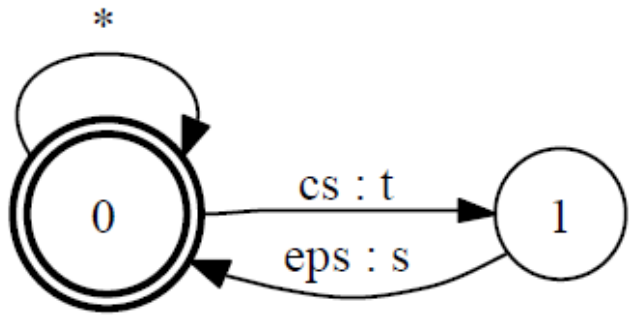
Megosztott állapotú környezetfüggő beszédhangmodellekkel

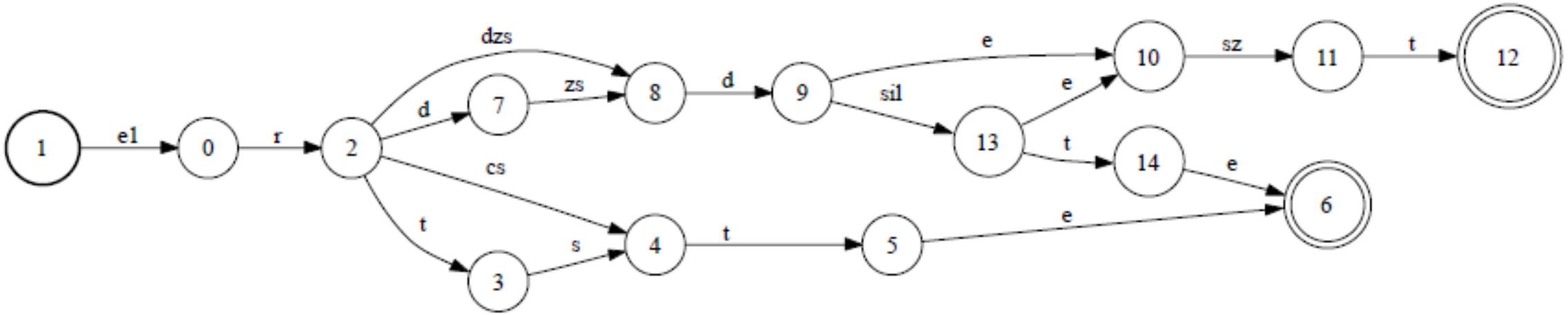
Egybeolvadás, hasonulás stb. modellezése

