UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI

Școala Doctorală de Electronică, Telecomunicații și Tehnologia Informației

Decizia Nr. 1048 **din** 10-07-2023

TEZĂ DE DOCTORAT
REZUMAT

Matei-Șerban MIHALACHE

TEHNICI DE ANALIZĂ ȘI PRELUCRARE A SEMNALULUI VOCAL PENTRU RECUNOAȘTEREA AUTOMATĂ A ELEMENTELOR PARALINGVISTICE, CU APLICAȚII ÎN EXPERTIZA CRIMINALISTICĂ A VORBIRII

SPEECH SIGNAL ANALYSIS AND PROCESSING TECHNIQUES FOR AUTOMATIC RECOGNITION OF PARALINGUISTIC ELEMENTS, WITH APPLICATIONS IN FORENSIC SPEECH

COMISIA DE DOCTORAT

Prof. dr. ing. Gheorghe BREZEANU Universitatea POLITEHNICA din București	Președinte
Prof. dr. ing. Dragoș BURILEANU Universitatea POLITEHNICA din București	Conducător de doctorat
Prof. dr. ing. Daniela TĂRNICERIU Universitatea Tehnică "Gh. Asachi" din Iași	Referent
Prof. dr. ing. Corneliu RUSU Universitatea Tehnică din Cluj-Napoca	Referent
Prof. dr. ing. Constantin PALEOLOGU Universitatea POLITEHNICA din București	Referent

BUCHAREST 2023

Cuprins

Capitolul 1

Introducere

În cadrul acestui proiect, modelele existente de învățare automată (*machine learning*, ML) și învățare profundă (*deep learning*, DL) sunt extinse, și modele și tehnici noi sunt propuse, dezvoltate și validate în contextul analizei și procesării semnalului vocal pentru recunoașterea automată a elementelor paralingvistice, cu aplicații în expertiza criminalistică a vorbirii.

1.1. Elemente paralingvistice și sarcini de recunoaștere

Conceptul de paralingvistică a fost prima oară formulat de lingvistul american George Trager, referindu-se la meta-informațiile prezente în comunicarea orală, nuanțele transmise dincolo de conținutul lexical / semantic, cu privire la dimensiunile afective [Bac95, Bac99] sau alte manifestări psihologice [Laz99, Vil12].

Cele mai importante și fundamentale elemente paralingvistice și sarcinile de recunoaștere automată corespondente pentru prelucrarea semnalului vocal sunt definite după cum urmează [Laz99, Mat09]:

- stres = o stare psihică și fiziologică prelungită de excitație, cu impact negativ asupra stării de spirit a subiectului
  $\Longrightarrow$ detecția stresului (*speech stress detection*, SSD);
- comportament nesincer = acțiuni precum ascunderea faptelor, transmiterea unor informații incomplete sau false (i.e. minciună) etc. cu scopul unui câștig individual pentru subiect, în general în defavoarea altor persoane
  $\Longrightarrow$ detecția minciunilor (*deceptive speech detection*, DSD);
- emoții = răspunsuri tranzitorii neuro-fiziologice la stimuli, determinând reacții coordonate fizice și mentale ce determină atitudinea subiectului
  $\Longrightarrow$ recunoașterea emoțiilor (*speech emotion recognition*, SER).

În acest context, principala dificultate provine din natura subiectivă a evaluării conținutului paralingvistic, întrucât exprimarea acestuia este puternic personală și într-un context afectiv diferit de cel al evaluatorului. Pentru recunoașterea automată, însă, sistemele trebuie antrenate folosind nu doar adnotări precise ale conținutului paralingvistic, ci și acesta din urmă necesită autenticitate. Interacțiunile înregistrate ar trebui să fie spontane, naturale, neghidate, nerestricționate și, pe cât posibil, variate.

Desigur, din motive practice, arareori este posibil acest lucru. Dezvoltarea unui set de date pentru paralingvistică este o sarcină dificilă în sine, iar asigurarea respectării tuturor criteriilor menționate anterior cu atât mai mult, fie din cauza absenței datelor, fie din cauza efortului disproporționat pentru a dezvolta un astfel de set de date. Astfel, cel mai adesea sunt angajați actori pentru a înregistra interacțiuni în care manifestarea emoțională sau a altor elemente paralingvistice este mimată pe cât de bine cu putință. Dar această natură de simulacru și faptul că textele rostite sunt în general scrise anterior și repetate conduce la robustețe și capacitate de generalizare scăzute pentru sistemele dezvoltate, întrucât există diferențe semnificative între datele disponibile la momentul antrenării și cele întâlnite în scenarii reale.

1.2. Aplicații de paralingvistică în expertiza criminalistică a vorbirii

Un domeniu pentru care se pretează direct recunoașterea automată a elementelor paralingvistice din vorbire este expertiza criminalistică sau, mai general, acțiuni ale autorităților de aplicare a legii. Prin natura acestora, atenția trebuie aplecată asupra detectării emoțiilor negative, a manifestărilor unui nivel înalt de stres, și mai ales a comportamentului nesincer și a minciunilor. Pentru diverse aplicații, determinarea evoluției conținutului afectiv este, de asemenea, foarte relevantă, în special când sunt luate în considerare acțiuni complexe, pe termen lung.

Câteva exemple de astfel de aplicații ar fi:

- investigații sub forma interviurilor, interogatoriilor sau preluării mărturiilor de către agenți de poliție din partea persoanelor de interes în cazuri criminale, suspecți, martori, victime;
- anticiparea și prevenția acțiunilor criminale sau teroriste prin monitorizarea și recunoașterea comportamentului suspect, în special prin evoluția manifestărilor afective pe intervale de timp lungi;
- monitorizarea comportamentului suspect în aeroporturi, zone de atracție turistică sau aglomerate, sau alte puncte de interes;
- sisteme pentru detectarea minciunilor folosind doar date audio; etc.

Trebuie subliniat faptul că, pentru toate aceste exemple, și relativ la orice aplicații de expertiză criminalistică, abordarea acestei teze nu implică automatizarea completă a acestor sarcini, eliminând elementul uman din ecuație. Un aspect-cheie din punct de vedere etic și deontologic este ca deciziile finale și acțiunile întreprinse în urma unor astfel de analize să fie luate de agenți umani, sistemele de inteligență artificială (AI) având doar rol de unelte eficace pentru a oferi date și informații contextuale.

Din cercetările anterioare, este clar că performanța modelelor ML/DL pentru sarcinile SSD, DSD și SER crește semnificativ atunci când se folosesc date multimodale (i.e., înregistrări audio-video, date fiziologice etc.). Totuși, rămân de interes științific abordările folosind doar prelucrarea semnalului vocal, prezentând un număr de avantaje: posibilitatea de a înregistra discret vorbirea, fără cunoștința subiectului, reducând șansele ca acesta să poată manipula datele, sau de a aborda situații în care doar date audio sunt disponibile (ex.: convorbiri telefonice).

1.3. Un domeniu de studiu interdisciplinar: scop și obiective

Scopul acestui doctorat poate fi înțeles din perspectiva:

- Dezvoltarea unor noi modele și tehnici de ML/DL performante și pentru detecție automată a stresului din vorbire (SSD), a minciunilor (DSD) și recunoașterea emoțiilor (SER), cu accent pus pe manifestări afective negative, de mare intensitate, și comportament nesincer.
- Dezvoltarea de seturi de trăsături extinse și robuste pentru semnalul vocal (i.e., parametri-cheie matematici și/sau mărimi fizice) relevante pentru recunoașterea automată a elementelor paralingvistice din vorbire.
- Dezvoltarea de noi seturi de date pentru recunoașterea elementelor paralingvistice, de calitate înaltă, în condiții realistice, care să depășească limitările și dezavantajele altor seturi de date publice.
- Determinarea unor evoluții pe termen lung ale emoțiilor pentru aplicații de expertiză criminalistică, și dezvoltarea de modele afective pentru studierea acestora în relație cu monitorizarea comportamentului suspect.

1.4. Structura tezei

Această teză este structurată în opt capitole, având următorul conținut:

Capitolul 1 are rolul de introducere în conceptele generale și aspectele particulare ale elementelor paralingvistice, ale sarcinilor de recunoaștere (automată) corespondente, și a modului în care acestea pot fi aplicate domeniului de expertiză criminalistică. De asemenea, sunt definite scopul, obiectivele și structura tezei.

Capitolul 2 reprezintă o sinteză detaliată a principalelor noțiuni teoretice necesare elaborării acestei lucrări, acoperind domeniile de analiză și prelucrare a semnalului vocal și învățare automată și profundă. Capitolul prezintă un set extins de trăsături de semnal vocal extrase algoritmic, utilizat cu succes pentru sarcinile de paralingvistică relevante acestei teze, modelele ML/DL folosite în dezvoltarea sistemelor, și metodologiile de antrenare și testare abordate pentru a asigura validarea corectă a performanțelor sistemelor propuse.

Capitolul 3 debutează cu o discuție mai detaliată a elementelor paralingvistice care sunt ținta acestei lucrări (stresul psihic, emoțiile, și comportamentul nesincer) și

relația dintre aceștia. Capitolul detaliază în continuare sarcina de detecție a stresului din vorbire, începând cu sinteza cercetătorilor anterioare pe domeniu, apoi prezentarea arhitecturilor sistemelor propuse, a seturilor de date și a metodologiei și rezultatelor experimentale, precum și analiza acestora din urmă.

