Aplicații Practice: Asistenți Vocali și Sisteme de Dialog

9.1 Introducere

Asistenții virtuali vocali reprezintă una dintre cele mai tangibile și larg răspândite aplicații ale inteligenței artificiale în procesarea vorbirii. De la agenți comerciali omniprezenți precum Siri (Apple), Alexa (Amazon) și Google Assistant, la aplicații specializate în domenii enterprise sau medicale, acești agenți conversaționali redefinesc interacțiunea om-mașină (HCI). Funcția lor primordială este de a facilita o interacțiune naturală și eficientă, utilizând vorbirea ca principal mediu de input și output.

Din punct de vedere academic, acești asistenți sunt implementări ale **Sistemelor de Dialog Vocal (Spoken Dialogue Systems - SDS)**. Arhitectura fundamentală a unui astfel de sistem a implicat, în mod tradițional, un pipeline modular complex, cuprinzând: recunoașterea automată a vorbirii (ASR), înțelegerea limbajului natural (NLU), gestiunea dialogului (DM), generarea limbajului natural (NLG) și sinteza vocală (TTS). Aceste module colaborează pentru a transforma o comandă vocală inițială într-un răspuns coerent și vocalizat. O analiză aprofundată a arhitecturilor sistemelor de dialog relevă o evoluție constantă de la modele bazate pe reguli la cadre statistice sofisticate.

Acest capitol analizează în detaliu această arhitectură, provocările tehnice ale fiecărei componente, tranziția recentă către modele lingvistice mari (LLM) ca orchestratori centrali și o aplicație conexă de mare complexitate: traducerea vocală automată (S2ST).

9.2 Arhitectura Modulară Tradițională (Pipeline)

Arhitectura clasică a unui sistem de dialog vocal, validată în decenii de cercetare este un pipeline secvențial în care output-ul unui modul devine input-ul următorului.

Modulele componente sunt:

1. **ASR (Automatic Speech Recognition):** Convertește semnalul audio într-o secvență de cuvinte (transcriere). Performanța acestui modul este critică, deoarece erorile de transcriere (măsurate prin WER) se propagă inevitabil și degradează calitatea întregului lanț de procesare.
2. **NLU (Natural Language Understanding):** Preia textul transcris și îl convertește într-o reprezentare semantică formală. Aceasta este, de obicei, descompusă în două sarcini:
    - **Clasificarea Intenției (Intent Classification):** Identificarea scopului principal al utilizatorului (de ex., SETEAZĂ_ALARMA, CERE_VREMEA).
    - Umplerea Sloturilor (Slot-Filling): Extragerea parametrilor (entităților) necesari pentru a îndeplini intenția (de ex., pentru SETEAZĂ_ALARMA, sloturile ar fi ora: "7" și perioada: "dimineața").

Modelele inițiale foloseau clasificatori separați și Câmpuri Aleatoare Condiționale (CRF) pentru slot-filling. Abordările moderne utilizează arhitecturi neuronale, adesea bazate pe BERT, care rezolvă cele două sarcini simultan (joint intent classification and slot filling), beneficiind de contextul bidirecțional (Chen, Zhu, & Wang, 2019). O provocare majoră aici este detecția Out-of-Domain (OOD), adică recunoașterea faptului că intenția utilizatorului nu face parte din setul de capabilități cunoscute ale sistemului (Larson et al., 2019). Framework-uri open-source precum RASA

implementează acest pipeline NLU (Bockl et al., 2020).

3. **DM (Dialog Management):** Acesta este "creierul" conversațional, responsabil cu menținerea contextului și decizia următorului pas. Gestiunea dialogului are două sub-componente:
    - **Urmărirea Stării Dialogului (Dialogue State Tracking - DST):** Menține o reprezentare a stării curente a conversației, acumulând informații (sloturi confirmate) peste mai multe tururi de dialog.
    - **Politica de Dialog (Dialogue Policy):** Selectează următoarea acțiune a sistemului (de ex., CERE_LOCAȚIA, CONFIRMĂ_ACȚIUNEA). Politicile variază de la automate cu stări finite (FSM), rigide și definite manual, la abordări statistice. Cadrul dominant în cercetarea academică este modelarea dialogului ca un **Proces Decizional Markov Parțial Observabil (POMDP)**. În acest cadru, starea dialogului este "parțial observabilă" din cauza incertitudinii introduse de erorile ASR și ambiguitatea NLU. Politica de dialog este antrenată, adesea folosind **Învățarea prin Întărire (Reinforcement Learning - RL)**, pentru a maximiza o funcție de recompensă pe termen lung (de ex., finalizarea cu succes a sarcinii într-un număr minim de tururi) (Young et al., 2013; Levin, Pieraccini, & Eckert, 2000).
4. **NLG (Natural Language Generation):** Preia acțiunea simbolică decisă de DM (de ex., confirm(alarm, time=7am)) și o formulează ca un răspuns textual coerent. Abordările variază de la sisteme bazate pe șabloane (template-based), care sunt sigure, dar repetitive și nenaturale, la modele neuronale (de ex., bazate pe RNN/LSTM sau Transformer) care pot genera răspunsuri fluente și stilizate, deși cu riscul de a omite informații critice (devieri semantice).
5. **TTS (Text-To-Speech):** Converteste răspunsul textual generat de NLG într-un semnal audio. După cum s-a discutat în Capitolul 6, modelele end-to-end moderne, cum ar fi Tacotron 2 sau VITS, produc o vorbire de înaltă fidelitate, care este esențială pentru percepția calității asistentului.