é r dzs d sil t e

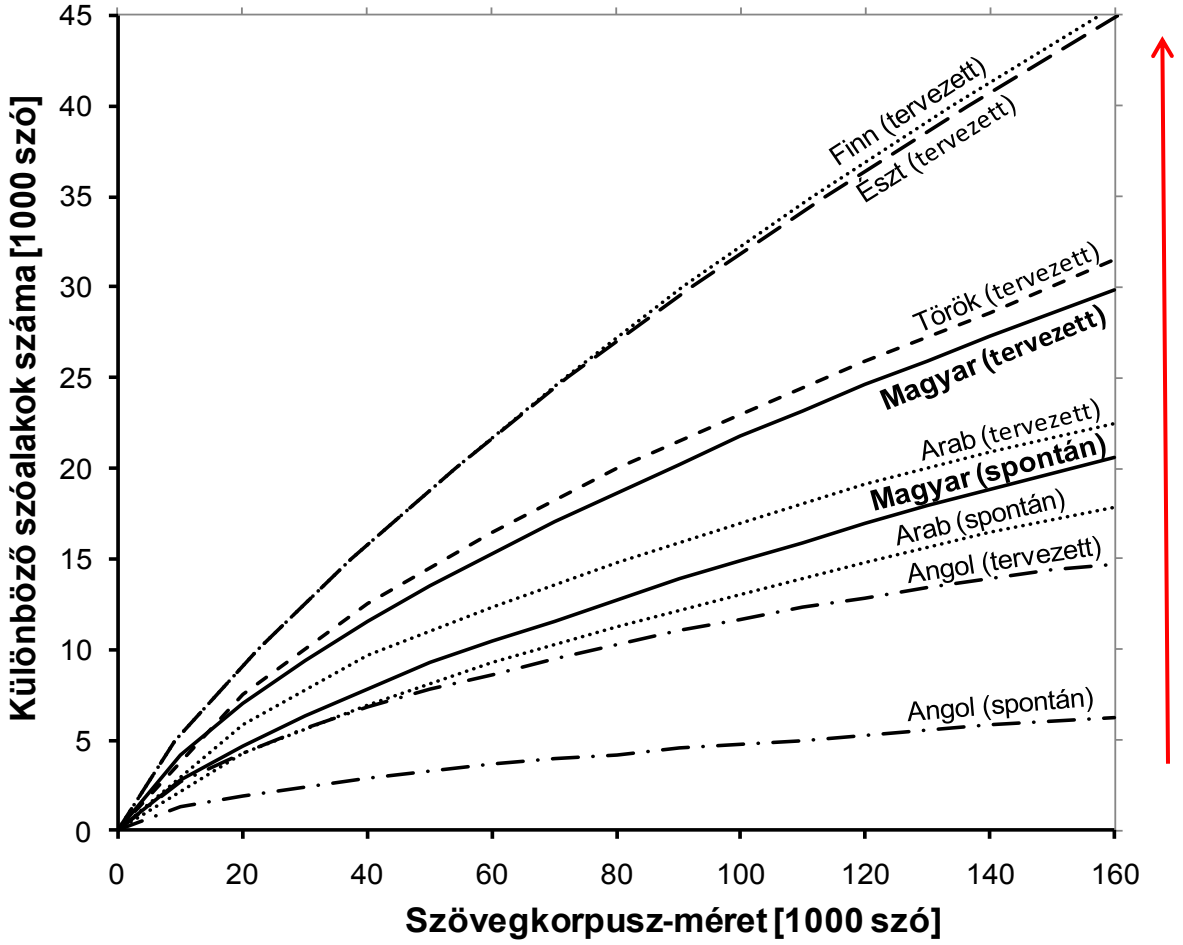
értsd [te|ezt] → *é r cs t e*

é r dzs d [sil] e sz t …





Szóalakszám csökkentése – morfológiai változatosság kezelése



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/

Szóadarabolás: szabály és/vagy statisztikai alapon

Szóalakok visszaállítása

- Szóvég jelekkel

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

- Non-initial jelekkel

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

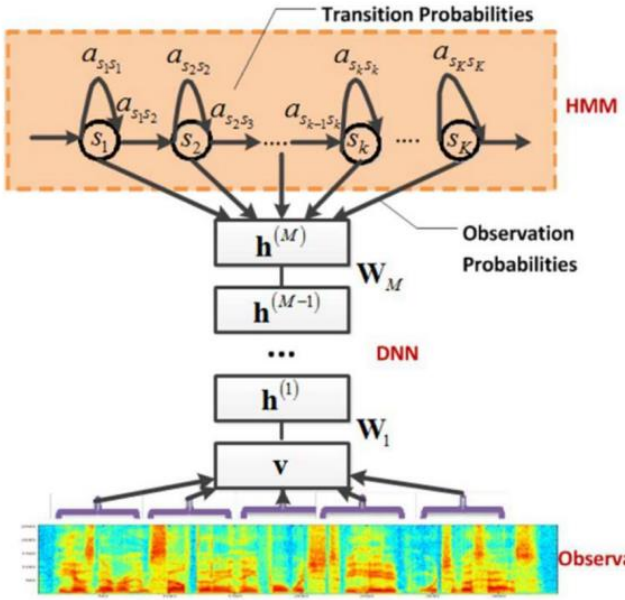
- Szóhatár jelekkel

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

Erőforrások – anno és ma…

- Operatív memória: n Mbyte -----> n Gbyte
- CPU: 80486, 50MHz, 50 MIPS/TMS320 DSP -----> core i7 3+ GHz, RTX 2080: 13.2 TFLOPS
- Beszédadatbázisok (magyarra): 30perc-3 óra ----- > 300-1000 óra
- Szövegadatbázisok (adott témakörből, magyarra): 200 ezer szó -----> 200 millió szó
- Szótárméretek: 200 szó -----> több millió szó
- (Projekt méretek: n x 100MFt -----> n x 10 MFt…)

Deep Learning



Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012): 30-42.