Capitolul 4 este primul din două capitole ce tratează sarcina de detecție a minciunilor. Capitolul începe cu o descriere a provocărilor și criteriilor implicate în dezvoltarea unor seturi de date realiste și de calitate înaltă pentru sarcini de paralingvistică, în particular detecția minciunilor, dar și o trecere în revistă a seturilor de date disponibile public pentru această sarcină. A doua jumătate a capitolului descrie în detaliu noul set de date Romanian Deva Criminal Investigation Audio Recordings (RODeCAR), dezvoltat ca parte integrantă a acestui doctorat, argumentând îmbunătățările pe care le aduce față de alte seturi similare.

Capitolul 5 extinde subiectul detecției minciunilor, urmărind aceeași structură ca a Capitolului 3, începând cu o privire de ansamblu asupra altor abordări raportate în literatură, arhitecturile sistemelor propuse, și metodologia și rezultatele experimentale obținute, inclusiv primele rezultate pentru setul RODeCAR introdus în Capitolul 4. Sistemele dezvoltate în cadrul acestei lucrări pentru detecția minciunilor implică un modul pentru detecția vorbirii, care este de asemenea descris de-a lungul acestui capitol, alături de arhitectura subsistemului și validarea experimentală.

Capitolul 6 prezintă principalele dezvoltări realizate de-a lungul doctoratului pentru sarcina de recunoaștere a emoțiilor din vorbire, acoperind trei tipuri de sisteme: abordări directe folosind trăsături extrase algoritmic și clasificatori pe bază de rețele neurale; abordări multidomeniu, în cadrul cărora se încearcă utilizarea modelelor dimensionale pentru a stabili o legătură între spațiul afectiv continuu și emoții privite drept categorii discrete, antrenând sistemul pentru a rezolva simultan o problemă de clasificare și una de regresie; și abordări bazate pe învățare prin transfer, pentru care modele de rețele neurale performante, de mare dimensiune, dezvoltate anterior pentru recunoașterea obiectelor în imagini, sunt adaptate și reantrenate astfel încât să recunoască evoluții relevante în spectrograme vocale. Capitolul urmează aceeași structură aleasă pentru Capitolul 3 și Capitolul 5: sinteza cercetărilor anterioare, arhitecturile sistemelor propuse, metodologia experimentală, rezultate și analiza acestora.

Capitolul 7 este o continuare a capitolului anterior, prezentând noțiunile teoretice și validarea experimentală pentru problema remanenței emoțiilor în vorbire, i.e. cele două ipoteze conexe conform cărora, pe măsură ce se apropie un eveniment solicitant din punct de vedere emoțional, subiecții vor manifesta reacții afective negative mai puternice care vor fi prezente (și detectabile) în vorbire pentru intervale de timp mai lungi după declanșarea lor.

Capitolul 8 este rezervat pentru concluzii, oferind un rezumat al dezvoltărilor realizate și a celor mai bune rezultate obținute în cadrul acestei lucrări, precum și o prezentare a contribuțiilor originale ale candidatului, lista articolelor și lucrărilor de conferință publicate în cadrul acestui doctorat, și perspective pentru dezvoltări ulterioare în domeniile de învățare automată și profundă și prelucrare a semnalului vocal, cu accent pus pe recunoașterea elementelor paralingvistice pentru expertiză criminalistică.

Capitolul 2

Noțiuni teoretice: analiza semnalului vocal, învățare automată

Acest capitol rezumă cunoștințele teoretice necesare pentru dezvoltarea sistemelor prezentate în cadrul acestei teze.

2.1. Un set extins de trăsături pentru procesarea semnalului vocal

Semnalul audio de intrare trece prin etapele de preprocesare (reeșantionare, normalizare, filtrare), semnalul original și varianta împărțită în cadre fiind ulterior desemnate ca vector *audio* și *vector cadru audio*. În această lucrare, împărțirea în cadre a implicat utilizarea ferestrelor Hamming cu durata de 25 ms și o suprapunere de 15 ms (60%).

Trăsăturile extrase în etapa următoare formează un set care extinde setul ComParE [Sch14] prin includerea altor trăsături care s-au dovedit relevante pentru sarcinile paralingvistice în literatura recentă, cât și în experimentele preliminare efectuate pentru această lucrare. Vectorul audio este segmentat, iar trăsăturile la nivel de segment (SWF) sunt extrase. Celelalte trăsături sunt extrase la nivel de cadru: (i) trăsături în domeniul timp (TDF); (ii) trăsături în domeniul frecvență (FDF); (iii) primii 13 coeficienți Mel-cepstrali (MFCC); și (iv) trăsături bazate pe micromodulații (MBF), propuse în [Cha14]. Coeficienții delta și delta-delta sunt calculați pentru MFCC, TDF și FDF, precum și pentru unele dintre MBF. Împreună cu SWF, acestea reprezintă trăsăturile la nivel de unitate de timp (TSF), asupra cărora se aplică setul de funcții F(·) (medie și deviație standard), rezultând trăsăturile la nivel de rostire (UWF). În cele din urmă, se aplică funcția N(·), normalizare de tip z-score, pentru fiecare vorbitor. Vectorul de trăsături normalizate (FV) obținut are o dimensiune totală de 2.260 și este utilizat ca

atare sau doar cu subseturi ale sale ca date de intrare pentru multe dintre sistemele dezvoltate în această lucrare. Etapele sunt ilustrate în Figura 2.2.

SWF (frecvența fundamentală, raportul armonice-zgomot, variația locală a frecvenței fundamentale, variația locală a valorii de vârf a semnalului vocal) sunt extrase cu ajutorul implementărilor Python ale algoritmului Yet Another Algorithm for Pitch Tracking (YAAPT) și Praat. TDF luate în considerare sunt energia RMS, care oferă o măsură a intensității fiecărui cadru al semnalului de vorbire, și rata de trecere prin zero (ZCR), care servește drept măsură a conținutului de înaltă frecvență al semnalului, valorile mai mari corespunzând, de obicei, regiunilor nesonore ale vorbirii. FDF luate în considerare sunt: energia de joasă frecvență (în subbanda 250 – 650 Hz), energia de înaltă frecvență (în subbanda 1 – 4 kHz) și, pentru mai multe subbenzi de frecvență distribuite conform scării Mel, centroizii spectrali, varianța spectrală, asimetria spectrală, entropia spectrală, fluxul spectral, panta spectrală și benzile de concentrare spectrală (calculate pentru pragurile de 25%, 50%, 75% și 90%). MBF sunt obținute prin tratarea semnalului vocal drept o serie de micromodulații în amplitudine și frecvență (AM-FM) și calcularea trăsăturilor pe baza amplitudinii și frecvenței instantanee estimate la fiecare unitate de timp prin demodularea semnalului. MBF surprind fenomenele neliniare și variabile în timp de producere a vorbirii, inclusiv structura formanților. MFCC sunt unele dintre caracteristicile eficiente și des utilizate pentru aplicațiile de analiză și prelucrare a vorbirii. Ele pot fi interpretate ca un model comprimat al tractului vocal, oferind o descriere a răspunsului în frecvență al tractului vocal în modelul sursă-filtru al vorbirii. Coeficienții delta, $\Delta(\cdot)$ ai unei trăsături F sunt măsuri ale variației sale locale, de la un cadru la altul. Astfel, ei pot fi interpretați ca o descriere a variației de ordinul întâi (în timp). Prin aplicarea funcției Δ coeficienților delta, se pot calcula coeficienții delta-delta.

2.2. Modele de învățare automată și profundă utilizate

Un sistem de învățare automată (ML) poate fi considerat un sistem adaptiv în care operațiile necesare pentru a stabili legătura dintre datele de intrare și cele de ieșire nu sunt determinate de un agent uman, ci sunt rezultatul convergenței automate către o soluție optimă numeric.

Modelul celor K medii (KMM) [Bis06] este unul dintre cei mai simpli, dar eficienți algoritmi de clasificare nesupervizată și poate fi interpretat drept un caz particular al algoritmului Baum-Welch. Ideea din spatele modelelor cu mixturi de Gaussiene (GMM) este de a modela distribuția de probabilitate a datelor ca o suprapunere (un amestec) de distribuții normale (Gaussiene) [Bis06].