9.3 Paradigma Emergentă: Agenți Conversaționali bazați pe LLM

Arhitectura modulară tradițională, deși interpretabilă, suferă de o mentenanță complexă și de propagarea erorilor în cascadă. Tendințele recente favorizează înlocuirea componentelor NLU, DM și NLG cu un singur model lingvistic mare (LLM).

În această paradigmă, LLM-ul nu este re-antrenat specific pentru sarcinile NLU/DM. În schimb, el utilizează **învățarea în context (in-context learning)**. Starea dialogului este pur și simplu istoricul conversației, injectat în prompt-ul modelului la fiecare tur. LLM-ul este instruit să completeze conversația, generând direct răspunsul textual al sistemului.

Provocarea fundamentală în această arhitectură este **"grounding-ul"** (ancorarea) – conectarea modelului lingvistic (care operează pe text nestructurat) la surse de date reale sau API-uri (de ex., accesarea calendarului, verificarea vremii). Modele recente, precum "Toolformer", demonstrează cum un LLM poate fi antrenat să genereze el însuși apelurile API necesare, să le insereze în raționamentul său textual și să utilizeze rezultatele pentru a formula răspunsul final (Schick et al., 2024).

9.4 Studiu de Caz Avansat: Traducerea Vocală Automată (S2ST)

O extindere notabilă a sistemelor vocale este traducerea vocală automată (Speech-to-Speech Translation - S2ST), procesul de conversie a vorbirii dintr-o limbă sursă direct în vorbire într-o limbă țintă.

9.4.1. Sisteme în cascadă pentru traducerea vorbirii în vorbire (S2ST)

Modelul tradițional pentru traducerea automată a vorbirii în vorbire (Speech-to-Speech Translation – S2ST) a fost reprezentat de un sistem modular în cascadă, structurat în trei etape distincte: recunoaștere automată a vorbirii (ASR) pentru transcrierea în limba sursă, urmată de o traducere automată a textului (MT) către limba țintă, iar în final, o sinteză vocală (TTS) care generează semnalul audio corespunzător în limba țintă (Gupta et al., 2024). Deși această arhitectură segmentată are avantajul controlului și optimizării fiecărei componente în mod independent, ea prezintă limitări structurale semnificative care afectează calitatea și naturalețea rezultatelor obținute.

Una dintre problemele fundamentale ale acestui model modular este propagarea erorilor. Astfel, orice eroare generată în prima etapă – cea de ASR – se transmite nefiltrat către componenta de traducere, amplificând distorsiunile semantice, iar ulterior acestea sunt reproduse în formă audio de modulul TTS, ceea ce duce la o degradare semnificativă a sensului original. Mai mult, această arhitectură suferă de o latență cumulativă ridicată, întrucât fiecare componentă introduce un timp de procesare adițional, făcând sistemul per ansamblu inadecvat pentru aplicații interactive sau conversaționale în timp real.

Un alt dezavantaj esențial constă în pierderea informației paralingvistice în timpul tranziției de la audio la text. În etapa de ASR, semnalul audio este redus la o transcriere textuală, ceea ce duce la eliminarea completă a elementelor precum intonația, emoția, ritmul, accentul sau identitatea vocală a vorbitorului. Aceste trăsături, deși non-lexicale, sunt esențiale pentru comunicare autentică și înțelegere empatică. În lipsa lor, componenta TTS generează o voce sintetică generică, adesea percepută ca plată sau robotică, diminuând considerabil naturalețea și expresivitatea mesajului tradus.

9.4.2. Sisteme directe end-to-end (E2E) în traducerea vorbirii

Pentru a depăși limitările arhitecturilor modulare, cercetarea recentă în domeniul traducerii automate a vorbirii s-a orientat către dezvoltarea sistemelor directe, de tip end-to-end (E2E). Acestea urmăresc realizarea unei mapări directe de la semnalul acustic al limbii sursă la semnalul acustic al limbii țintă, evitând etapele intermediare de conversie în text. Avantajul principal al acestei abordări constă în capacitatea de a păstra trăsăturile prosodice și identitatea vocală a vorbitorului original, facilitând astfel o experiență auditivă mult mai naturală și fidelă originalului (Jia et al., 2021).