Szóalakok visszaállítása

- Szóvég jelekkel

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

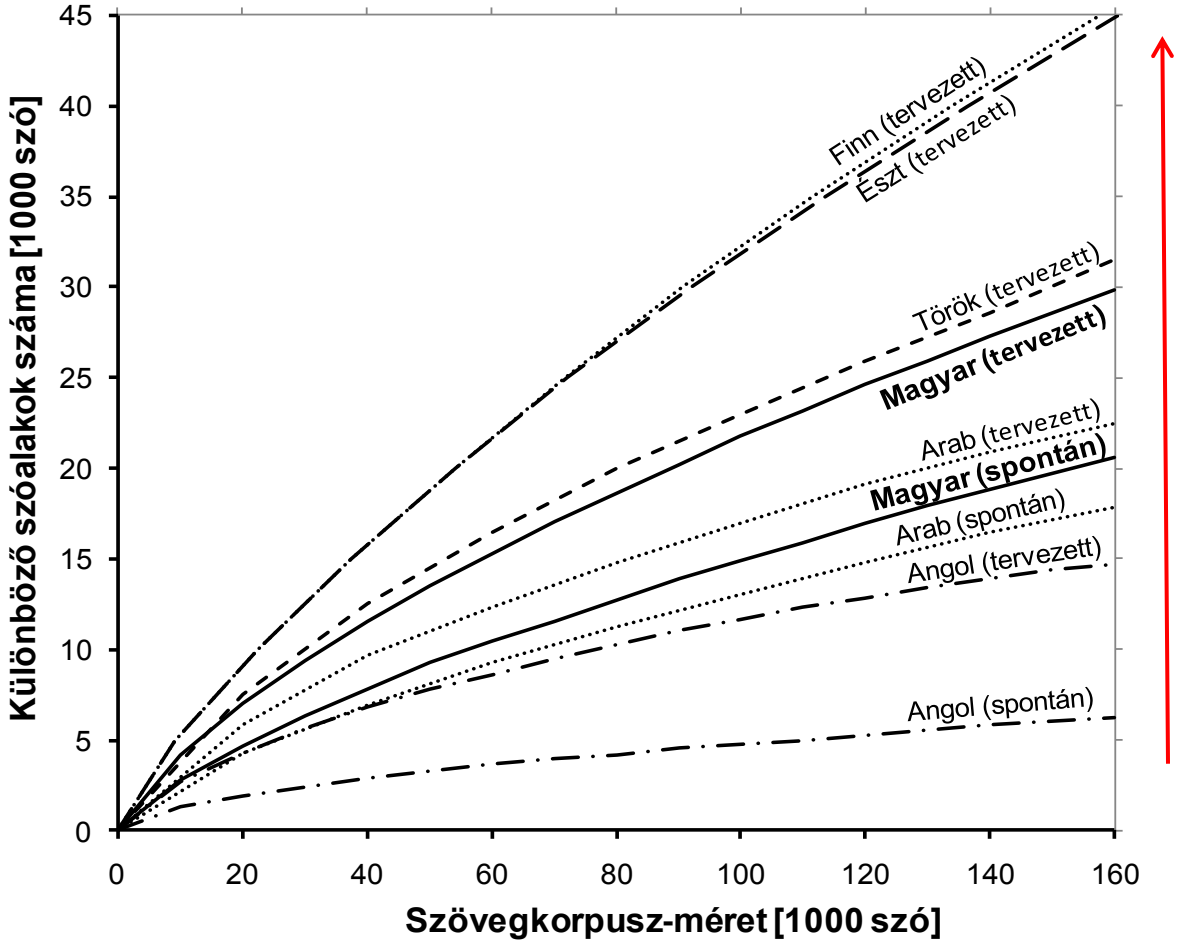
- Non-initial jelekkel

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

- Szóhatár jelekkel

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

Szóalakszám csökkentése – morfológiai változatosság kezelése



láthattuk kóstolhattuk milyen jól tudnak főzni ha akarnak