Pentru modelul SVM, presupunând că datele de intrare sunt liniar separabile, adică există un hiperplan care separă perfect instanțele aparținând fiecăreia dintre cele două clase, hiperplanul este ales astfel încât marja (distanța minimă dintre limita de decizie și punctele cele mai apropiate de limita de decizie, adică *vectorii suport*) să fie maximizată. Modelul inițial a fost construit astfel încât să includă o transformare neliniară $\Phi(\cdot)$, aplicată spațiul trăsăturilor, deoarece datele de intrare adesea nu sunt

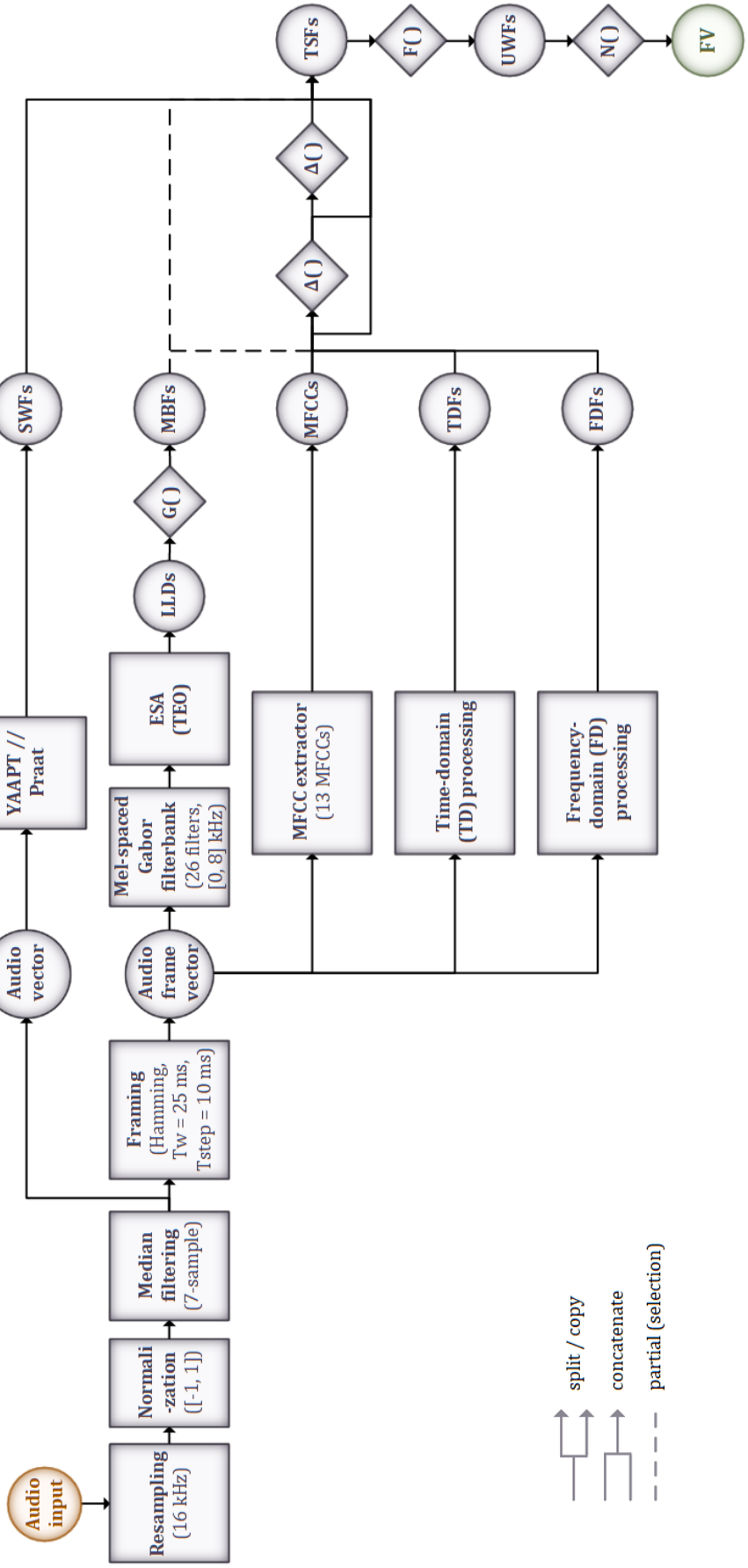

Figura 2.2 – Schema bloc detaliată a etapelor de preprocesare și extragere de trăsături.

separabile liniar în spațiul inițial al trăsăturilor, dar pot separabile fi într-un spațiu mai mare dimensional, corespondența dintre cele două fiind dată de $\Phi(\cdot)$. Calculul direct al lui $\Phi(\cdot)$ este prohibitiv de costisitor. În schimb, o funcție nucleu (kernel) permite utilizarea indirectă a lui $\Phi(\cdot)$ ca un simplu produs scalar în spațiul inițial al trăsăturilor, reducând considerabil complexitatea de calcul. Pentru probleme de clasificare în K clase, se utilizează mai multe SVM-uri în ansamblu prin: strategia unul-vs.-restul (OvR), care presupune antrenarea a K SVM-uri separat, câte unul pentru fiecare clasă, prin gruparea mai întâi a tuturor instanțelor care nu fac parte din clasa „pozitivă" curentă într-o clasă „negativă"; sau strategia unul-vs.-unul (OvO), prin antrenarea separată a câte un clasificator pentru fiecare dintre cele $K\cdot(K–1)/2$ combinații de clase.

Elementul de bază al unei rețele neurale complet conectate (FCNN) este *neuronul*, care modelează celula biologică echivalentă din sistemul nervos uman. Neuronul artificial ia ca intrare variabilele provenind de la neuronii anteriori cu ponderile asociate, stimulul total (*activarea*), reprezentând combinația liniară a intrărilor, peste care se aplică o *funcție de activare* neliniară. Un singur neuron nu ar oferi o putere de procesare relevantă, așa că mai mulți sunt organizați împreună în *straturi*. Principala abordare euristică care s-a dovedit a fi din ce în ce mai performantă în dezvoltarea contemporană a domeniului este ca, în loc să se încerce extragerea informațiilor printr-o singură transformare între spațiul caracteristic al datelor de intrare și datele de ieșire, această să fie făcută în mai multe etape, adică folosind (un număr mare de) straturi ascunse, fiecare dintre acestea obținând, în mod ideal, o abstractizare de nivel superior a datelor de intrare prin transformări suprapuse.

Rețelele neurale convoluționale (CNN) reprezintă esența sistemelor moderne de inteligență artificială cu învățare profundă (DL). Ele pot fi considerate un derivat al FCNN, bazându-se pe câteva principii cheie, inclusiv principiul local, invarianța și abstractizarea profundă [Bis06, Goo16]. Aceste principii se traduc prin faptul că nu toți neuronii dintr-un strat sunt conectați la fiecare dintre neuronii din stratul consecutiv, iar fiecare strat creează o nouă reprezentare a datelor de intrare, numită *hartă de trăsături*. În mod ideal, fiecare strat ulterior ar extrage trăsături care să ofere un nivel de abstractizare din ce în ce mai ridicat pentru date.

2.3. Metodologii de antrenare și testare

Metodologia fundamentală implică împărțirea datelor în mai multe subseturi [Has06]. În mod ideal, trei: pentru antrenare, pentru validare (*dev* / *val*) și pentru evaluarea generală (testare finală; *eval* / *test*). Dimensiunea totală a setului de date ar trebui să fie cât mai mare. Multe seturi de date disponibile public sunt, însă, relativ mici, alternativa fiind aplicarea *validării încrucișate*: împărțiri repetate în subseturi de antrenare-validare. O serie de tehnici avansate au fost adoptate pentru sistemele dezvoltate în această lucrare pentru a îmbunătăți procesul de antrenare și pentru a crește performanța modelelor: regularizare, antrenare selectivă a neuronilor, normalizarea datelor pe grupuri și altele.

Pentru problemele de regresie, în afară de valoarea funcției de cost în sine, măsurile de performanță sunt coeficientul de corelație, ρ, definit în (2.91), unde σ_y și $\sigma_{\hat{y}}$

sunt deviațiile standard ale valorilor țintă și ale valorilor de ieșire ale modelului, iar $K_{y\hat{y}}$ reprezintă covarianța dintre cei doi vectori, și coeficientul de concordanță a corelației, ρ_c, definit în (2.92), unde μ_y , $\mu_{\hat{y}}$, σ_y^2 și $\sigma_{\hat{y}}^2$ sunt mediile și variațnele valorilor țintă și ale valorilor de ieșire ale modelului. Pentru sarcinile de clasificare, considerând N instanțe făcând parte din K clase, cu N_k numărul de instanțe din fiecare clasă, fie H_k numărul de predicții corecte făcute de model pentru clasa k, dat de (2.98a), unde funcția $\mathrm{h}_{(k)}(\cdot, \cdot)$ este definită în (2.98b); și F_k numărul de predicții incorecte făcute pentru clasa k, dat de (2.99a), unde funcția $\mathrm{f}_{(k)}(\cdot, \cdot)$ este definită în (2.99b). Precizia (P), senzitivitatea (R), și acuratețea neponderată și ponderată (UA / WA) se definesc apoi în (2.100) – (2.103).

$$\rho = \frac{K_{y\hat{y}}}{\sigma_y \cdot \sigma_{\hat{y}}} \tag{2.91}$$

$$\rho_c = \frac{2\rho \cdot \sigma_y \cdot \sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)^2} \tag{2.92}$$

$$H_k = \sum_{n=0}^{N-1} \mathrm{h}_{(k)}\left(y_n, \hat{y}_n\right) \tag{2.98a}$$

$$\mathrm{h}_{(k)}\left(y_n, \hat{y}_n\right) = \begin{cases} 1, & \hat{y}_n = y_n = k \\ 0, & \text{otherwise} \end{cases} \tag{2.98b}$$

$$F_k = \sum_{n=0}^{N-1} \mathrm{f}_{(k)}\left(y_n, \hat{y}_n\right) \tag{2.99a}$$

$$\mathrm{f}_{(k)}\left(y_n, \hat{y}_n\right) = \begin{cases} 1, & \hat{y}_n = k \quad \text{and} \quad y_n \neq k \\ 0, & \text{otherwise} \end{cases} \tag{2.99b}$$

$$\mathrm{P}_k = \frac{H_k}{H_k + F_k} \tag{2.100}$$

$$\mathrm{R}_k = \frac{H_k}{N_k} \tag{2.101}$$

$$\mathrm{WA} = \frac{1}{N} \sum_{k=0}^{K-1} H_k = \frac{1}{K} \sum_{k=0}^{K-1} \frac{K \cdot N_k}{N} \cdot \frac{H_k}{N_k} \tag{2.102}$$

$$\mathrm{UA} = \frac{1}{K} \sum_{k=0}^{K-1} \frac{H_k}{N_k} \tag{2.103}$$

2.4. Concluziile capitolului

În acest capitol, a fost prezentat un rezumat al principalelor cunoștințe teoretice utilizate în timpul dezvoltării acestei lucrări, inclusiv un set extins de trăsături extrase algoritmic, modele ML/DL și metodologii, tehnici și măsuri fundamentale și avansate pentru antrenare și testare, utilizate pentru a asigura performanța corespunzătoare a sistemului.

