Analiza Semantică și Biometrică a Vorbirii

7.1. Introducere

Pe măsură ce sistemele de procesare a vorbirii devin tot mai sofisticate, ele depășesc simpla transcriere (ASR) sau generare (TTS) a conținutului lingvistic. Tehnologiile moderne explorează straturi mai profunde de informație conținute în semnalul vocal, răspunzând la întrebări precum "Cine vorbește?" și "Cum se simte vorbitorul?". Aceste două direcții, **recunoașterea vorbitorului** (o formă de biometrică vocală) și **recunoașterea emoțiilor din vorbire (Speech Emotion Recognition - SER)**, reprezintă frontiere active de cercetare și dezvoltare, cu aplicații de la securitate și autentificare la interfețe om-calculator empatice și analiza avansată a conversațiilor.

Acest capitol oferă o incursiune detaliată în aceste două domenii, prezentând conceptele fundamentale, evoluția tehnicilor de la modele statistice clasice la embedding-uri neuronale de ultimă generație și provocările specifice fiecărei sarcini. Vom analiza arhitecturi de referință precum i-vector și x-vector pentru recunoașterea vorbitorului și vom explora diversele abordări de extragere a trăsăturilor și de clasificare pentru SER.

În paralel, un sistem tehnologic este la fel de valoros pe cât de riguros poate fi evaluat. Prin urmare, o componentă esențială a acestui capitol este dedicată metricilor de evaluare a performanței. Vom recapitula și aprofunda standardele industriei pentru ASR și TTS — Word Error Rate (WER) și Mean Opinion Score (MOS) — și vom introduce metricile specifice utilizate pentru a cuantifica acuratețea în recunoașterea vorbitorului (ex: Equal Error Rate - EER) și a emoțiilor (ex: Weighted/Unweighted Accuracy). Obiectivul este de a oferi un cadru complet pentru înțelegerea, implementarea și evaluarea sistemelor vocale avansate.

7.2. Recunoașterea Vorbitorului (Speaker Recognition)

Recunoașterea vorbitorului este procesul de identificare sau autentificare a unei persoane pe baza caracteristicilor unice ale vocii sale. Vocea umană este un semnal biometric complex, care poartă amprenta atât a caracteristicilor fiziologice ale tractului vocal, cât și a particularităților comportamentale învățate. Sistemele automate exploatează această unicitate pentru a crea "amprente vocale" digitale.

7.2.1. Identificare vs. Verificare

În funcție de scopul final, sarcinile de recunoaștere a vorbitorului se împart în două categorii distincte:

1. **Identificarea Vorbitorului (Speaker Identification):** Aceasta este o problemă de clasificare **1:N** (unu-la-mulți). Dat un enunț vocal de la un vorbitor necunoscut, sistemul trebuie să determine care dintre cei N vorbitori dintr-o bază de date pre-înregistrată este cel mai probabil să fi produs acel enunț. Este similar cu a răspunde la întrebarea "Cine din această mulțime vorbește?".

2. **Verificarea Vorbitorului (Speaker Verification):** Aceasta este o problemă de clasificare binară **1:1** (unu-la-unu), cunoscută și ca autentificare vocală. Sistemul primește un enunț vocal și o identitate declarată (ex: "Eu sunt utilizatorul X"). Sarcina sa este de a confirma sau infirma această afirmație, răspunzând la întrebarea "Este această persoană cine pretinde a fi?". Sistemul

compară enunțul de test cu modelul vocal (amprenta vocală) al utilizatorului X și produce un scor de similaritate. Decizia de "acceptare" sau "respingere" se ia prin compararea acestui scor cu un prag predefinit.

Costurile erorilor în verificarea vorbitorului sunt adesea asimetrice. O falsă acceptare (False Acceptance - FA), unde un impostor este acceptat ca fiind utilizatorul legitim, este de obicei o eroare de securitate mult mai gravă decât o falsă respingere (False Rejection - FR), unde utilizatorul legitim este respins. Metrica standard pentru evaluarea sistemelor de verificare este Rata de Eroare Egală (Equal Error Rate - EER), care reprezintă punctul de operare pe curba ROC (Receiver Operating Characteristic) sau DET (Detection Error Tradeoff) unde rata de false acceptări (FAR) este egală cu rata de false respingeri (FRR). O altă metrică importantă este Funcția de Cost de Detecție (Detection Cost Function - DCF), care permite ponderarea costurilor celor două tipuri de erori în funcție de cerințele aplicației.

7.2.2. Evoluția Tehnicilor: De la i-vectors la Embedding-uri Neuronale

Metodele de recunoaștere a vorbitorului au evoluat de la abordări generative, statistice, la cele discriminative, bazate pe rețele neuronale profunde.