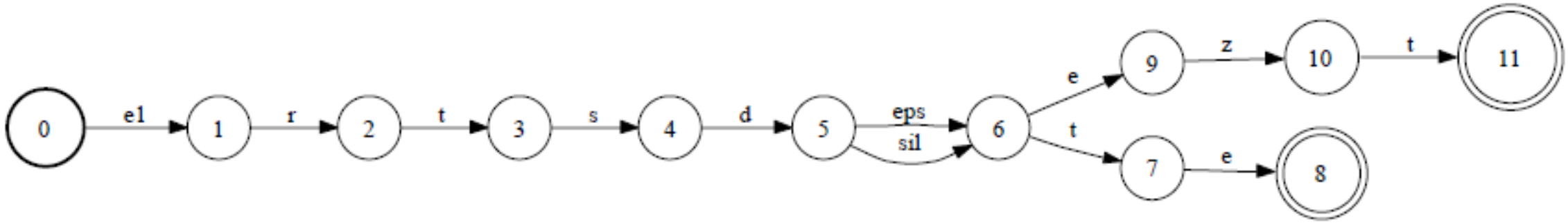
lát hat tuk kóstol hat tuk mily en jól tud nak főz ni ha akar nak

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

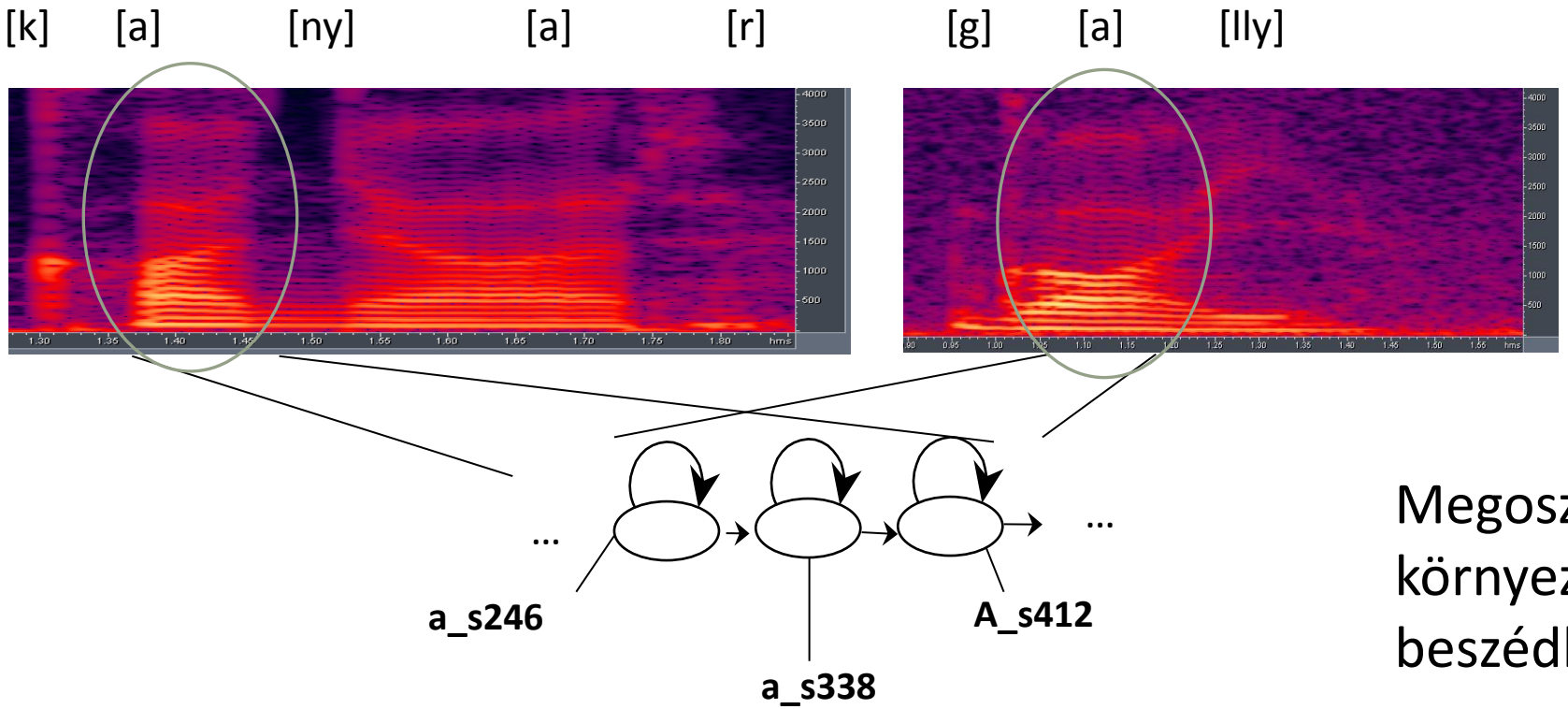
Szóadarabolás: szabály és/vagy statisztikai alapon

~~Egybeolvadás, hasonulás stb. modellezése~~

értsd [te|ezt] → *értsd [te|ezt]*



Koartikuláció modellezés ?