Capitolul 3

Detecția stresului din vorbire

Acest capitol se referă la sarcina de detecție a stărilor de stres psihologic din discursul unui subiect, numită și *detecția stresului din vorbire (*SSD). Părți din conținutul de față au fost publicate într-un articol de conferință de către candidat [Mih21b].

3.1. Context și sinteza cercetării existente în domeniu

Este important de înțeles că există o suprapunere conceptuală considerabilă între starea de a fi supus unui stres psihologic, externalizarea stărilor afective (emoții) și abordarea unui comportament nesincer. Acest lucru conduce la două principii fundamentale care trebuie luate în considerare pentru sarcinile paralingvistice:

1) O separare completă între aceste concepte / stări nu este nici posibilă, nici dezirabilă, deoarece multe aplicații de expertiză criminalistică (și nu numai) vor avea obiective finale axate pe comportamente sau manifestări de nivel superior (de exemplu, monitorizarea comportamentului suspect), în care toate aceste elemente sunt relevante atât distinct, cât și holistic.
2) Fiecare concept / stare este relevant și nu poate fi pur și simplu inclus în oricare dintre celelalte, deoarece multe cazuri individuale întâlnite în cazuri reale pot foarte bine să se încadreze doar într-una dintre stări sau în combinații de două dintre ele, cu doar mici manifestări ale celei de-a treia.

Printre trăsăturile care s-au dovedit a oferi rezultate promițătoare, cele spectrale și cepstrale sunt des întâlnite, cum ar fi descompunerea în spectrograme [He09] sau în wavelet-uri [Zao14] și coeficienții Mel-cepstrali (MFCC) [Cas06, Li07], precum și alte trăsături acustice, de exemplu, frecvența fundamentală [Cas06] și variația acesteia și a valorii de vârf a semnalului vocal [Li07]. S-a demonstrat, de asemenea, că valorificarea seturilor extinse de trăsături îmbunătățește adesea acuratețea sistemului pentru SSD.

Au fost raportate mai multe abordări care utilizează modele tradiționale de învățare automată (ML), inclusiv modele Markov ascunse (HMM) [Cas06], modele cu mixturi de Gaussiene (GMM) [He09, Zao14], mașini cu vectori suport (SVM) [Bes16] sau modele hibride HMM-GMM [Li07]. Mai recent, au fost raportate, de asemenea, soluții de învățare profundă (DL), cum ar fi rețele neurale convoluționale (CNN) și rețele neurale hibride convoluțional-recurente (CRNN) [Avi19, Shi20].

3.2. Arhitecturile sistemelor propuse

Sistemul de bază DL propus constă în utilizarea unei rețele neurale profunde (DNN) care ia ca intrare un set extins de caracteristici obținute prin aplicarea unor funcții statistice de nivel înalt asupra descriptorilor acustici, spectrali și cepstrali extrași algoritmic [Mih21b]. Clasificatorul DNN este un model de rețea neurală complet conectată (FCNN), care utilizează între 2 și 4 straturi ascunse, cu numere diferite de noduri pe strat, și un strat de ieșire a cărui dimensiune este egală cu numărul de clase luate în considerare pentru fiecare grup de experimente (4 clase, 3 clase sau clasificare binară). Au fost luate în considerare două structuri pentru straturile ascunse: arhitectura „constant”, care constă în selectarea aceluiași număr de noduri pentru fiecare strat ascuns; și arhitectura „log2dec”, care constă în selectarea unui număr progresiv mai mic de noduri pentru fiecare strat activ, urmând o lege logaritmică în baza 2 descrescătoare.

Se propune, de asemenea, un sistem mai avansat care utilizează clasificatori în ansamblu. Acesta este prezentat în Figura 3.3. În acest scop, a fost utilizată o strategie de clasificare în ansamblu de tip „unul-vs.-unul” (OvO), inspirată de abordarea corespondentă a modelelor SVM. Un total de $K{\cdot}(K{-}1)/2$ clasificatori (unde K este numărul de clase) au fost antrenați independent pentru fiecare pereche de clase, datele lor de ieșire fiind folosite, împreună cu valorile rotunjite (predicțiile binare intermediare ale fiecărui clasificator), ca intrări pentru un DNN pentru clasificarea finală.

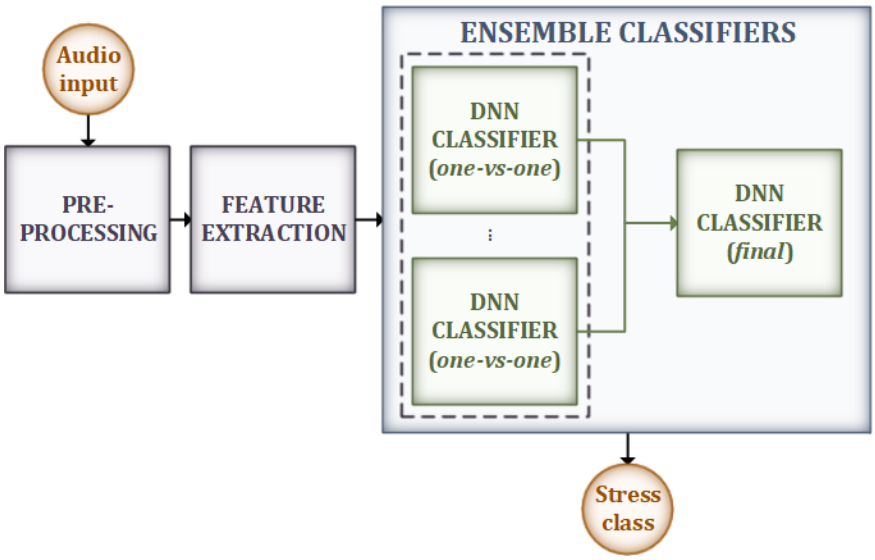

Figura 3.3 – Clasificatori în ansamblu, strategie unul-vs.-unul (OvO).

3.3. Metodologie și rezultate experimentale

Baza de date Speech Under Simulated and Actual Stress (SUSAS) [Han98] conține aproximativ 14.600 de înregistrări în limba engleză, cu o durată medie de 0,6 s. Prima jumătate a corpusului, având 9 vorbitori (toți bărbați), cuprinde înregistrări în condiții de stres simulat, rezultând 11 clase: 7 referindu-se la stilul de vorbire (Fast, Slow, Soft, Loud, Clear, Angry, Question), 3 la mediul în care erau plasați vorbitorii în momentul înregistrării (Cond50 – înregistrat în timp ce rezolvau o sarcină de dificultate medie; Cond70 – înregistrat în timp ce rezolvau o sarcină de dificultate mai mare; Lombard – înregistrat în timp ce ascultau un zgomot roz de intensitate mare care declanșează efectul Lombard) și o clasă neutră. A doua jumătate, având 7 vorbitori (3 femei, 4 bărbați), conține înregistrări realizate în condiții reale de stres, rezultând 5 clase: 2 care se referă la rezolvarea unor sarcini complexe (MeS – sarcină de dificultate medie; HiS – sarcină de dificultate mai mare), 2 care se referă la participarea la activități din parcuri de distracții (Freefall și Scream) și o clasă neutră.

Pentru corelarea cu alte lucrări prezentate în literatura de specialitate, precum și cu principala aplicație SSD vizată, și anume detecția binară a stresului psihologic din vorbire, au fost create următoarele seturi de date prin împărțirea bazei de date SUSAS:

- **Setul A** – 4 clase în condiții *reale* de stres: Scream (SCRM), HiS, MeS și Neutral (NEU); total: 3.567 de înregistrări.
- **Setul B** – 3 clase din setul A: HiS, MeS și NEU; total: 3.179 de înregistrări.
- **Setul C** – 2 clase din setul A: STRS (grupând SCRM, HiS și MeS) vs. NEU; aceeași dimensiune ca setul A.
- **Setul D** – 4 clase în condiții *simulate* de stres: Angry (ANG), Lombard (LOM), Loud (LOU) și Neutral (NEU); total: 2.518 de înregistrări.
- **Setul E** – 2 clase în condiții *simulate* de stres: STRS (grupând ANG, LOM, LOU, Cond50 și Cond70) vs. NEU; total: 7.556 de înregistrări.

