Dr. Zainkó Csaba

A gépi beszédkommunikáció gyakorlati kihívásai

Beszédkommunikáció 01

Megoldott problémák

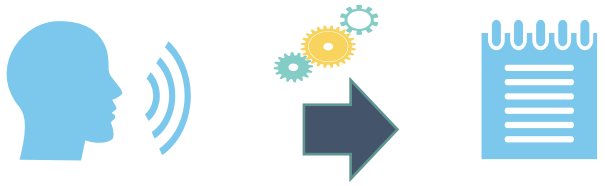
User Interface User eXperience 02

Van segítség

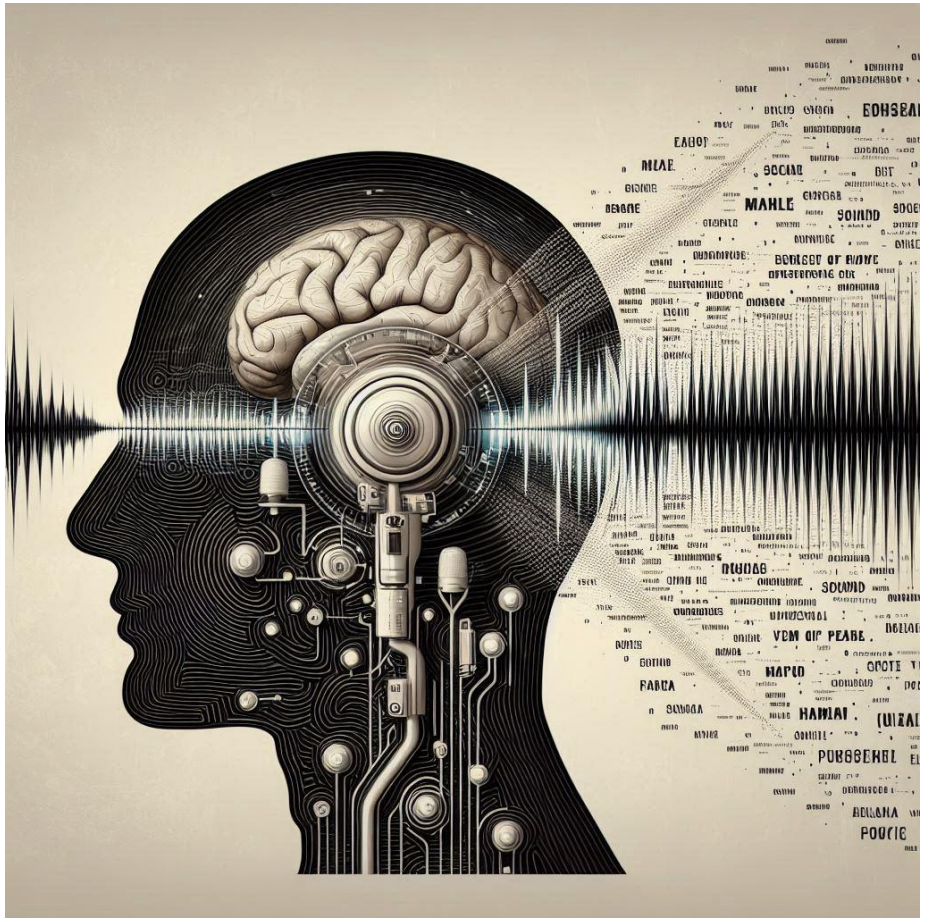
Mire kell figyelni? 03

Jó rendszer

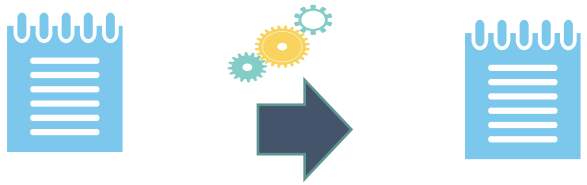
Beszédfelismerés



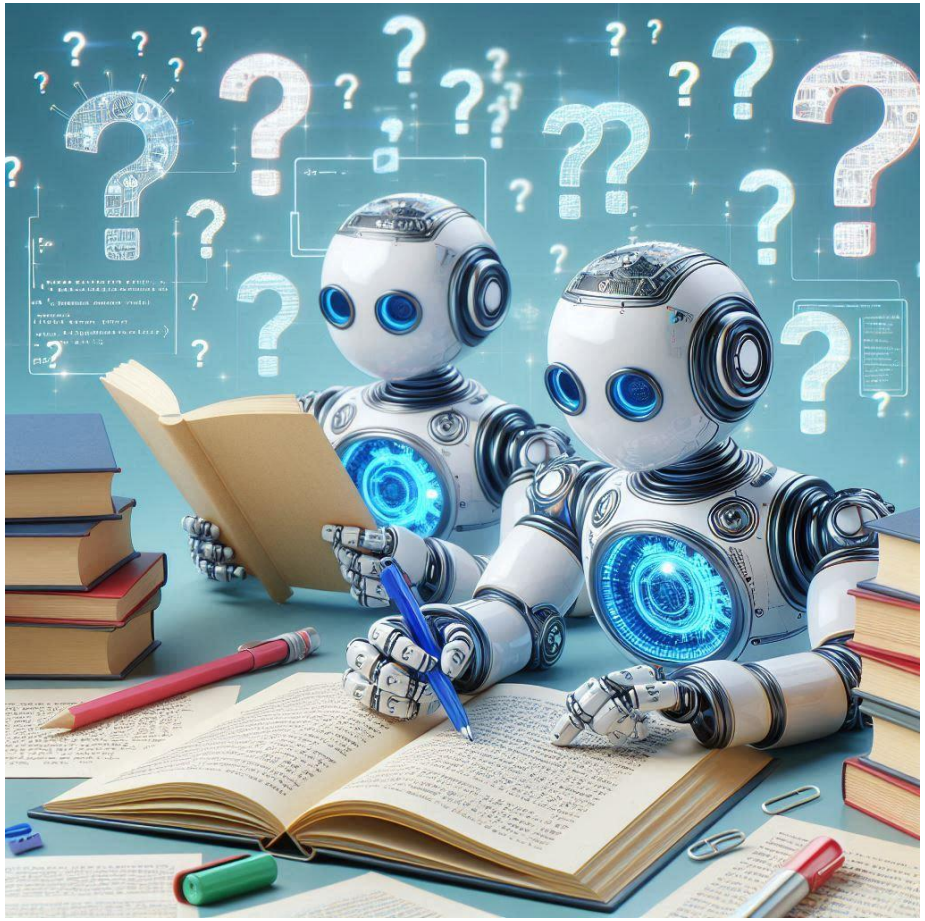
- Beszéd Szöveg



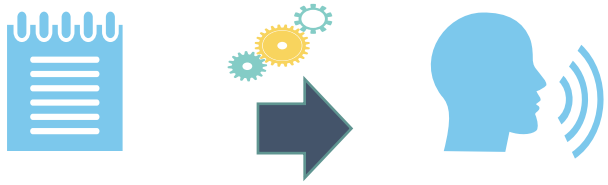
Dialógusvezérlés



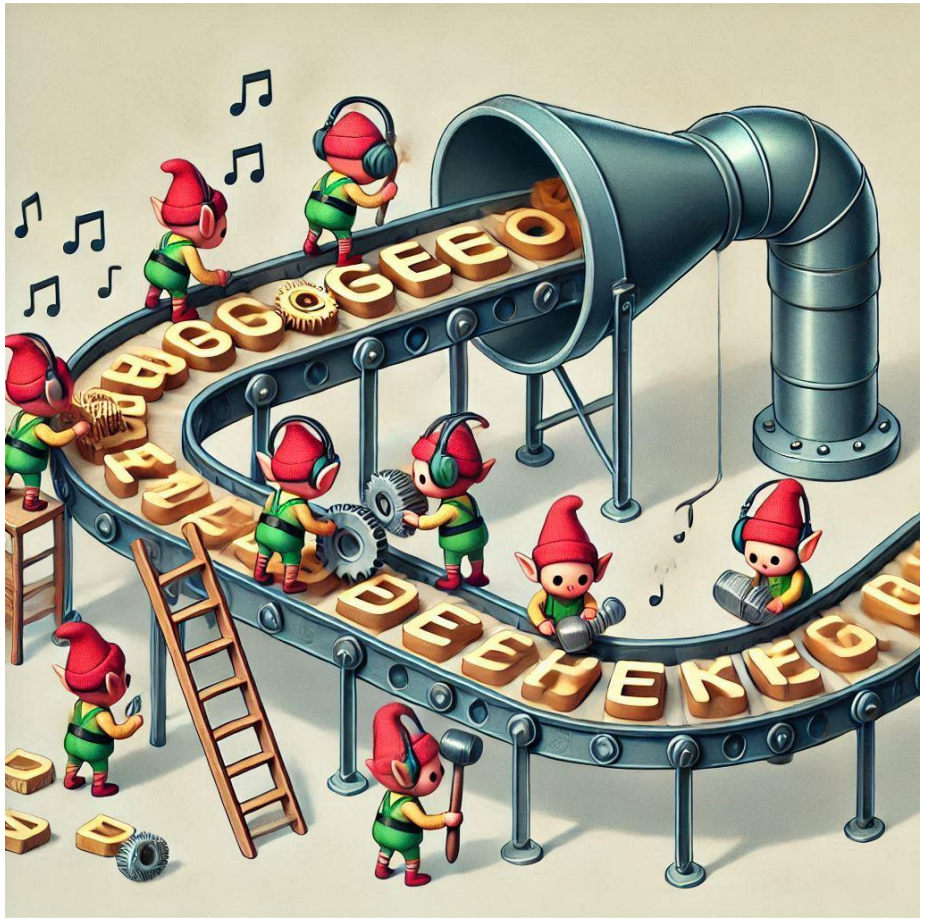
- Szöveg Válasz



Beszédszintézis



- Szöveg Beszéd



