Deep learning alapú beszédszintézis

Dr. Zainkó Csaba

Deep learning - Beszédszintézis 01

Alapok

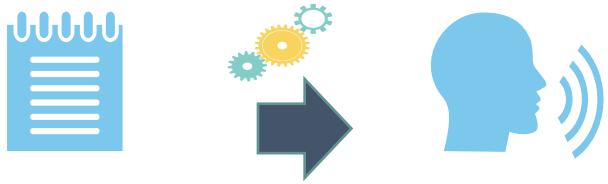
Mit várjunk, várhatunk el? 02

Jelen állapot

Merre tovább? 03

Jövő

Beszédszintézis



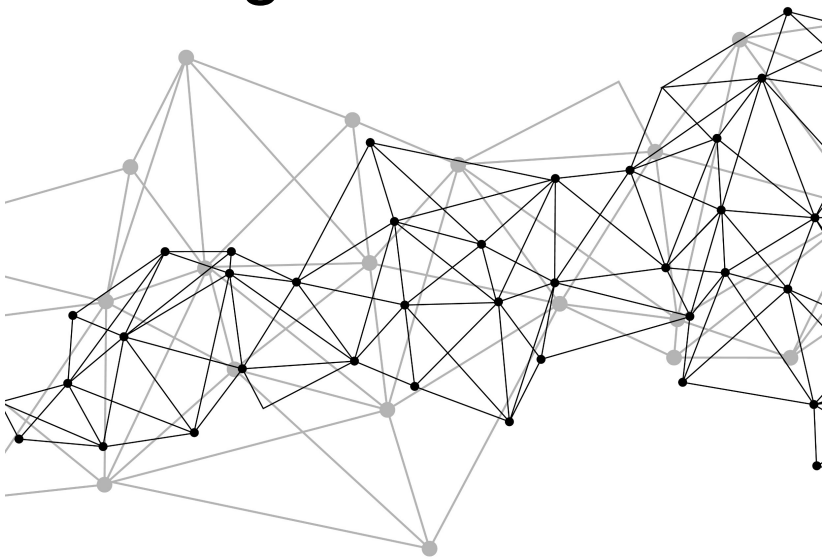
- Szöveg Beszéd

„Nem is gondolnánk milyen gyakran használunk már ma is neurális hálózatokat vagy azok eredményeit a mindennapokban.”