Numărul de straturi ale clasificatorului DNN a fost variat între 2 și 4, cu un număr inițial de noduri (pentru primul strat ascuns) de 256 sau 128. Alți hiperparametri aleși au fost: funcția de activare ReLU pentru straturile ascunse și funcția de activare softmax pentru stratul de ieșire; și Adam ca algoritm de optimizare. Aceleași configurații au fost testate pentru clasificatorii OvO DNN, clasificatorul DNN final având o adâncime fixă egală cu 2 sau 3, precum și 6 sau 12 noduri pe strat ascuns, pentru experimentele cu 3 clase și 4 clase. Pentru experimentele de stres *real* (seturile de date A, B și C), a fost selectată o schemă de testare prin validare încrucișată cu 10 repetări, cu o împărțire a setului de date în 70% / 30% pentru antrenare-validare, rezervând 2 din 7 vorbitori (o femeie, un bărbat) pentru fiecare subset de validare. Pentru experimentele de stres *simulat* (seturile de date D și E), s-a utilizat de asemenea validarea încrucișată cu repetări, dar cu o împărțire a setului de date în 66% / 33%, rezervând 3 din 9 vorbitori. Pentru toate cazurile dezechilibrate, a fost utilizată ponderarea claselor.

O comparație a rezultatelor obținute față de alte lucrări din literatura de specialitate este realizată în Tabelul 3.9, pentru toate cazurile disponibile, și anume,

condiții de stres *real* cu 4 clase (setul A), condiții de stres *real* cu 3 clase (setul B), condiții de stres *simulat* cu 4 clase (setul D) și condiții de stres *simulat* cu 2 clase (setul E). Sistemul propus prezintă o creștere semnificativă a performanței (evidențiată în verde) pentru cazurile de stres *real* cu 4 clase (setul A), stres *simulat* cu 4 clase (setul D) și stres *simulat* cu 2 clase (setul E). Se remarcă faptul că, în cazul stresului *real* cu 3 clase (setul B), singurele rezultate raportate [He09] au fost pentru antrenarea modelelor KNN și GMM (acesta din urmă demonstrând rezultate mai bune) pe spectrograme extrase din eșantioane de vocale. Acestea au fost extrase din doar 547 de instanțe de date, față de numărul mult mai mare de 3.179 de instanțe utilizate în această lucrare. Această discrepanță mare în ceea ce privește dimensiunile subseturilor de antrenare și de validare poate explica acuratețea mai mare obținută în [He09].

3.4. Concluziile capitolului

În acest capitol, au fost propuse sisteme de învățare profundă, bazate pe utilizarea mai multor rețele neurale profunde (DNN-uri) complet conectate, legate între ele în cadrul unei configurații de strategie de clasificare în ansamblu „unul-vs.-unul" (OvO) și utilizând ca intrare un set extins de trăsături acustice, spectrale și cepstrale extrase algoritmic. Sistemele au fost testate pe baza de date SUSAS, pentru 5 subseturi de grupare a claselor (SSD cu 4 clase, 3 clase și 2 clase pentru condiții *reale* de stres; 4 clase și 2 clase pentru condiții *simulate* de stres).

S-au obținut îmbunătățiri semnificative ale performanțelor față de alte rezultate relevante de ultimă generație raportate anterior în literatura de specialitate, cu o acuratețe (neponderată/ponderată) (UA/WA) de **68,8% / 65,5%** pentru condiții de stres *real* cu 4 clase, **59,2% / 62,4%** pentru condiții de stres *real* cu 3 clase, **66,7% / 81,4%** pentru condiții de stres *real* cu 2 clase, **75,5% / 75,5%** pentru condiții de stres *simulat* cu 4 clase, și **76,1% / 78,4%** pentru condiții de stres *simulat* cu 4 clase.

Tabelul 3.9 – Comparație de performanță între cele mai bune rezultate obținute în această lucrare și alte rezultate relevante publicate în literatură.

Setul de date	**Metodă**	**Performanță**			
		P medie [%]	**F1 medie [%]**	**R medie ≡ UA [%]**	**WA [%]**
A	[Zao14] – GMM	–	–	64,0	–
	Această lucrare	69,7	68,9	**68,8**	65,5
B	[He09] – GMM	–	–	–	**73,8**
	Această lucrare	61,0	59,5	59,2	62,4
D	[Cas06] – HMM	–	–	72,9	–
	[Avi19] – CNN	–	–	–	71,0
	Această lucrare	75,6	75,3	**75,5**	**75,5**
E	[Avi19] – CNN	–	–	–	76,0
	Această lucrare	79,5	76,8	76,1	**78,4**

Capitolul 4

RODeCAR: Un set de date nou pentru detecția minciunilor din vorbire

Acest capitol se referă la dezvoltarea completă a unui set de date nou, obiectiv și de înaltă calitate pentru *detectarea minciunilor din vorbire (*DSD): setul de date Romanian Deva Criminal Investigation Audio Recordings (RODeCAR). Părți din conținutul de față au fost publicate într-un articol de conferință de către candidat [Mih19b].

4.1. Context și sinteza cercetării existente în domeniu

Rezultatele unui test poligraf nu sunt admisibile ca probe științifice într-o instanță de judecată. În cel mai bun caz, acestea pot ghida un agent să urmărească o anumită pistă sau o linie de interogare. Deși nici un sistem de detecție a minciunilor din vorbire nu ar avea valoare juridică, rezultatele s-ar putea dovedi de o acuratețe mai mare, deoarece unele trăsături ale vorbirii sunt, în general, mai greu de modificat în mod voluntar decât parametrii urmăriți de testele poligraf convenționale sau de metodele mai noi de analiză psihologică, care pot fi manipulate pentru a produce rezultate false [Ver09]. Mai mult, chiar dacă o abordare multimodală ar conduce la performanțe mai mari, un detector de minciuni doar pe bază audio poate fi utilizat discret pentru majoritatea scenariilor relevante, reducând gradul de conștientizare pentru subiect, minimizând șansele de manipulare a rezultatelor testului.

Pentru a dezvolta sisteme de învățare automată capabile să detecteze minciunile, sunt necesare seturi mari de date cu adnotări precise și metodice. Natura conținutului audio ar trebui să fie cât mai autentică. Datele simulate pot afecta capacitatea sistemului de a generaliza pentru condiții reale [Mor12]. Pentru un context sensibil precum detecția minciunilor, necesitatea de a dispune de date din situații reale este cu atât mai mare.

Principalele probleme evidențiate pentru seturile de date pentru DSD disponibile public sunt utilizarea actorilor sau a participanților antrenați anterior (comportament simulat); abordarea unui obiectiv specific într-un mediu familiar sau relaxat (scenariu simulat); abordarea unor sarcini cu miză redusă (stimulare redusă); și utilizarea unor metode de autoevaluare sau subiective pentru adnotarea datelor (adnotare subiectivă) [Mih19b]. Pentru a depăși aceste dezavantaje semnificative comune, în această lucrare se propune o abordare diferită:

1) Participanții nu trebuie să fi fost îndrumați asupra comportamentului așteptat și trebuie să aibă control total asupra conținutului răspunsurilor.
2) Nu ar trebui creat un scenariu specific; în schimb, în limitele rezonabilului, ar trebui să se utilizeze un cadru liber pentru a reduce previzibilitatea discursului participanților și/sau a liniei de intervievare (dacă este cazul).
3) Participanții ar trebui să fie conștienți de consecințe relativ severe (cu miză mare) atât pentru implicarea într-un comportament nesincer, cât și pentru recunoașterea unor adevăruri incriminatoare.
4) Adnotarea datelor ar trebui să fie efectuată de un expert, după efectuarea unei cercetări ulterioare pentru a determina în mod obiectiv și clar veridicitatea declarațiilor participanților.
5) În cazul în care incertitudinea privind etichetarea este minimizată, dar nu eliminată, ar trebui utilizat un scor de încredere asociat fiecărei interacțiuni.

4.2. Setul de date Romanian Deva Criminal Investigation Audio Recordings

Pentru a răspunde primelor trei cerințe prezentate anterior, se poate argumenta că una dintre cele mai bune surse de material ar fi înregistrările unor activități reale de investigare a aplicării legii, în care participanții sunt suspecți, martori etc., iar contextul este dat de audieri și interviuri efectuate de agenți profesioniști instruiți în domeniul aplicării legii. Astfel, participanții au foarte puține cunoștințe prealabile cu privire la posibila linie de interogare și există un mare stimulent atât pentru cei onești să prezinte în mod convingător o relatare veridică, cât și pentru suspecți să ascundă în mod convingător fapte incriminatoare. Natura sensibilă și adesea confidențială a unor astfel de înregistrări este un factor de descurajare, dar, cu toate acestea, setul de date RODeCAR a fost construit folosind fișiere care acoperă 9 cazuri criminale rezolvate, în timpul cărora s-au desfășurat anchete pentru omor, agresiune sexuală și fraudă. A patra cerință a fost îndeplinită prin efectuarea unei analize meticuloase a înregistrărilor și a însemnărilor asociate împreună cu procurorul care investigase inițial cazurile, pentru a determina veridicitatea conținutului. În sfârșit, incertitudinea inevitabilă cu privire la detalii sau care rezultă din informații indisponibile cu privire la investigații a fost abordată asociind un nivel de încredere între 70% la 100% fiecărei interacțiuni (fișier).

După o filtrare inițială a conținutului, toți cei 20 de vorbitori implicați au fost identificați manual și asociați cu un număr de identificare, cu o valoare specială rezervată procurorului (care nu a fost luată în considerare pentru conținutul final al

setului de date). De asemenea, s-a ținut cont de genul acestora, având 4 femei și 16 bărbați. Piesele audio au fost apoi extrase cu ajutorul FFmpeg și salvate în format PCM pe 16 biți, la o frecvență de eșantionare de 16 kHz.