Un exemplu emblematic în această direcție este modelul **Translatotron 2**, propus de cercetătorii de la Google Research. Acesta integrează o arhitectură compusă dintr-un encoder acustic, care procesează conținutul lingvistic al semnalului sursă, și un encoder vocal, care extrage embedding-ul vocal al vorbitorului – o reprezentare latentă a trăsăturilor unice ale vocii acestuia. Ulterior, un decoder TTS condiționat de aceste două componente generează semnalul audio în limba țintă, reușind să păstreze atât conținutul semantic, cât și particularitățile vocale ale vorbitorului original. Acest tip de integrare permite conservarea identității vocale în procesul de traducere, oferind nu doar o translație lingvistică, ci și una afectivă și stilistică (Jia et al., 2021).

Pe lângă Translatotron, cercetările recente au introdus sisteme de amploare și mai mare. Modelul **SeamlessM4T**, dezvoltat de Meta AI (2023), reprezintă un sistem multimodal de tip fundational, capabil să proceseze vorbire, text și traducere pentru peste 100 de limbi, într-un singur flux integrat. SeamlessM4T este conceput pentru a menține coerența translingvistică și pentru a gestiona diferite forme de intrare – inclusiv text și vorbire – într-o manieră fluentă și scalabilă. În mod particular, acest model adresează provocările legate de variația accentelor, condițiile acustice adverse și alternarea codurilor (code-switching).

Într-o direcție complementară, modelul **Hibiki**, propus de Labiausse et al. (2025), se concentrează pe scenarii de traducere în streaming, unde latența joacă un rol critic. Acest model este optimizat pentru a păstra izocronia – adică alinierea temporală între vorbirea sursă și cea țintă – aspect esențial în aplicații precum dublajul video, unde sincronizarea labială și ritmul sunt esențiale pentru naturalețea percepută a traducerii. Hibiki integrează mecanisme de anticipare și adaptare temporală, astfel încât traducerea să fie generată în timp real, fără a compromite fidelitatea semantică sau ritmul enunțului.

Prin urmare, tranziția de la arhitecturi în cascadă la sisteme end-to-end marchează un punct de cotitură în domeniul traducerii automate a vorbirii, reflectând o maturizare a tehnologiei și o orientare către modele mai expresive, mai robuste și mai apropiate de complexitatea comunicării umane.

9.5 Evaluarea Sistemelor de Dialog

Evaluarea sistemelor de dialog vocale constituie o provocare metodologică semnificativ mai complexă decât evaluarea izolată a componentelor individuale, precum recunoașterea automată a vorbirii (ASR) sau sinteza vocală (TTS). În timp ce performanța ASR poate fi cuantificată prin indicatori standard, cum ar fi rata de eroare la cuvânt (Word Error Rate – WER), iar calitatea TTS este adesea evaluată prin scoruri de tip Mean Opinion Score (MOS), un sistem complet de dialog necesită o analiză holistică care să surprindă nu doar precizia tehnică, ci și calitatea percepută a interacțiunii și eficiența comunicării.

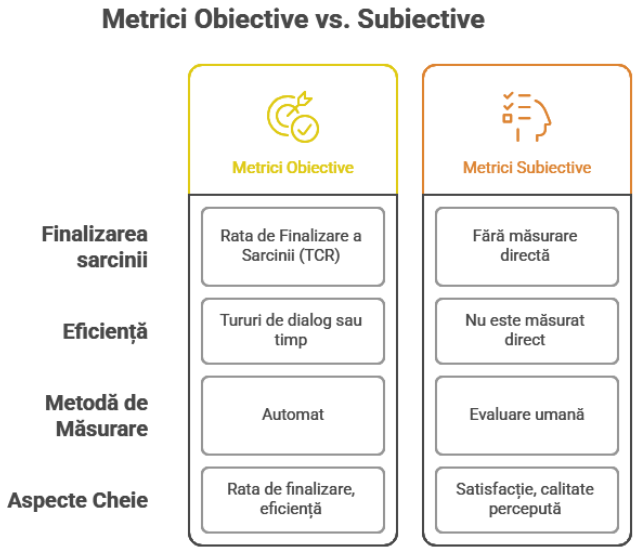

Fig. 3 Comparare metrici de evaluare

În acest context, evaluarea se desfășoară pe două dimensiuni principale: obiectivă și subiectivă. Metricile obiective sunt calculate automat și reflectă performanța sistemului în îndeplinirea scopurilor definite. Una dintre cele mai utilizate măsuri este rata de finalizare a sarcinii (Task Completion Rate – TCR), care exprimă proporția dialogurilor în care obiectivul utilizatorului este atins cu succes. Această metrică este deosebit de relevantă în cazul sistemelor orientate spre sarcini, precum cele utilizate pentru rezervări, asistență tehnică sau programări. Un alt indicator obiectiv important este eficiența dialogului, măsurată adesea prin numărul de tururi de conversație (turns) necesare pentru rezolvarea unei sarcini sau prin timpul total de interacțiune. Astfel, un sistem care atinge obiectivul într-un număr minim de pași este considerat mai eficient.