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

Kész?


ChatGPT 4o
“Hello.”
00:00
Hello! How's it going?
“Tudunk magyarul beszélgetni?”
00:01
Persze, nagyon szívesen! Miről szeretnél beszélgetni?
“Beszéljük kommunikációról, mondjál nekem valamit.”
00:03
Üzene

Példák kritikus pontokra

- Nyelvdetekció
  - Néha elméletileg sem kitalálható
- Beszédfelismerés
  - Nem 100%-os
- Dialógus
  - Szándék azonosítása
- Beszédszintézis
  - Félreértelmezés (hangsúly)
  - Kevert nyelvű mondatok
  - Minőség

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)

Beszéd

- End-to-End megoldások

Table 1. Comparison of evaluated MOS with 95% confidence intervals on the LJ Speech dataset.

Model	MOS (CI)
Ground Truth	4.46 (±0.06)
Tacotron 2 + HiFi-GAN	3.77 (±0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 (±0.07)
Glow-TTS + HiFi-GAN	4.14 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 (±0.07)
VITS (DDP)	4.39 (±0.06)
VITS	**4.43 (±0.06)**

Forrás: https://arxiv.org/pdf/2106.06103
https://jaywalnut310.github.io/vits-demo/index.html

Hogyan készítsünk jó rendszert?

- Grafikus UI-ra kidolgozott
- Beszédkommunikációra is használható
- Felhasználó, Feladatok, Környezet