Apoi, cele 26 de înregistrări audio prelucrate (în total, aproximativ 7,5 ore de material) au fost clasificate în trei categorii distincte, în funcție de tipul de conținut și de modul în care sunt implicați participanții:

- Interogatorii (Q): interogatoriile participanților de către procuror într-un mediu formal și în urma unei proceduri stricte; acesta este cel mai stresant scenariu pentru participanți.
- Interviuri (I): interacțiuni între procuror și participant într-un mediu informal (adesea mai familiar pentru subiect); acesta este cel mai puțin stresant scenariu pentru participanți.
- Mărturii (M): relatări / mărturisiri neîntrerupte, în formă liberă, făcute de participanți, adesea ca urmare a unui interogatoriu anterior.

Fiecare fișier a fost segmentat semiautomat. Un segment este definit ca fiind o porțiune de vorbire de la un singur vorbitor, fie (i) separată printr-o pauză de cel puțin 200 ms de alte porțiuni de vorbire de la același vorbitor; fie (ii) separată de alte porțiuni de vorbire de la un vorbitor diferit, indiferent de durata pauzei inițiale.

Adnotarea binară (adevărat, T / neadevărat, UT) s-a făcut pentru fiecare segment, dar într-un sens global; de exemplu, un segment scurt care conține informații corecte faptic, aflat în cadrul unui segment mai lung al unui vorbitor care are un comportament nesincer, este etichetat ca neadevărat. Argumentul este că starea de spirit în care se află participanții atunci când mint va fi menținută de obiectivul pe termen lung de a înșela procurorul, indiciile fiind încă prezente în discursul participanților.

Setul de date, în forma sa completă și publică, este format din 4 ore și 46 de minute total, obținut de la 20 de vorbitori (4 femei, 16 bărbați) în timpul unor mărturii, interviuri și interogatorii efectuate de organele de aplicare a legii din România, în care toți participanții au fost persoane de interes (vinovați, suspecți, martori etc.). Din durata totală, 3 ore și 28 de minute reprezintă segmentele participanților; 2 ore și 6 minute (60,5%) reprezintă conținutul adevărat, iar 1 oră și 22 de minute (39,5%) reprezintă conținutul neadevărat. Setul de date RODeCAR este disponibil la cerere și poate fi accesat aici: https://speed.pub.ro/downloads/paralinguistic-datasets/.

4.3. Concluziile capitolului

În acest capitol, a fost prezentat setul de date Romanian Deva Criminal Investigation Audio Recordings (RODeCAR): un set de date ce conține rostiri adevărate și neadevărate, construit prin analiza, procesarea și examinarea înregistrărilor originale arhivate în urma unor investigații criminale. Cel mai important avantaj de valorificat prin utilizarea acestui set de date este natura obiectivă a conținutului: toți vorbitorii au fost suspecți sau martori în cazuri criminale reale, iar toate interacțiunile au fost spontane și au făcut parte din contexte autentice de aplicare a legii.

Capitolul 5

Detecția minciunilor din vorbire

Acest capitol se referă la sarcina de *detecție a minciunilor din vorbire* (DSD). Părți din conținutul de față au fost publicate de candidat într-o lucrare de conferință [Mih21a] și un articol de jurnal [Mih22a].

5.1. Context și sinteza cercetării existente în domeniu

Pentru DSD, cercetările anterioare au fost efectuate utilizând trăsături și descriptori ai semnalului vocal extrași algoritmic, inclusiv media și deviația standard a frecvenței fundamentale [Sen22], MFCC-uri și coeficienții delta și delta-delta ai acestora [Fat21a, Men17], raportul armonice-zgomot (HNR) [Jai16] sau alte trăsături acustice și prozodice [Men17, Vel19] bazate pe setul ComParE. În ceea ce privește modelele ML și DL utilizate, acestea includ mașini cu vectori de suport (SVM) [Jai16, Men17, Mon16, Sen22], ansambluri de arbori de decizie (RF) [Men17, Sen22, Vel19, Zha20], FCNN [Kop19, Men17, Sen22, Vel19], regresie logistică [Kop19, Men17] sau metode în ansamblu cu clasificatori multipli și vot mediu/majoritar [Vel19].

5.2. Detecția vorbirii ca subsarcină

Una dintre primele sarcini care trebuie abordate în cadrul unui lanț de procesare este detecția vorbirii (VAD) [Mih21a], abordată în această lucrare ca o sarcină secundară pentru DSD pentru extragerea unor trăsături prozodice. În această lucrare, *rostirile* sunt definite ca fiind conținutul intervalelor unui semnal audio în care este prezentă vorbirea, separate de alte astfel de intervale de vorbire prin pauze cu o durată de cel puțin 200 ms.

Modelele de rețele neurale profunde (DNN) analizate sunt FCNN, RNN bazate pe LSTM și CNN, împreună cu trei tehnici optimizate de postprocesare: praguri cu

histerezis, filtrarea duratei minime și extinderea bilaterală. Subsistemele VAD propuse bazate pe FCNN utilizează două straturi ascunse și folosesc trăsături extrase algoritmic, și anume energia, rata de trecere prin zero (ZCR), HNR, coeficientul de autocorelație normalizat și primele 13 MFCC, care sunt grupate în mai multe subseturi de trăsături. Subsistemele bazate pe RNN utilizează un strat LSTM, urmat de straturi complet conectate pentru clasificarea propriu-zisă, și au ca intrare aceleași seturi de trăsături. Subsistemele bazate pe CNN implementează trei perechi de straturi convoluționale și de agregare, urmate de straturi complet conectate pentru clasificare. Eșantioanele semnalului în domeniul timp sau reprezentarea sa în domeniul frecvență, dată de transformarea Fourier discretă (DFT) calculată în 127 de puncte, sunt furnizate către CNN. Toate trăsăturile sunt calculate la nivel de cadru. Ieșirea modelului va reprezenta probabilitatea per cadru să fie prezentă vorbirea. O fereastră glisantă care cuprinde 3 cadre consecutive este utilizată pentru a calcula media acestor probabilități, iar valoarea rezultantă este comparată cu un prag. Dacă valoarea este peste prag, fereastra va fi considerată *pozitivă* (conține vorbire). Ferestrele *pozitive* consecutive determină momentele de început și de sfârșit ale rostirii. Pentru a crește performanța, se utilizează histerezisul pentru a obține două praguri separate, cel superior fiind implicat pentru tranziția de la predicții negative la pozitive. În plus, în cazul în care o rostire ar avea o durată mai mică decât valoarea de referință Δt_{min}, este eliminată. Pentru rostirile rămase, se implementează o extindere bilaterală a duratei acestora pentru a compensa tendința subsistemului de a subestima lungimea lor. Extinderea constă în scăderea momentul de timp de început cu o valoare Δt_{ext} , în timp ce momentul final crește cu aceeași valoare.

Testarea finală a fost efectuată pe subsetul cu zgomot ambiental real din Corpus and Environment for Noisy Speech Recognition (CENSREC-1-C) [Kit07], rezultatele detaliate fiind prezentate în Tabelul 5.3, în comparație cu alte lucrări. Subsistemul VAD a fost adaptat ulterior pentru seturile de date DSD.

Tabelul 5.3 – Subsistemul VAD: acuratețea la nivel de rostire [%] în funcție de tipul de model, setul de date CENSREC-1-C.

Model	**Tipul zgomotului ambiental**				
	Restaurant		**Autostradă**		**Medie**
	SNR crescut	**SNR scăzut**	**SNR crescut**	**SNR scăzut**	
CENSREC-1-C (standard) [Kit07]	74,20	56,50	39,40	41,40	52,88
[Esp11]	76,75	63,02	92,44	79,64	77,96
[Fuj10]	92,75	65,51	**100,00**	**100,00**	**89,57**
[Fuj14]	75,65	21,45	95,94	49,86	60,73
FCNN	88,11	57,46	56,65	54,34	64,14
RNN	74,25	39,10	65,80	51,50	57,66
CNN-DFT1	85,21	64,92	82,60	75,65	77,10
CNN-DFT2	97,10	59,13	95,36	89,56	85,29
CNN-DFT3	**99,13**	**68,69**	**97,97**	**90,72**	**89,13**

5.3. Arhitecturile sistemelor propuse

Sistemul de bază de învățare profundă propus pentru sarcina principală, detecția minciunilor din vorbire (DSD), constă într-un DNN care primește ca intrare o versiune modificată a setului extins de 2.258 trăsături de semnal vocal descris în Capitolul 2. Clasificatorul DNN este un model FCNN, care utilizează între 2 și 3 straturi ascunse cu 64 sau 128 de noduri per strat, dar cu un strat de ieșire de dimensiune egală cu numărul de clase luate în considerare, i.e. 2 (*adevărat* și *neadevărat)*, și care utilizează funcția de activare softmax în loc de un singur neuron care aplică funcția de activare sigmoidă.

În ceea ce privește modificarea setului de trăsături, aceasta se referă la includerea suplimentară a două trăsături prozodice (UPF) [Mih22a]: durata rostirii și durata pauzei inițiale, adică intervalul de timp dintre sfârșitul rostirii anterioare și începutul celei curente, ambele dovedite a fi relevante pentru sarcina DSD.