Pe de altă parte, metricile subiective sunt derivate din percepțiile utilizatorilor reali sau evaluatori umani. Satisfacția utilizatorului reprezintă un criteriu central și este frecvent măsurată prin chestionare aplicate post-interacțiune, precum scalele Likert sau instrumente standardizate precum System Usability Scale (SUS). Calitatea percepută a dialogului este de asemenea analizată prin evaluări umane asupra unor dimensiuni precum naturalețea (gradul în care vocea pare umană), coerența răspunsurilor (consistența logică în contextul conversațional), și inteligența percepută a sistemului (impresia că sistemul „înțelege” intențiile utilizatorului).

Un cadru de referință recurent în literatura de specialitate este modelul PARADISE (PARAdigm for DIalogue System Evaluation), propus de Walker et al. (1997), care oferă o metodologie formală de integrare a metricilor obiective și subiective. Acest cadru utilizează analiza de regresie pentru a construi un model predictiv al satisfacției utilizatorului, pornind de la o combinație ponderată de măsuri cuantificabile – precum TCR – și factori de cost ai dialogului, cum ar fi latența, numărul de erori generate de componentele ASR sau NLG (Natural Language Generation), și redundanța întrebărilor de clarificare. Astfel, se obține un model comprehensiv care corelează performanțele tehnice cu experiența umană, facilitând optimizarea globală a sistemului în raport cu scopurile sale funcționale.

În concluzie, evaluarea sistemelor de dialog nu poate fi redusă la simple scoruri cantitative. Ea trebuie să reflecte complexitatea interacțiunii om–mașină, integrând atât date obiective despre eficiență, cât și feedback-ul subiectiv al utilizatorilor, pentru a construi sisteme vocale robuste, utile și acceptate social.

9.6 Provocări Actuale și Direcții Viitoare

În ciuda progreselor rapide, dezvoltarea asistenților vocali și a sistemelor S2ST se confruntă cu provocări semnificative:

1. **Robustețea în Condiții Reale:** Performanța ASR degradează dramatic în prezența zgomotului de fond, a reverberațiilor (recunoaștere "far-field") sau a accentelor și dialectelor ne-standard.
2. **Disponibilitatea Datelor:** Modelele S2ST E2E necesită volume uriașe de date vocale paralele, care sunt extrem de rare. Modelele fundamentale precum SeamlessM4T încearcă să rezolve acest deficit prin pre-antrenare masivă pe date neetichetate și slab supervizate (Meta AI, 2023).
3. **Latența:** Pentru o interacțiune naturală, răspunsul sistemului trebuie să fie aproape instantaneu. Latența este o problemă majoră atât pentru modelele LLM masive, cât și pentru

pipeline-urile S2ST în streaming (Labiausse et al., 2025).

4. **Păstrarea Nuanțelor (Grounding & Prosody):** În S2ST, păstrarea identității vocale și a prosodiei (emoției) rămâne o provocare[32]. În sistemele bazate pe LLM, "grounding-ul" corect al limbajului în acțiuni API reale și evitarea "halucinațiilor" sunt probleme deschise.
5. **Bias și Echitate (Bias and Fairness):** Așa cum este detaliat în Capitolul 8, sistemele vocale pot manifesta bias-uri, având performanțe scăzute pentru grupuri demografice sau lingvistice subreprezentate în datele de antrenament.
6. **Evaluarea Robustă:** Pe măsură ce sistemele devin mai creative (în special cele bazate pe LLM), metricile tradiționale precum TCR devin insuficiente. Evaluarea interacțiunilor deschise (open-domain) rămâne o provocare majoră.

9.7 Concluzii

Sistemele de dialog vocal și traducerea vocală automată sunt printre cele mai avansate și cu cel mai mare impact aplicații ale tehnologiilor de prelucrare a vorbirii[35]. Ele transformă modul în care interacționăm cu tehnologia, oferind interfețe naturale și eliminând barierele lingvistice[36]. Trecerea de la pipeline-uri modulare rigide la modele E2E unificate și la agenți bazați pe LLM deschide calea către asistenți virtuali cu adevărat globali, conștienți de context și expresivi. Cu toate acestea, atingerea acestui obiectiv necesită rezolvarea provocărilor tehnice și etice considerabile legate de robustețe, latență, disponibilitatea datelor și echitate.