Abordarea i-vector: Timp de aproape un deceniu, paradigma dominantă a fost cea bazată pe i-vectors (identity vectors). Această metodă este generativă și se bazează pe modelarea statistică a variabilității totale dintr-un semnal vocal. Pipeline-ul de extragere a unui i-vector implică următorii pași:

1. Se antrenează un Model Universal de Fundal (Universal Background Model - UBM), de obicei un GMM cu un număr mare de componente (ex: 2048), pe un set de date vast și divers. UBM-ul învață să modeleze distribuția generală a trăsăturilor acustice, independent de vorbitor.
2. Pentru un enunț specific, statisticile Baum-Welch sunt colectate prin alinierea trăsăturilor acustice ale enunțului la componentele UBM.
3. Se postulează că supervectorul mediei GMM adaptat pentru un anumit enunț, M, poate fi descompus ca: $M = m + T \cdot w$, unde m este supervectorul mediei UBM, T este o matrice de rang redus numită matricea de variabilitate totală, iar w este i-vectorul.
4. Matricea T este antrenată pentru a captura principalele direcții de variație în spațiul GMM, care includ atât variabilitatea inter-vorbitor, cât și cea intra-vorbitor (de canal, de zgomot etc.). I-vectorul w este o proiecție de dimensiune redusă (tipic 400-600) care codifică deviația unui enunț specific de la media universală, în acest spațiu de variabilitate totală.
5. Pentru clasificare, se utilizează de obicei Analiza Liniară Discriminantă Probabilistică (Probabilistic Linear Discriminant Analysis - PLDA) pentru a modela și a separa variabilitatea inter- și intra-vorbitor din i-vectori.

Abordarea x-vector: Un salt de performanță semnificativ a fost adus de introducerea **x-vectors**, propusă de Snyder et al. (2018) în lucrarea "X-Vectors: Robust DNN Embeddings for Speaker Recognition".[5] Această abordare este complet discriminativă și înlocuiește întregul pipeline GMM-UBM cu o singură rețea neuronală profundă.

- Arhitectura: Rețeaua este de obicei o Rețea Neuronală cu Întârziere în Timp (Time Delay Neural Network - TDNN). Straturile inițiale ale TDNN procesează cadre acustice cu un context temporal limitat. Un strat de agregare statistică (statistics pooling) calculează media și deviația standard a activărilor de la un strat anterior pe întreaga durată a enunțului, transformând astfel o secvență de lungime variabilă într-un vector de dimensiune fixă. Acest vector este apoi procesat de câteva straturi complet conectate.
- Antrenarea: Rețeaua este antrenată pe o sarcină de clasificare multi-clasă, unde trebuie să identifice vorbitorul corect dintr-un set de mii de vorbitori din datele de antrenament.
- Extracția: După antrenare, stratul final de clasificare (softmax) este eliminat. X-vectorul este embedding-ul extras de la unul dintre straturile complet conectate dinaintea ieșirii (de ex., segment6 în arhitectura originală). Aceste embedding-uri, de dimensiune fixă (ex: 512), sunt optimizate pentru a fi discriminative între vorbitori. Similar cu i-vectors, x-vectorii sunt apoi procesați cu un backend PLDA pentru scorare.

Sistemele bazate pe x-vector, în special atunci când sunt antrenate cu tehnici de augmentare a datelor (adăugarea de zgomot și reverberație artificială), au demonstrat performanțe superioare față de sistemele i-vector, în special pe enunțuri scurte și în condiții acustice dificile.

Arhitecturi Moderne (ECAPA-TDNN și dincolo): Cercetarea a continuat să rafineze arhitectura de extragere a embedding-urilor. ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in TDNN) este o arhitectură de referință care introduce mai multe îmbunătățiri :

- Module SE-Res2Net: Înlocuiește straturile TDNN standard cu blocuri care combină arhitectura Res2Net cu mecanisme de atenție pe canale (Squeeze-and-Excitation), permițând modelului să învețe reprezentări multi-scală.
- Agregarea Trăsăturilor Multi-strat: Concatenează ieșirile de la diferite straturi ale rețelei, permițând pooling-ului să aibă acces la trăsături de la diferite niveluri de abstractizare.
- Pooling Atentiv: Înlocuiește pooling-ul statistic standard cu un mecanism de atenție care învață să pondereze importanța diferitelor cadre temporale în formarea embedding-ului final.
-

7.2.3. Provocări și Direcții Viitoare

Deși sistemele moderne au atins performanțe impresionante, recunoașterea vorbitorului în condiții reale ("in the wild") rămâne o provocare. Degradarea performanței este cauzată de [3]:

- Zgomot și Reverberație: Condițiile acustice necontrolate pot masca trăsăturile specifice vorbitorului.
- Variabilitatea Canalului: Diferențele între microfoane și canale de transmisie (ex: telefon mobil vs. microfon de studio) introduc distorsiuni.
- Durata Scurtă a Enunțurilor: Extragerea unui embedding robust din doar câteva secunde de vorbire este dificilă.

Direcțiile actuale de cercetare se concentrează pe creșterea robusteței prin tehnici precum învățarea multi-task (antrenarea simultană a modelului pentru recunoașterea vorbitorului și a altor sarcini, cum ar fi recunoașterea fonemelor sau a limbii) și denoising-ul embedding-urilor (utilizarea

de autoencodere pentru a "curăța" x-vectorii extrași din semnale zgomotoase).