Este propus un al doilea sistem DSD, care valorifică natura extragerii automate a trăsăturilor oferite de CNN-uri. Modelul primește la intrare spectrograma fiecărei rostiri preprocesate, extrasă folosind ferestre Hamming cu durata de 25 ms și suprapunere de 15 ms, împărțite liniar în 257 de benzi de frecvență (corespunzător jumătății frecvenței de eșantionare). Ulterior, se aplică trei rânduri de straturi convoluționale 2D cu un câmp receptiv mic, cu straturi de agregare aplicate după fiecare pentru a reduce dimensiunea datelor. Ieșirea stratului final de agregare reprezintă setul de hărți de trăsături extrase automat, care sunt apoi concatenate într-un vector 1D și trecute printr-o secvență de straturi ascunse complet conectate. Împreună cu stratul de ieșire, acestea formează secțiunea propriu-zis de clasificare și adoptă o configurație (număr de straturi, număr de noduri per strat) conform celei mai bune structuri determinată anterior.

În cele din urmă, pentru a crește și mai mult performanța, se propune o rețea hibridă CNN-MLP (secțiunea MLP fiind reprezentată de setul final de straturi complet conectate de după stratul de concatenare), combinând trăsăturile extrase automat furnizate de straturile convoluționale cu cel mai bun subset de trăsături extrase algoritmic, determinat cu ajutorul sistemului DSD de bază bazat pe DNN. Aceste caracteristici, extrase la nivelul rostirilor așa cum s-a descris anterior, sunt furnizate ca intrare suplimentară și sunt concatenate cu trăsăturile extrase automat înainte de a fi introduse în secțiunea de clasificare (complet conectată) a rețelei hibride.

5.4. Metodologie și rezultate experimentale

Setul de date Real-Life Trial Data for Deception Detection (RLDD) [Per15] cuprinde 121 de înregistrări audio-video în limba engleză ale inculpaților și martorilor, obținute din procese desfășurate în Statele Unite, 61 de înregistrări fiind etichetate ca *neadevărate* și 60 ca *adevărate*. Durata totală este de 56 min, cu o durată medie a de 28 s. Excluzând procurorii, avocații și alți intervievatori, numărul de vorbitori este de 56 (22 femei, 34 bărbați). Al doilea set de date, RODeCAR, a fost descris în Capitolul 4.

Alte cercetări publicate anterior cu privire la DSD pentru setul de date RLDD utilizează o abordare la nivel de vorbitor [Sen22] sau la nivel de înregistrare [Fat21a,

Jai16, Vel19]. Prima implică determinarea atitudinii generale a fiecărui vorbitor (56 de cazuri), în timp ce a doua clasifică fiecare înregistrare audio întreagă (121 de cazuri) ca fiind *adevărată* sau nu. În această lucrare, accentul principal este pus pe o abordare diferită, mai dificilă, la nivel de rostire, anume determinarea rostirilor *adevărate* vs. *neadevărate* din fiecare înregistrare. Pentru aceasta, din fiecare înregistrare audio din seturile de date RLDD și RODeCAR, au fost extrase toate rostirile, în total 931 (467 *adevărate* și 464 *neadevărate*) și 5.859 (3.136 *adevărate* și 2.723 *neadevărate*).

Dacă nu este specificat altminteri, pentru experimentele DSD s-a utilizat ca metodologie de testare validarea încrucișată cu 10 repetări, separând vorbitorii, cu o împărțire 80% / 20%, asigurându-se același raport de instanțe *adevărate* și *neadevărate* pentru fiecare subset, precum și același raport de vorbitori de gen masculin și feminin.

Modelele bazate pe FCNN au fost evaluate pentru o adâncime (numărul de straturi ascunse) între 2 și 3, cu 64 sau 128 de noduri per strat ascuns. Alți hiperparametri aleși au inclus: funcția de activare ReLU pentru straturile ascunse; Adam ca algoritm de optimizare; regularizare L1 cu parametrul de regularizare egal cu 10^{-4} . Deoarece setul de date RODeCAR este ușor dezechilibrat în ceea ce privește distribuția claselor (53,5% dintre enunțuri sunt *adevărate*), a fost utilizată ponderarea claselor.

Sistemele de bază au fost testate pentru setul complet de 2.260 de trăsături, descris în Secțiunea 5.3, precum și pentru mai multe subseturi de trăsături. Apoi, utilizând un nou algoritm de selecție a trăsăturilor bazat pe testul Kolmogorov-Smirnov (KS), au fost obținute încă 5 subseturi prin selectarea celor mai relevante 10, 20, 50, 100 și 200 de trăsături din punct de vedere al statisticii KS, rezultând un număr total de 36 de subseturi de trăsături. Pentru celelalte două sisteme DSD propuse, i.e. cele bazate pe CNN sau cele care utilizează o rețea hibridă CNN-MLP, spectrogramele de intrare trebuie să aibă toate aceeași dimensiune, necesitând astfel completare până la durata celei mai lungi rostiri din fiecare set de date (în număr de cadre). A fost utilizată, în schimb, regularizarea L2. Secțiunea complet conectată este formată din două straturi, cu 32 de noduri per strat. Toți ceilalți hiperparametri urmează aceeași configurație ca în cazul sistemului DSD de bază, bazat pe FCNN, descris anterior.

Pentru a compara modelul propus și celelalte sisteme DSD raportate în literatura de specialitate, rezultatele la nivel de rostire au fost postprocesate la nivel macro pentru a corespunde rezultatelor alternative la nivel de vorbitor și la nivel de înregistrare, după caz. Pentru setul de date RLDD (singurul pentru care se pot face astfel de comparații, deoarece până la data redactării acestei teze nu au fost publicate alte rezultate pentru setul de date RODeCAR de către părți externe), se parcurg următorii pași:

- în cazul abordării la nivel de vorbitor, toate rostirile care aparțin fiecărui vorbitor sunt grupate împreună, iar votul majoritar este luat asupra etichetelor la nivel de rostire prezise de modelul CNN-MLP;
- în cazul abordării la nivel de înregistrare, se efectuează un pas similar, dar pentru fiecare înregistrare în loc de fiecare vorbitor.

Pentru RLDD, rezultatele la nivel de vorbitor și de înregistrare sunt prezentate în Tabelul 5.11 și Tabelul 5.12, în funcție de acuratețea (ponderată). Pentru RODeCAR, performanța pentru adaptarea sistemului la nivel de vorbitor a fost determinată ca fiind 83,5%. Abordarea la nivel de înregistrare este inaplicabilă din cauza naturii setului de

Tabelul 5.11 – Acuratețea pe subsetul de test la nivel de vorbitor [%] comparație cu alte lucrări publicate în literatură: RLDD.

Sistem	Acuratețe [%]
[Sen22] – RF	71,2
[Sen22] – MLP	61,0
Această lucrare: CNN-MLP; intrare: spectrogramă la scară liniară de dimensiune 1,183×257 / 2,392×257 + 10 / 340 trăsături; 16 kHz eșantionare	**85,6**

Tabelul 5.12 – Acuratețea pe subsetul de test la nivel de înregistrare [%] comparație cu alte lucrări publicate în literatură: RLDD.

Sistem	Acuratețe [%]
[Fat21a] – SVM	81,5
[Vel19] – Ansamblu (KNN + RF + MLP)	70,0
[Jai16] – SVM	34,2
Această lucrare: CNN-MLP; intrare: spectrogramă la scară liniară de dimensiune 1,183×257 / 2,392×257 + 10 / 340 trăsături; 16 kHz eșantionare	**88,6**

date RODeCAR, deoarece fișierele sale individuale au durate foarte lungi (de la câteva zeci de minute până la o oră).

5.5. Concluziile capitolului

În acest capitol, au fost propuse, implementate și validate mai multe subsisteme de detecție a vorbirii (VAD) bazate pe rețele neurale profunde (DNN). Subsistemul VAD este utilizat pentru extragerea trăsăturilor prozodice la nivel de rostire.

Pentru sarcina principală DSD, s-a demonstrat că abordarea la nivel de rostire este mai potrivită decât alte abordări la nivel de vorbitor sau de înregistrare pentru aplicații de expertiză criminalistică. Au fost propuse, implementate și validate mai multe sisteme DSD bazate pe rețele neurale. Cea mai performantă arhitectură a fost o nouă rețea hibridă de tip CNN-MLP, care a valorificat o fuziune a hărților de trăsături extrase automat și a unor subseturi de trăsături tradiționale, selectate pe baza unui nou algoritm de selecție propus. În abordarea la nivel de rostire, sistemul atinge o acuratețe de **63,7%** în cazul setului de date RLDD și de **62,4%** în cazul setului de date RODeCAR.

Pentru RLDD, performanța la nivel de vorbitor a fost de **85,6%**, reprezentând o creștere de 20,22% față de alte sisteme comparabile raportate în literatură; iar cea la nivel de înregistrare a fost de **88,6%**, o creștere corespunzătoare de 8,71%.

Pentru setul de date RODeCAR, abordarea la nivel de înregistrare este incompatibilă, dar abordarea la nivel de vorbitor a condus la o acuratețe de **83,5%**.

Capitolul 6

Recunoașterea emoțiilor din vorbire, urmărirea comportamentului suspect

Acest capitol se referă la sarcina de determinare a conținutului afectiv prezent în vorbire, denumită și *recunoașterea emoțiilor din vorbire* (SER). Părți din conținutul de față au fost publicate într-un articol de conferință [Mih19a] și într-un articol de revistă [Mih21c] de către candidat. Părți din conținutul de față au fost susținute de participarea candidatului ca asistent de cercetare în cadrul proiectului PN-III-P2-2-2.1-SOL-2016-02-0002, acord 2SOL/2017, finanțat de Guvernul României prin UEFISCDI: *Sisteme inteligente de analiză video și audio - Tehnologii și sisteme video inovative pentru reidentificarea persoanei și analiza comportamentului disimulat* (SPIA-VA) [Mih20].