Technológia

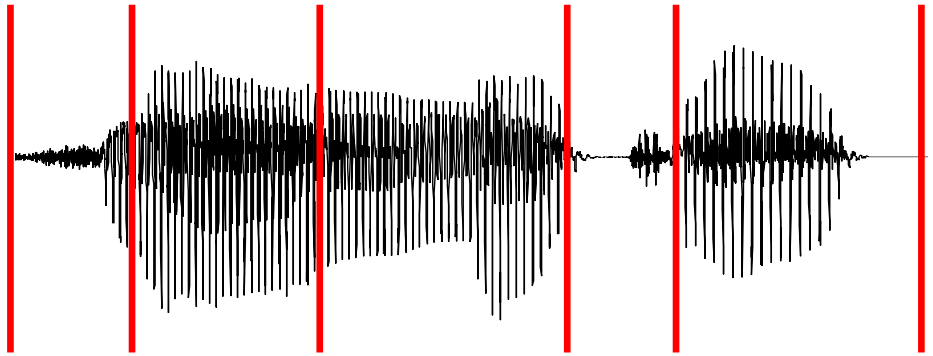
- Parametrikus

- Deep learning



http://www.freepik.com

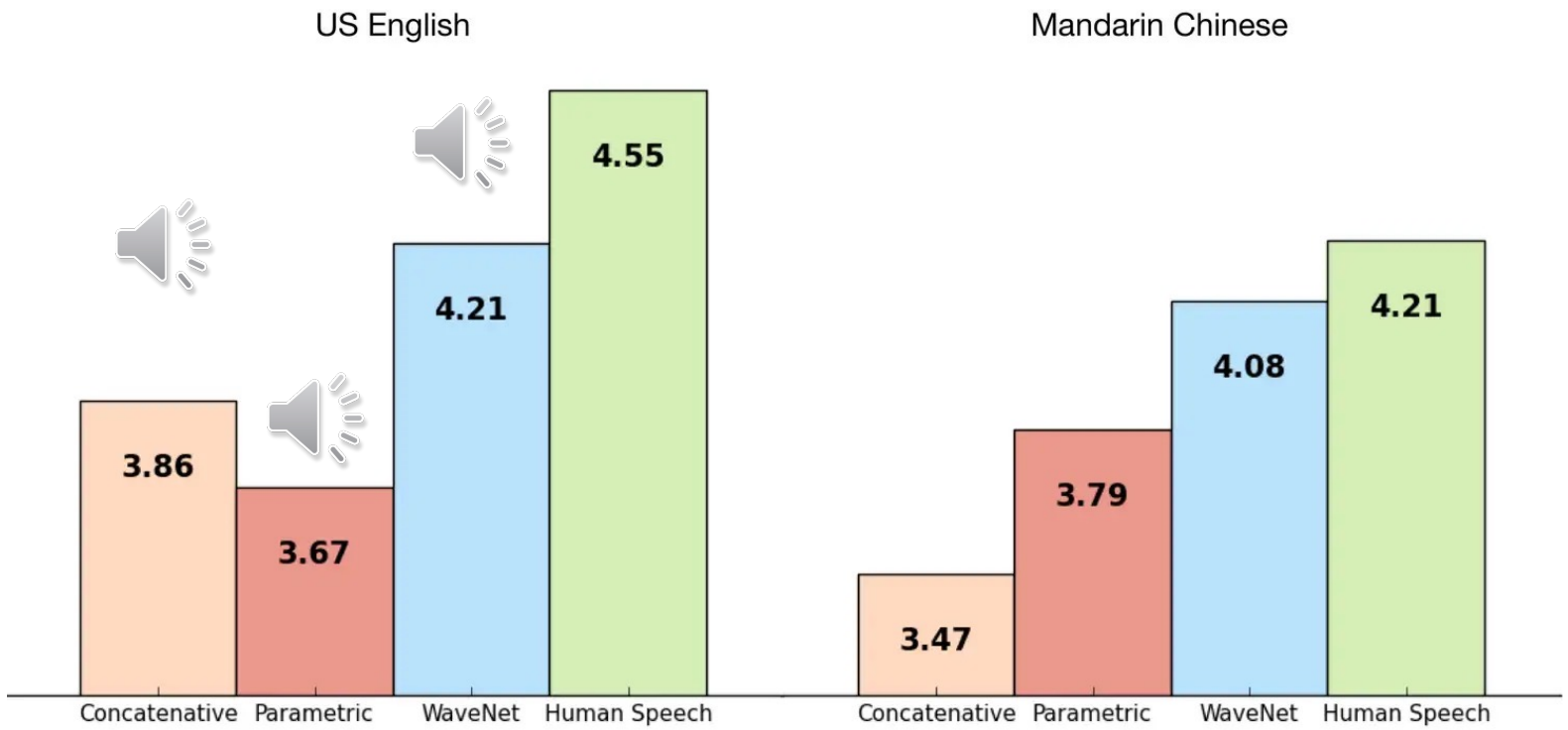
- Összefűzéses



Deep learning

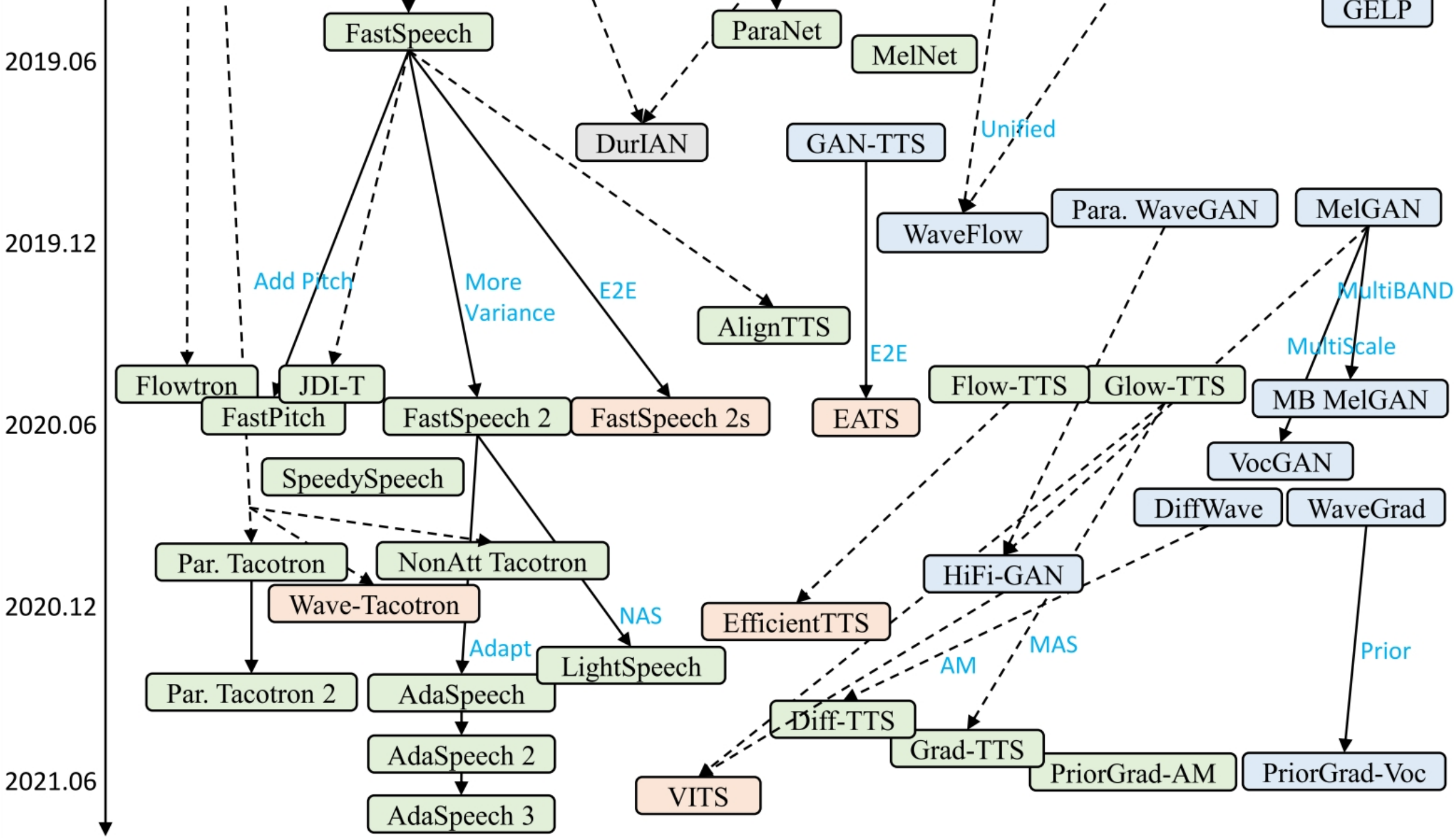
- 2016
- Google Deepmind
- WaveNet

https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio



Minőség – Hogyan mérjük?

- Szubjektív
  - MOS – Mean Opinion Score
    - 1 – legrosszabb
    - 5 – legjobb
  - MUSHRA - Multiple Stimuli with Hidden Reference and Anchor
    - Referencia minőség
      - Alsó horgony
      - Felső horgony
- Objektív
  - Referencia
  - Felismerővel (WER)
  - Gépi tanulásos mód (VoiceMOS)


2019.06
2019.12
2020.06
2020.12
2021.06
FastSpeech
ParaNet
MelNet
GELP
DurIAN
GAN-TTS
Unified
WaveFlow
Para. WaveGAN
MelGAN
Add Pitch
More Variance
E2E
AlignTTS
MultiBAND
MultiScale
Flowtron
JDI-T
FastPitch
FastSpeech 2
FastSpeech 2s
EATS
Flow-TTS
Glow-TTS
MB MelGAN
VocGAN
SpeedySpeech
DiffWave
WaveGrad
Par. Tacotron
NonAtt Tacotron
HiFi-GAN
Wave-Tacotron
EfficientTTS
NAS
Adapt
LightSpeech
AM
MAS
Prior
Par. Tacotron 2
AdaSpeech
Diff-TTS
AdaSpeech 2
Grad-TTS
PriorGrad-AM
PriorGrad-Voc
VITS
AdaSpeech 3

DNN-TTS

- End-to-End: szöveg (hangsorozat) → Hullámforma
- 2 lépcsős:
  szöveg (hangsorozat) → spektrális (mel-spectrum)→ Hullámforma

- Tanítás: GPU sok óra beszéd + szöveg
  - Tanítás idő: néhány óra, 1-2 nap (1-2 GPU-n)
- Generálás: CPU/GPU

Mit tud, mit kell tudni?

