

Prelucrarea și recunoașterea vorbirii

Prezentat de Tudor Bumbu

Recunoașterea și sinteza vorbirii

În tehnologiile vocale există două procese fundamentale care lucrează împreună:

Recunoașterea Automată a Vorbirii (ASR)

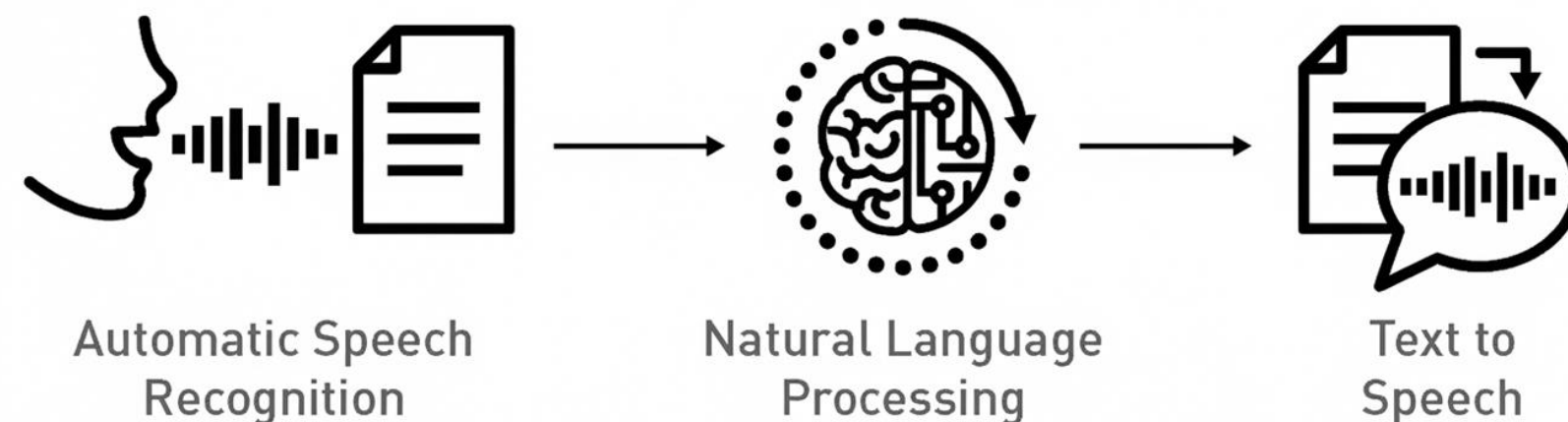
Transformă vocea în text (speech-to-text).

Funcționează ca „urechile” sistemului.

Sinteza Text-Vorbire (TTS)

Transformă textul în voce (text-to-speech).

Funcționează ca „gura” sistemului.



Recunoașterea vorbirii (ASR)

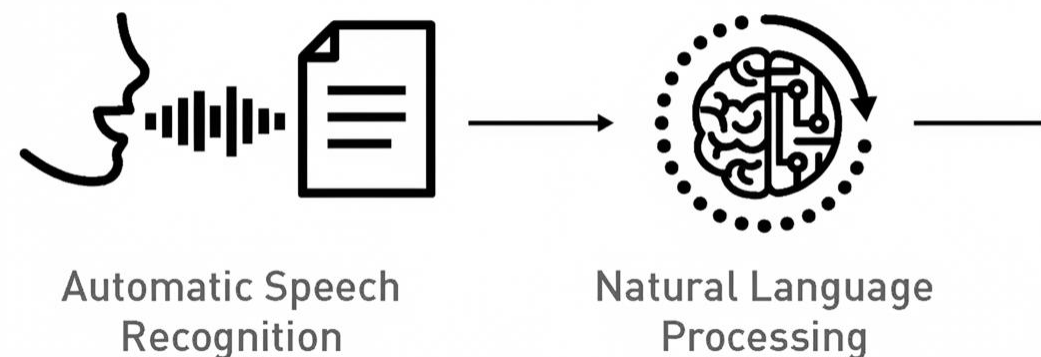
Un sistem de ASR primește un semnal vocal continuu, foarte variabil, și trebuie să producă o secvență de cuvinte.

Formulare:

Vrem să găsim propoziția cea mai probabilă dat fiind semnalul audio. Format din model acustic și model lingvistic (probabilitate cuvinte).

Provocarea: variabilitatea

- oamenii vorbesc diferit (accent, viteză, timbru),
- aceeași persoană nu sună la fel mereu (oboseală, emoții); condiții de mediu: zgomot, microfon slab, ecou.