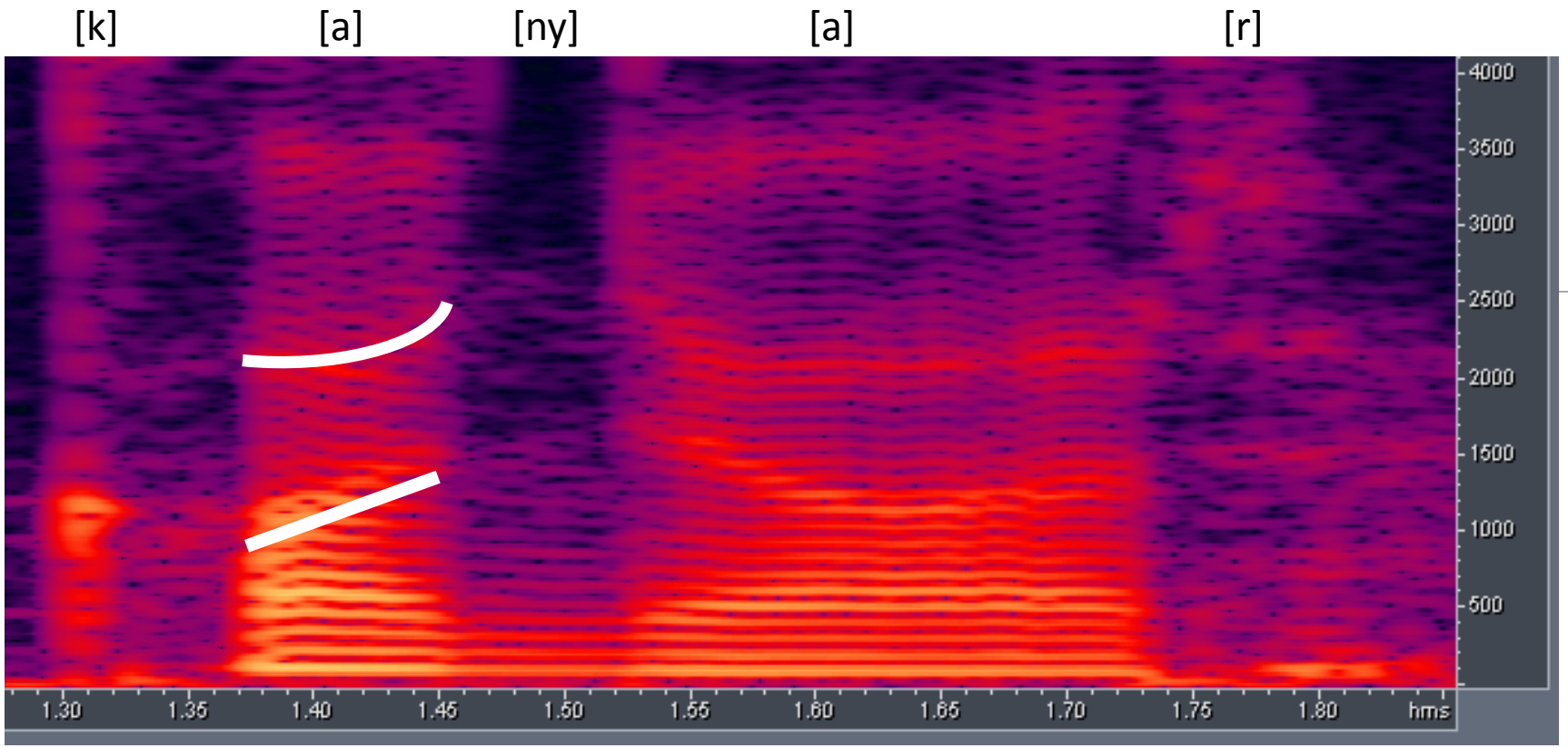


Megosztott állapotú környezetfüggő beszédhangmodellekkel

Kiejtési szótár

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…

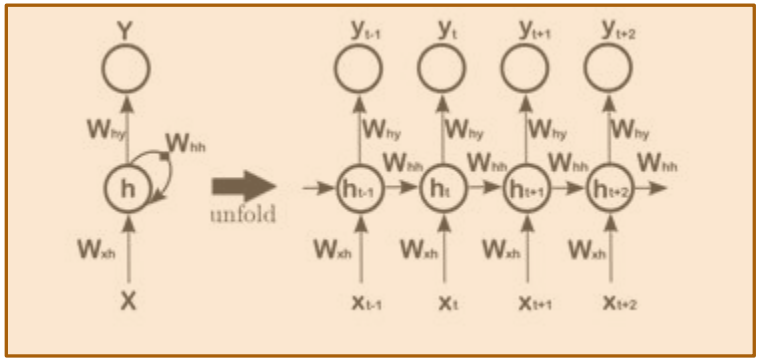


+ Δ?

+ ΔΔ?

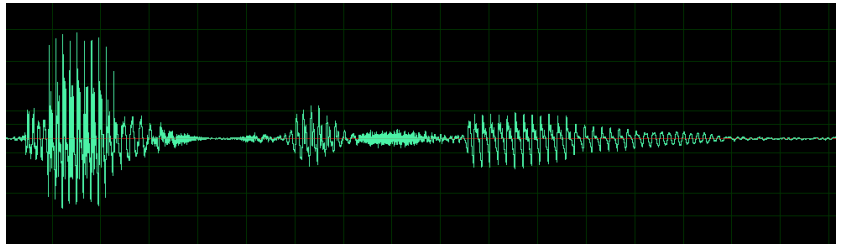
+ ΔΔΔ?