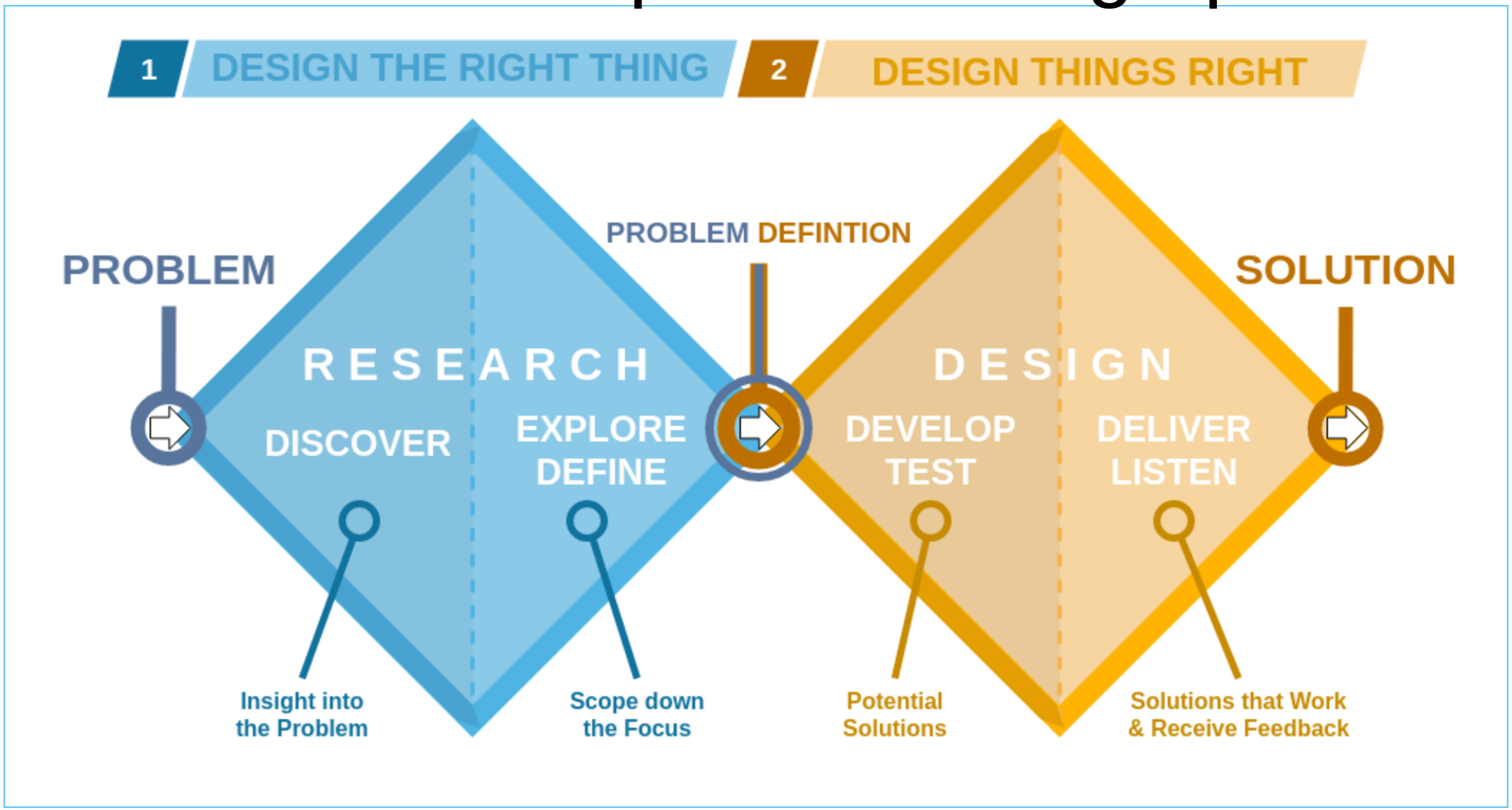
Iteratív modell

Design

Prototyping

User testing and evaluation

Double Diamond product design process



Forrás: https://en.wikipedia.org/wiki/Double_Diamond_%28design_process_model%29

Miért kell a felhasználó?

- Lakásvásárlás
- Weboldal
- Kóla

Szép → Jó

- Minőség
- Érthetőség

Beszédszintézis: Célközönség

- Igények figyelembe vétele
  - Férfi/női
  - Beszélő választás
  - Egyedi hang
  - Stílus
  - Érzelem???

- *Robotos v. Természetes*

Beszélő hangja

- Színész, bemondó, celeb?
  - Scarlett Johansson vagy hasonló
- Robotos, gépies

Hogyan tudok jó rendszert készíteni?