7.3. Recunoașterea Emoțiilor din Vorbire (Speech Emotion Recognition - SER)

SER este subdomeniul procesării vorbirii care are ca scop clasificarea stării afective a unui vorbitor (ex: furie, tristețe, bucurie) pe baza caracteristicilor acustice și prozodice ale vocii. Emoțiile modulează fundamental modul în care vorbim, afectând tonul (pitch), intensitatea, ritmul și calitatea vocii.

7.3.1. Provocări Teoretice și Modele Emoționale

SER este o sarcină intrinsec dificilă din mai multe motive [3]:

- Ambiguitatea Emoțiilor: Emoțiile umane sunt complexe, adesea amestecate și dependente de context. Etichetarea datelor este subiectivă și poate varia semnificativ între diferiți anotatori umani.
- Variabilitate: Modul în care emoțiile sunt exprimate vocal variază enorm între indivizi, culturi și chiar pentru aceeași persoană în contexte diferite (variație intra- și inter-subiect).
- Dependența de Conținutul Lingvistic: Prozodia emoțională este adesea împletită cu prozodia lingvistică (ex: intonația interogativă vs. declarativă).

Există două paradigme principale pentru modelarea emoțiilor:

1. Modele Categoriale: Emoțiile sunt clasificate în categorii discrete, precum cele șase emoții de bază ale lui Ekman (furie, dezgust, frică, fericire, tristețe, surpriză), la care se adaugă adesea starea neutră și calmă.
2. Modele Dimensionale: Emoțiile sunt reprezentate ca puncte într-un spațiu continuu, de obicei definit de două sau trei axe, cum ar fi valența (pozitiv vs. negativ) și activarea (calm vs. excitat).

7.3.2. Extracția Trăsăturilor și Reprezentări Profunde

Abordările pentru SER pot fi împărțite în cele bazate pe extragerea manuală de trăsături și cele end-to-end.

Abordări Bazate pe Trăsături: Această abordare tradițională implică extragerea unui set de trăsături acustice predefinite, despre care se știe din studiile de fonetică că sunt corelate cu starea emoțională. Aceste trăsături sunt apoi folosite pentru a antrena un clasificator standard (ex: SVM, Random Forest). Trăsăturile comune includ :

- Trăsături Prozodice: Conturul frecvenței fundamentale (F_0 sau pitch), energia (intensitatea) semnalului și durata fonemelor/pauzelor.
- Trăsături de Calitate a Vocii: Jitter (variația perioadei pitch-ului), shimmer (variația amplitudinii) și raportul armonic-zgomot (HNR).
- Trăsături Spectrale: Coeficienții MFCC, formanții, centroizi spectrali etc.

De obicei, nu se folosesc valorile brute ale acestor trăsături, ci funcționale statistice aplicate pe contururile lor (ex: media, deviația standard, minim, maxim, panta regresiei liniare). Seturi de trăsături standardizate, precum cele din toolkit-ul openSMILE, pot conține mii de astfel de

funcționale.

Abordări Bazate pe Învățare Profundă: Abordările moderne tind să elimine extragerea manuală de trăsături și să utilizeze rețele neuronale pentru a învăța reprezentări relevante direct din date.

- Intrare: Modelele primesc ca intrare spectrograma (sau mel-spectrograma) semnalului, tratând-o ca pe o imagine. Unele modele mai recente operează chiar direct pe forma de undă brută.
- Arhitecturi: Arhitecturile hibride CNN-LSTM sunt foarte populare. Straturile convoluționale (CNN) învață să extragă trăsături locale spectro-temporale, în timp ce straturile recurente (LSTM) modelează dependențele temporale pe termen lung ale acestor trăsături.
- Mecanisme de Atenție: Adăugarea unui mecanism de atenție permite modelului să învețe să se concentreze pe segmentele cele mai relevante din punct de vedere emoțional ale unui enunț, îmbunătățind performanța. O lucrare influentă a lui Zhang et al. (2018) a propus o rețea complet convoluțională cu un mecanism de atenție care a obținut rezultate de referință pe setul de date IEMOCAP.

7.3.3. Metrici de Evaluare și Provocări

Performanța sistemelor SER este de obicei măsurată folosind acuratețea. Deoarece seturile de date emoționale sunt adesea dezechilibrate (unele emoții sunt mai frecvente decât altele), se raportează două tipuri de acuratețe :

- **Weighted Accuracy (WA):** Acuratețea medie generală, unde fiecare eșantion de test contribuie în mod egal. Aceasta poate fi înșelătoare dacă o clasă majoritară domină setul de date.
- **Unweighted Accuracy (UA):** Media acurateților pe fiecare clasă. Este echivalentă cu metrica "balanced accuracy" sau "macro-average recall" și oferă o imagine mai corectă a performanței pe clase minoritare.

Provocările majore în SER rămân legate de calitatea și cantitatea datelor. Majoritatea seturilor de date disponibile (ex: IEMOCAP, RAVDESS) conțin emoții simulate de actori, care pot să nu generalizeze bine la emoțiile spontane, din viața reală. Colectarea și etichetarea datelor spontane este extrem de dificilă și costisitoare.