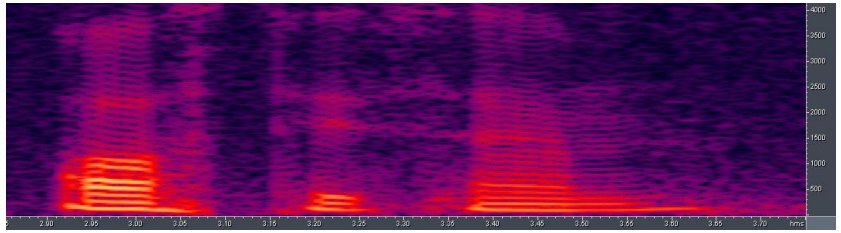
Beszéddinamika



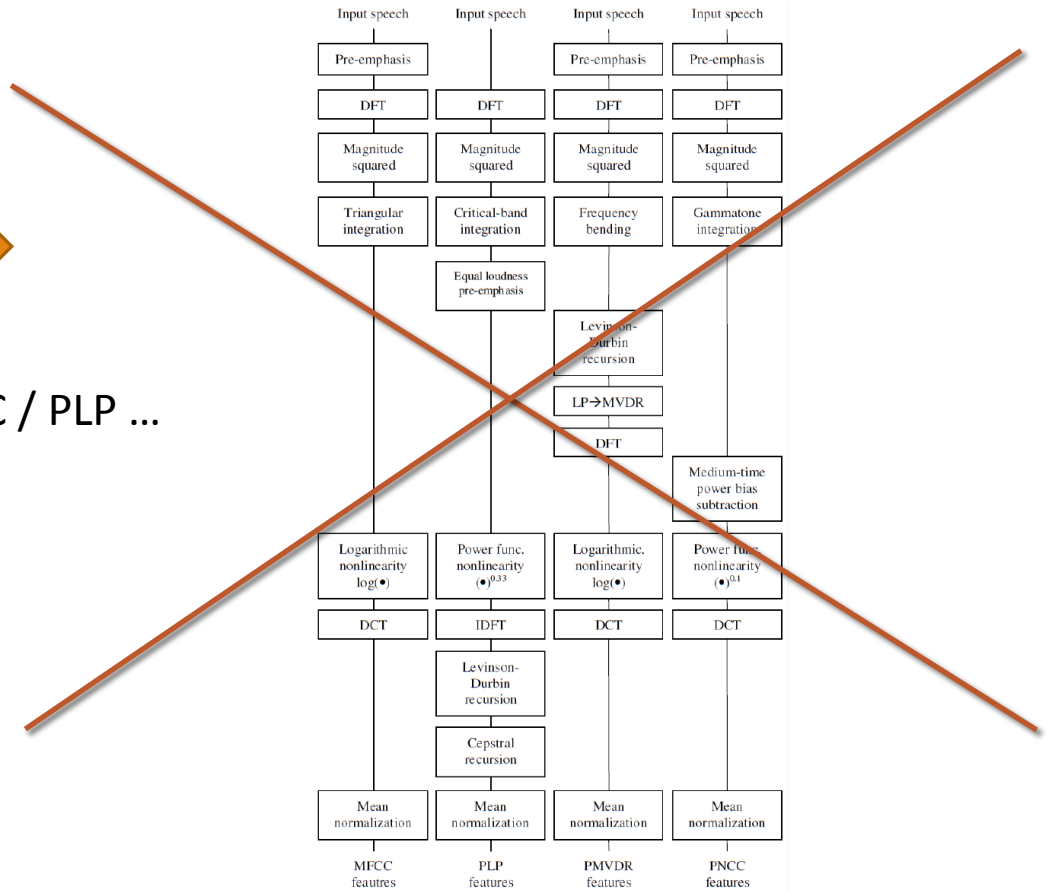
A hangjel megfelelő reprezentálása

„ a z t h i sz e m ”





