Paradigme moderne în TTS

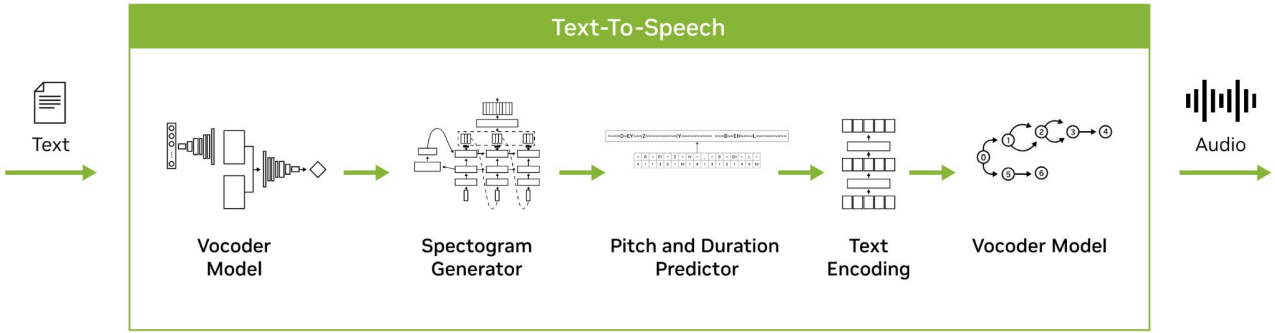
Tudor Bumbu

Sinteza vorbirii în ziua de azi

În sistemele Text-to-Speech (TTS) moderne, generarea vocii folosește frecvent o reprezentare intermediară: mel-spectrograma.

Arhitectura clasică modernă are două etape:

1. Modelul acustic transformă textul scris într-o mel-spectrogramă (care conține informația vizuală a ritmului și intonației).
2. Vocoderul transformă mel-spectrograma într-un semnal audio real (forma de undă).



TTS bazat pe modele LLM

Integrarea LLM-urilor în TTS adaugă „inteligență lingvistică” generării sunetului. Astfel, vocea capătă expresivitate, emoție și un ritm uman natural.

Exemple recente de sisteme:

- *Qwen3-TTS*
- *Orpheus-TTS*

O platformă pentru a testa și compara astfel de modele este *TTS Arena*.

Aplicații

Vocea generată artificial este extrem de calitativă și a devenit omniprezentă:

- Citirea automată a textului pentru persoanele cu deficiențe de vedere.
- Narare automată pentru cărți și materiale interactive.
- IAsistenți virtuali, navigație GPS, roboți, sisteme IVR.
- Voci pentru personajele din jocuri sau videoclipuri.

Dubbing

Dubbing-ul înseamnă înlocuirea vocii dintr-un material video cu o voce într-o altă limbă.

Această sarcină poate fi automatizată complet:

1. ASR: Transcrie fișierul audio sursă (ex. folosind Whisper).
2. Traducere automată: Textul transcris este tradus automat în limba țintă.
3. TTS: Noul text este generat vocal, sincronizând ritmul și stilul cu videoclipul original.

Acest flux demonstrează integrarea tehnologiilor limbajului uman (HLT).

Clonarea vocii

Datorită avansului modelelor AI, a apărut conceptul de Voice Cloning (clonarea vocii).

- Vocea unei persoane poate fi reprodusă automat din doar câteva secunde de înregistrare.

Atenție! Poate fi folosită pentru deepfake-uri audio, fraude financiare, ocolirea autentificării vocale sau dezinformare.