Sinteza vorbirii (TTS)

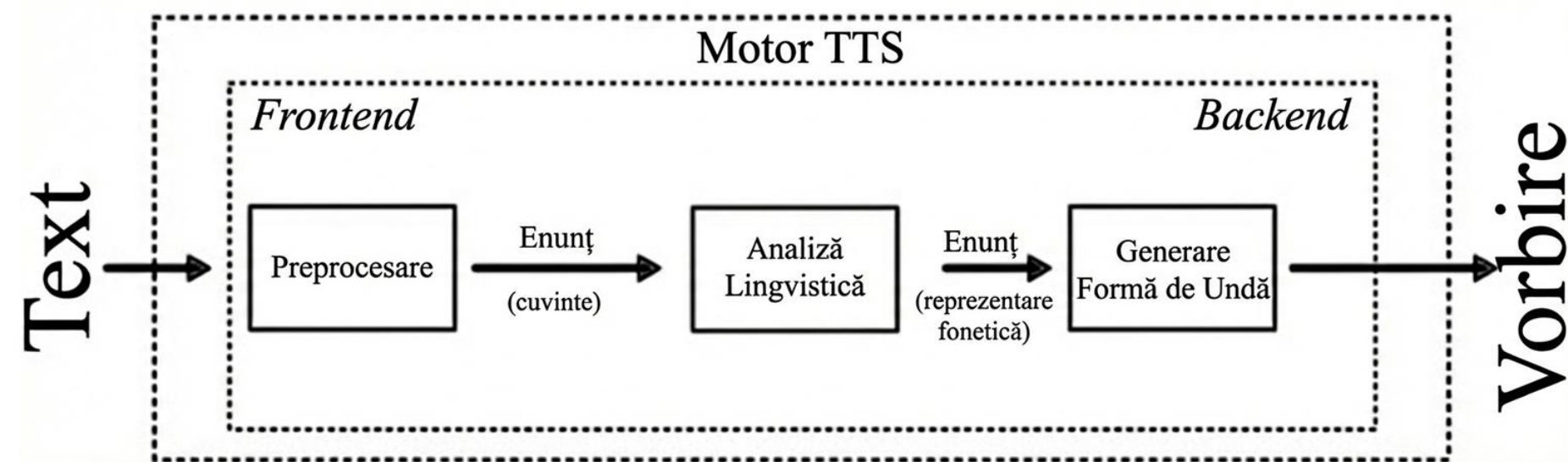
În sistemele TTS din text obținem o voce care sună natural.

Front-end:

normalizează textul (numere, abrevieri),
și decide prozodia (pauze, intonație, durate).

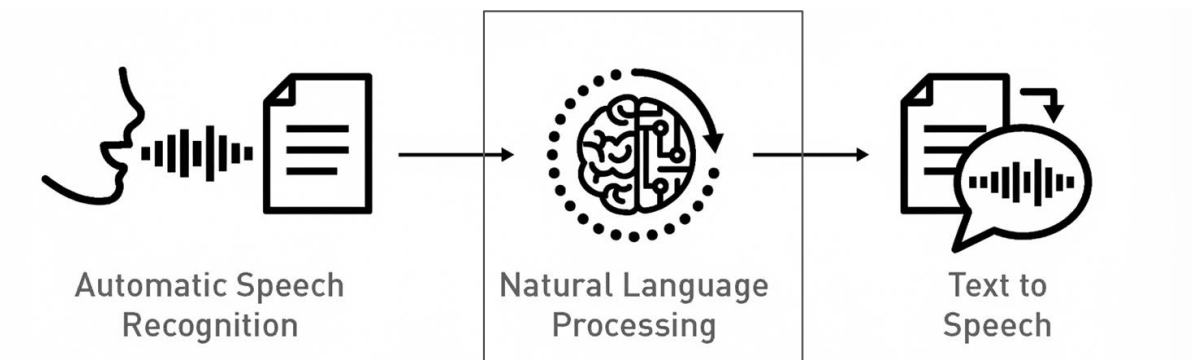
Back-end (Vocoder):

transformă reprezentarea acustică în semnal audio.
Exemple: WaveNet, HiFi-GAN, VITS.



„Creierul” unui sistem conversațional

ASR transcrie vocea, iar TTS o generează, însă „înțelegerea” comenzii are loc în modulul de ***Procesare a Limbajului Natural (NLP)***.



Pentru un dialog eficient, NLP folosește două direcții:

NLU (Înțelegerea Limbajului Natural)

Identifică intenția utilizatorului (ex: *SetAlarm*) și extrage entitățile (ex: *mâine, ora 7*).