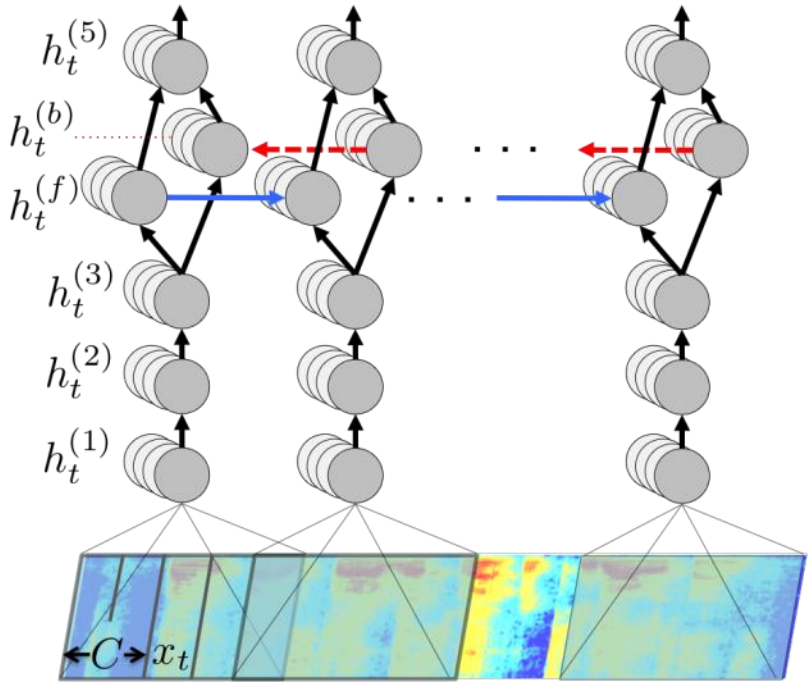
MFCC / LPC / PLP …



A folyamatos beszédfelismerés jövője

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

End-to-end (tisztán neuronhálós) megközelítés…



`80-as évek vége: **NN**