- Közel emberi hangminőség
- Kezelni:
  - Számok
  - Dátum
  - Összegek
  - Időpontok
  - Rövidítések
  - …
- Gyors: legalább valós idő
- Kérdések (eldöntendő)
- Kontrollálható kimenet
  - kivétel szótár

Mit tud(hatna) a mostani technológia?

- Testreszabás
  - Férfi/női
  - Beszélő választás
  - Egyedi hang
  - Stílus
  - Érzelem???

- Szövegjavítás

Egyedi hang

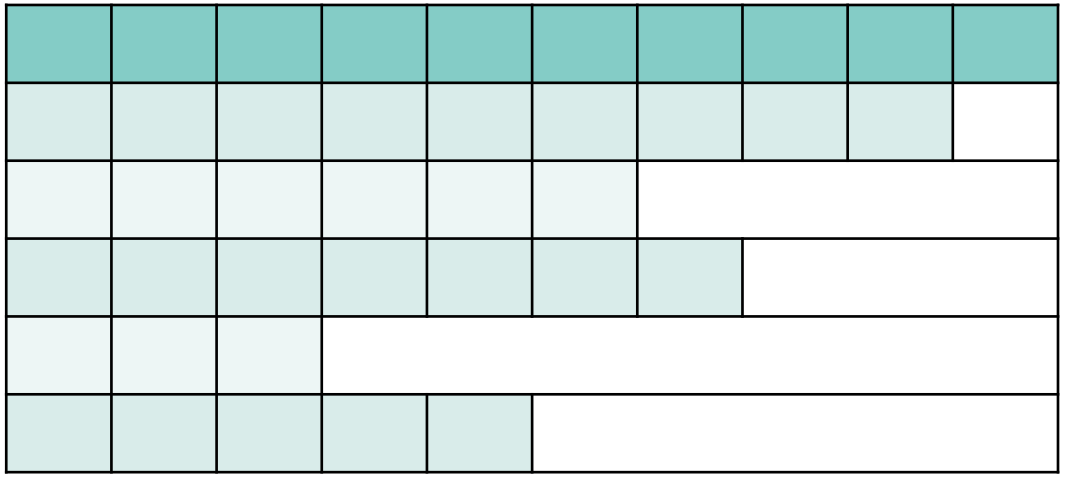
- Új tanítás: 1-2 óra beszéd
- Adaptálás: 5-10 perc beszéd

- És ha csak 1 mondat van?

Ez a szoveg mindenki szamara jol ertheto.

Mit tud(hatna) a mostani technológia?

- Válaszidő
  - CPU/GPU
  - max 100-200ms késleltetés
- Batch feldolgozás (GPU)
  - Több azonos jellegű feladat
- Párhuzamos generálás



Chat - Párbeszéd

- Sima szövegfelolvasás helyett: Párbeszéd
- Alkalmazkodni a másik félhez, integrált TTS
  - Stílusban
  - Beszédtempóban
  - Beszélő (ffi/nő)

Különálló független komponensként egyre kevésbé értelmezhető

Merre tovább?

- 5 éves gyerek szintje
- Szótár, kiejtés

PÉLDA:

„Ezekre a kérdésekre keressük a választ az idei dataSTREAM konferencián."

Tudjuk-e automatizálni az ember hozzáadott tudását?

LLM – Large Language Model

- 100 óra beszéd (kb. 4,3 milló karakter, <1 milló szó)
- GPT-3 – 10 milliárd szó!

- Nagy modellek adhatnak megoldást
  - Wave2vec
  - Bert alapú
  - GPT típusú modellek

Mit várunk?

- Jobb szubjektív élményt
- Pontosabb nyelvfelismerést
- Szövegkorrekciót (hogyan teszteljem?)
  - Ügyfélélmény v. Szöveghű viselkedés

- Előzmények figyelembevétele
- Szöveg megértése ???