NLG (Generarea Limbajului Natural):

Formulează un răspuns în limbaj natural

(„*Am setat alarma pentru mâine la ora 7*”).

Ansamblul ASR, NLP, TTS formează domeniul ***Tehnologiilor Limbajului Uman (HLT)***.

Arhitectura unui asistent vocal

Un asistent vocal (Spoken Dialogue System) funcționează modular unde ieșirea unui modul devine intrarea următorului.

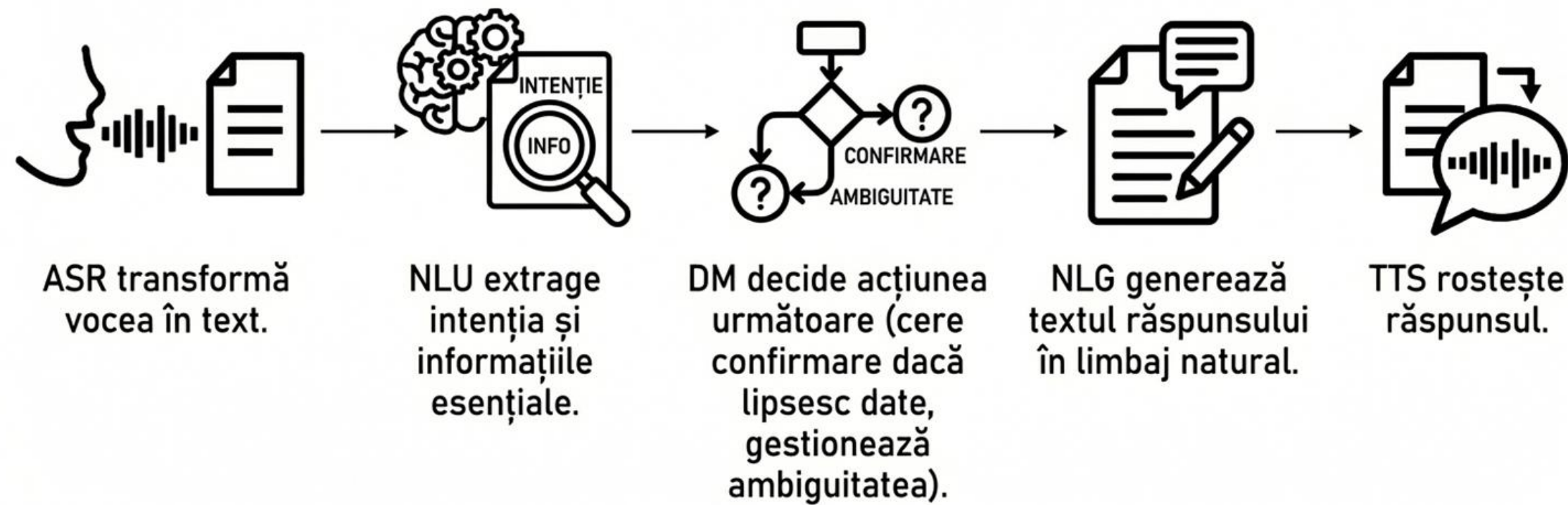
Fluxul de procesare:

1. ASR transformă vocea în text.
2. NLU extrage intenția și informațiile esențiale.
3. DM (*Dialogue Management*) decide acțiunea următoare (cere confirmare dacă lipsesc date, gestionează ambiguitatea).
4. NLG generează textul răspunsului în limbaj natural.
5. TTS rostește răspunsul.

Atenție! Într-o astfel de arhitectură, erorile se propagă. O transcriere greșită la nivel de ASR va duce la decizii greșite în modulul DM.

Arhitectura unui asistent vocal

Un asistent vocal (Spoken Dialogue System) funcționează modular unde ieșirea unui modul devine intrarea următorului.



Atenție! Într-o astfel de arhitectură, erorile se propagă. O transcriere greșită la nivel de ASR va duce la decizii greșite în modulul DM.

Evoluția asistenților vocali

Recunoașterea vorbirii a fost una dintre cele mai dificile probleme din IA. Progresul a fost susținut de volume masive de date, putere de calcul și modele noi.



Etape istorice:

În anii '50-'60 apar primele sisteme (Audrey, Shoebox) bazate pe potrivirea de șabloane.

Începând cu 2010 apar asistenții moderni (Siri, Alexa). Adoptarea pe scară largă a accelerat colectarea datelor și antrenarea modelelor ASR.

În prezent avem modele *End-to-End* (Whisper). Acestea învață direct maparea *audio-text*, fiind robuste la zgomot și accente.