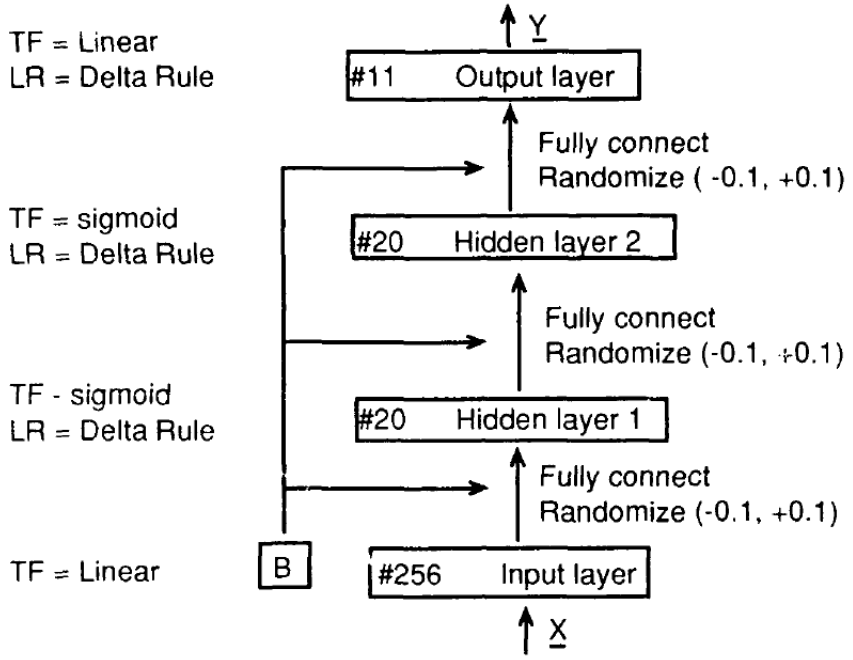
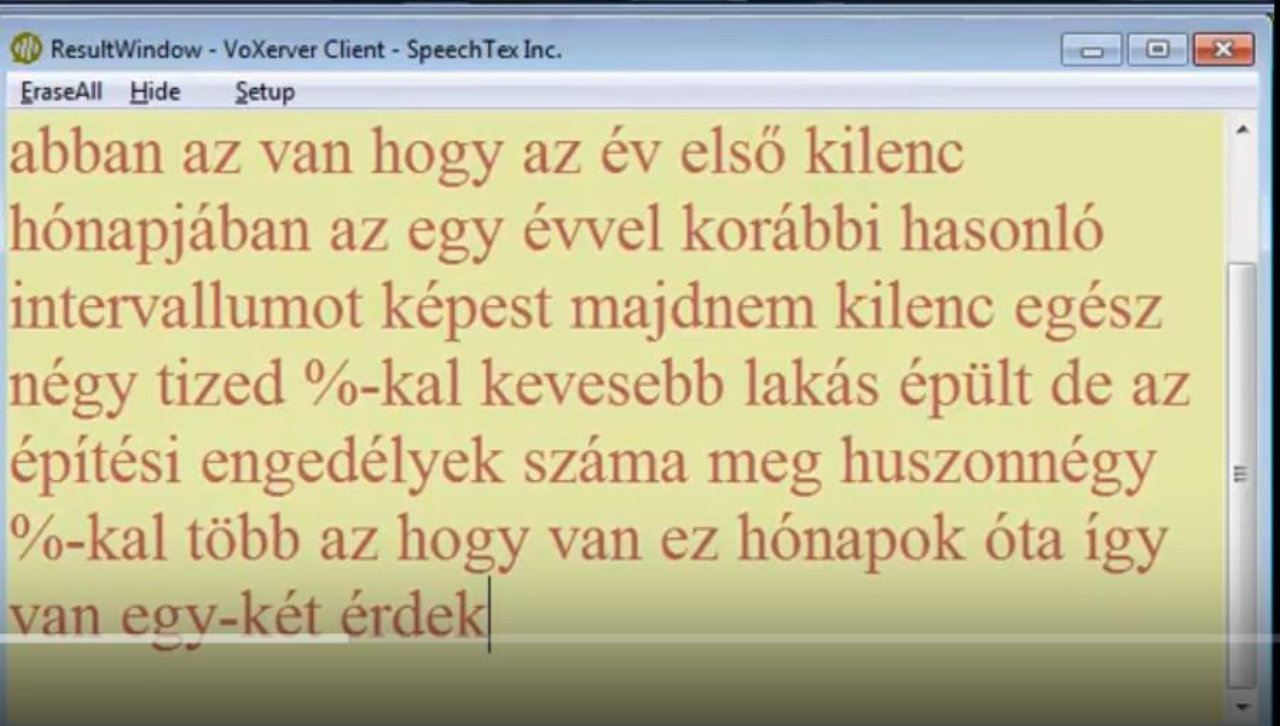


Figure 3. Back-propagation neural network topology



Jobb-e már a gép mint az ember?

https://www.youtube.com/watch?time_continue=1&v=p_oKK4xzZg8

Köszönöm a figyelmet!