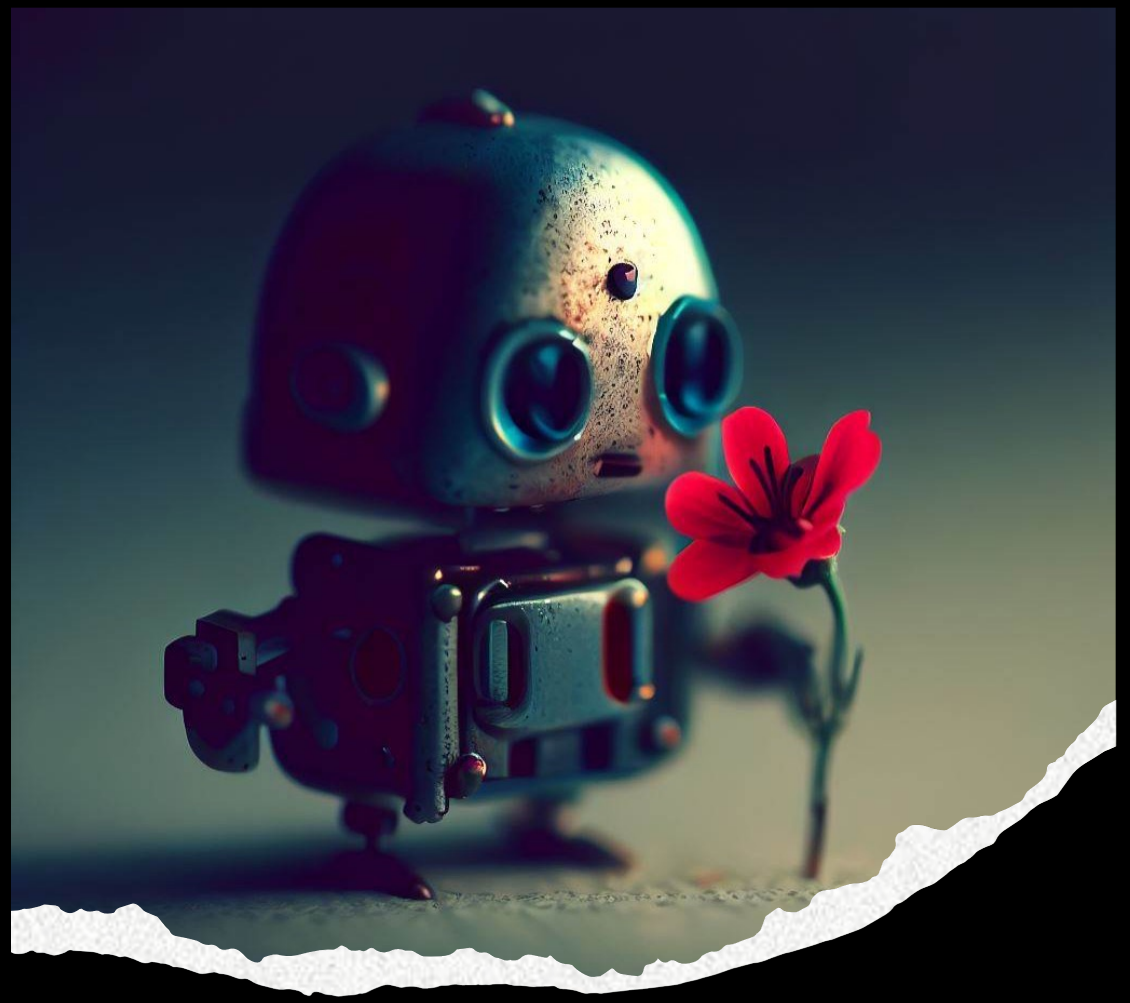
Veszélyek

- FakeVoice
- Hibás értelmezés
- Félreérthető hangsúlyozás
- Sértő viselkedés (szomorú hír vidám hangon felolvasása)

- Lehet, hogy „Gépies” hangot kell majd fejleszteni?

Köszönöm

Deep learning alapú
beszédszintézis



Dr. Zainkó Csaba zainko@tmit.bme.hu

Tartalom	Human: Dr. Zainkó Csaba
Szöveg	Human: Dr. Zainkó Csaba
TTS beszédminták	**AI:** BME Profivox-DNN
Design	**AI** + Human: (MS powerpoint + ZCs)
Rajzok (4db)	Human: Dr. Zainkó Csaba
Illusztráció (1db)	?: https://www.freepik.com
Képek (5db)	**AI:** https://www.bing.com/images