- Célközönség meghatározása
- Prototípus készítése
- Tesztelés

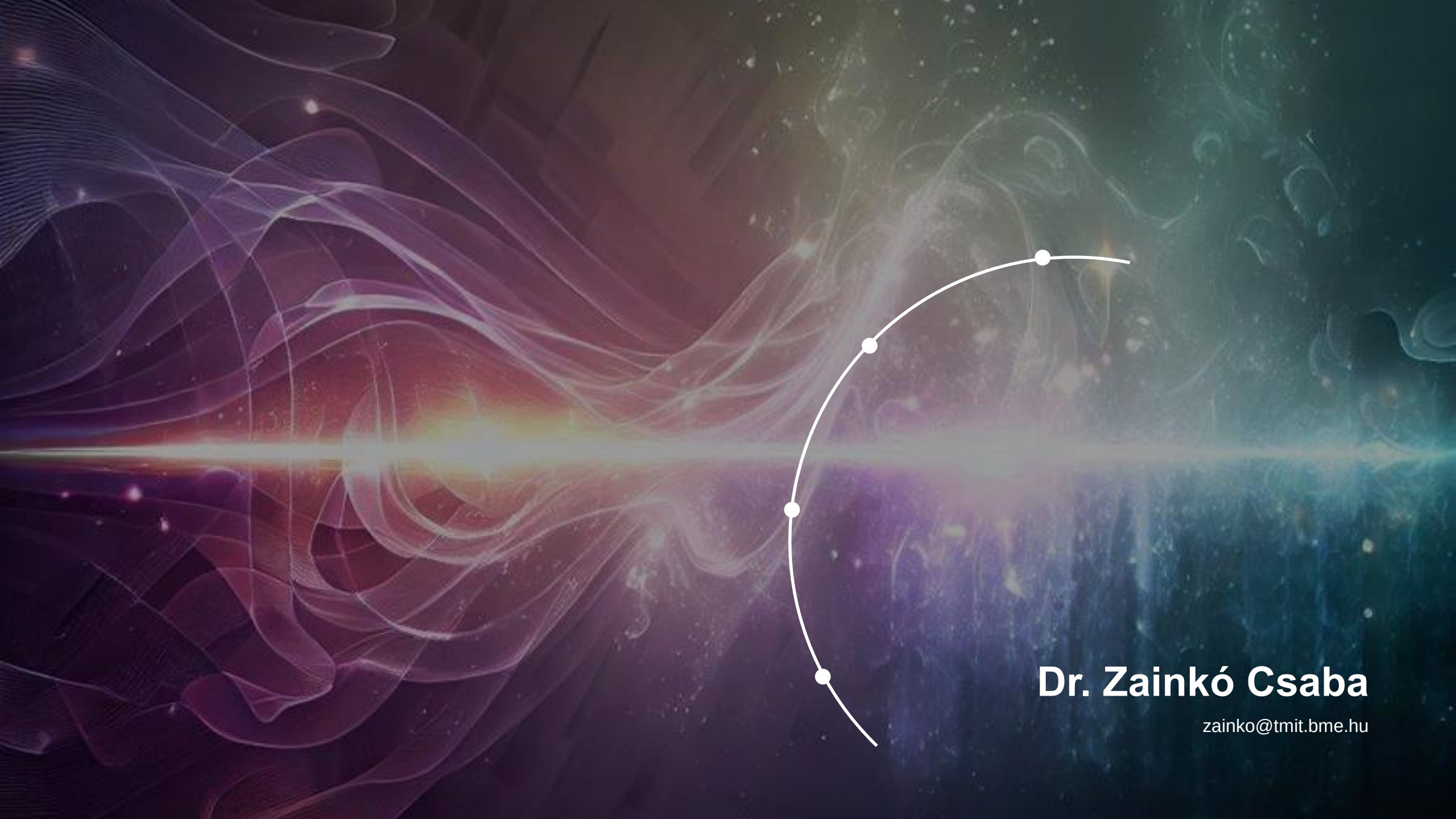
- Példák:
  - Komponensek megfelelő sebesség: 2mp szabály
  - Felhasználó irányít (érzés)
  - Választási lehetőségek korlátozása (7±2)
  - Mondatok stílusának meghatározása és konzekvens használata

Kihívások

- Beszédfelismerő
  - Zaj
  - Mit mondhat a felhasználó?
- Dialógusrendszer
  - Hogyan irányítsuk a felhasználó viselkedését
- Beszédszintézézis
  - Értelem szerinti hangsúlyozás
  - Kérdések

Összefoglalás

- Alaptechnológiák használhatók már
- Rendszerben kell gondolkodni
- Tesztelni,
- Tesztelni,
- …
- Tesztelni (felhasználókkal)



Dr. Zainkó Csaba

zainko@tmit.bme.hu