6.1. Context și sinteza cercetării existente în domeniu

La proiectarea sistemelor SER, există două școli principale de gândire în psihologie care stabilesc modelarea conceptuală a emoțiilor:

- clase discrete [Laz99], în care fiecare emoție (sau, mai degrabă, fiecare clasă de emoții) se distinge holistic de celelalte; și
- modele dimensionale [Wat99], în care un număr de mărimi psihologice continue (intensitatea, valența) formează un spațiu afectiv multidimensional (de obicei 2D), fiecare emoție fiind o zonă în cadrul acestuia.

În acest sens, în literatura de specialitate au fost raportate rezultate promițătoare folosind modele și tehnici de învățare automată (ML) și profundă (DL), inclusiv modele Markov ascunse [Sha23b], mașini cu vectori de suport (SVM) [Jin15], DNN-uri de tip perceptron multistrat (MLP) [Atm20, Lat20a, Rao17], rețele neurale recurente (RNN) cu celule LSTM [Gha19, Liu20, Mir17], rețele neurale convoluționale (CNN) [Tan21,

Zha18a], modele hibride [Fah20] sau rețele neurale recurent-convoluționale avansate (CRNN) [Che18, Zha19], cu extragere a trăsăturilor fie algoritmică, fie automată.

6.2. Modele dimensionale pentru reprezentarea continuă-discretă a emoțiilor

Deoarece modelele dimensionale convenționale derivate în domeniul psihologiei permit doar o vagă „cartografiere" a spațiului afectiv, fără poziționare exactă sau delimitare numerică a claselor de emoții, ar fi util să se construiască modele mai precise (dar compacte și de complexitate redusă) folosind tehnici ML. Aceste modele ar permite determinarea clasei de emoție a unei instanțe pe baza poziției sale în spațiul afectiv (de exemplu, valorile sale de *intensitate* și *valență).* O abordare alternativă constă în adoptarea unor strategii multidomeniu diferite, în care paradigmele discrete și continue să fie legate direct între ele a priori [Mih21c].

Principala sursă de date pentru adaptarea modelului dimensional ar fi un set de date cu o dublă adnotare discretă și continuă a conținutului emoțional (etichete pentru clasele de emoții și valori numerice pentru dimensiunile afective). Singurul corpus de acest tip disponibil este setul de date Interactive Emotional Dyadic Motion Capture (IEMOCAP) [Bus08], care a fost utilizat aproape exclusiv pentru clasificarea emoțiilor. Se poate utiliza o sursă suplimentară, precum corpusul Warriner-Kuperman-Brysbaert (WKB) [War13], care include adnotări ale dimensiunilor afective pentru o serie de cuvinte, cele relevante reprezentând clasele de emoții, precum „furie" (i.e. „conceptul de furie"), etc. În abordarea propusă, aceste valori adnotate sunt valorificate pentru a inițializa centroizii (mediile) claselor. Raționamentul este că includerea corpusului WKB crește considerabil robustețea și duce la o generalizare mai bună.

Trei algoritmi ML au fost testați pentru dezvoltarea modelului dimensional al emoțiilor: modelul celor K medii (KMM), modele cu mixturi de Gaussiene (GMM) și mașini cu vectori suport (SVM). Pentru KMM și GMM, centroizii (mediile) claselor au fost inițializați în unul din două moduri: (i) utilizând valorile estimate prin calcularea mediei pe datele IEMOCAP (inițializare nativă); (ii) utilizând valorile estimate pe baza datelor WKB (inițializare WKB). Prima opțiune permite o mai bună potrivire a datelor, dar, în al doilea caz, generalizarea este mai mare. SVM-urile conduc la rezultate și mai bune datorită transformării spațiului afectiv într-unul mai mare dimensional, dar sunt limitate la IEMOCAP. Toate experimentele au fost efectuate prin validare încrucișată cu 5 repetări, rezervând un grup de vorbitori pentru testare (20% din date). Rezultatele sunt prezentate folosind ca măsuri acuratețea neponderată (UA) și cea ponderată (WA).

În Tabelul 6.2, abordarea propusă pentru dezvoltarea modelelor dimensionale este comparată cu alte lucrări care utilizează sisteme de clasificare standard pentru emoții discrete. După cum se poate observa, modelele dimensionale pot conduce la cele mai bune performanțe atâta timp cât există date fiabile privind dimensiunile afective; cu alte cuvinte, dacă valorile spațiului afectiv pentru segmentele de vorbire pot fi prezise corect de un model de regresie.

Tabelul 6.2 – Compararea performanțelor maxime cu alte lucrări care utilizează sisteme de clasificare standard pentru emoții discrete.

Context	Cele mai bune rezultate
[Che18]	UA = 64,7%
[Fah20]	UA = 66,0%, WA = 70,5%
[Jin15]	WA = 68,6%
[Lat20a]	UA = 61,0%
[Liu20]	UA = 65,0%, WA = 66,1%
[Mir17]	UA = 58,8%, WA = 63,5%
[Zha18a]	UA = 63,9%, WA = 70,4%
[Zha19]	UA = 67,0%, WA = 68,1%
Modelare dimensională (DM)	**UA = 74,3%, WA = 72,5%**

6.3. Arhitecturile sistemelor propuse

În această secțiune sunt prezentate mai multe arhitecturi de sistem complete care pot fi utilizate pentru SER, inclusiv tipuri bazate pe strategii multidomeniu care valorifică modelele dimensionale detaliate în Secțiunea 6.2.

Arhitecturile de sistem propuse pentru sarcina SER au fost dezvoltate iterativ și prezintă un nivel de complexitate crescător, încadrându-se în șase categorii (*abordări*):

1) modele DNN de sine stătătoare, pentru clasificare sau regresie;
2) strategii de clasificare în ansamblu incorporând mai multe modele DNN;
3) sisteme multidomeniu, reunind paradigmele de recunoaștere a emoțiilor;
4) modele DNN de clasificare adaptate prin învățare prin transfer (TL);
5) clasificare prin fuziune eterogenă folosind modele TL-DNN;
6) clasificare în ansamblu omogenă folosind modele TL-DNN.

Abordarea cu modele DNN de sine stătătoare primește la intrare un set extins de 2.258 de trăsături acustice, spectrale și cepstrale. Clasificatorul este un model de rețea neurală complet conectată (FCNN), care utilizează între 1 și 4 straturi ascunse, cu numere diferite de noduri per strat și cu un strat de ieșire de dimensiune egală fie cu numărul de clase, fie de doi neuroni, care corespund dimensiunilor spațiului afectiv 2D. Cel de-al doilea tip de sistem, mai avansat, revizitează abordarea de clasificare în ansamblu prezentată în Capitolul 3, extinzând-o către cele două strategii principale adoptate de modelele SVM pentru probleme multiclasă: unul-vs.-unul (OvO) și unul-vs.-restul (OvR). Pentru sistemele multidomeniu, sunt propuse șapte forme:

- tipul 1: forma nativă de învățare corelată;
- tipul 2: primele straturi active sunt comune pentru cele două sarcini, apoi există un strat de ieșire pentru una dintre sarcini, iar a doua secțiune activă este utilizată pentru cealaltă sarcină – (A) prima secțiune modelează regresia; (B) prima secțiune modelează clasificarea;

- tipul 3: prima secțiune activă se concentrează pe una dintre sarcini, iar a doua secțiune activă este antrenată pentru cealaltă sarcină pe aceleași date inițiale de intrare împreună cu ieșirile primei secțiuni active – (A) regresia este modelată de prima secțiune; (B) clasificarea este modelată de prima secțiune;
- tipul 4: aceste abordări secvențiale valorifică modelele dimensionale (DM) pre-antrenate, așa cum este descris în Secțiunea 6.2, pentru a stabili legătura dintre spațiul continuu al afectelor și poziția în cadrul acestuia (determinată prin regresie) și clasa de emoție – (A) se utilizează aplicarea directă a DM pentru a determina clasa de emoție pe baza ieșirii modelului de regresie DNN; (B) un al doilea DNN este antrenat pe datele de intrare inițiale împreună cu clasificarea preliminară furnizată de DM.

Învățarea prin transfer (TL) este o tehnică de învățare profundă (DL) care valorifică transformările spațiului de date corespunzătoare unei sarcini pentru care au fost proiectate DNN-urile, adaptându-le pentru o sarcină diferită, conexă. În contextul SER, acest lucru se realizează prin adoptarea unor modele de recunoaștere a imaginilor foarte performante și prin reprezentarea instanțelor de date într-o formă compatibilă cu imaginile, de exemplu, spectrograme. Modelele moderne de recunoaștere a imaginilor (denumite în continuare TL-DNN) sunt: Xception, VGG16 și VGG19, ResNet50, ResNet50V2, InceptionV3, InceptionResNetV2, NASNetMobile și NASNetLarge și EfficientNetB0 până la EfficientNetB7, antrenate pe setul de date ImageNet.

În această lucrare, prima abordare propusă bazată pe TL constă în reantrenarea straturilor superioare ale fiecărui TL-DNN pentru a dezvolta modele de clasificare de sine stătătoare. Mergând mai departe, se propune o formă de reprezentare a informațiilor prin fuziune sub forma unui sistem TL-DNN eterogen: nucleul fiecăruia dintre modelele TL-DNN este utilizat pentru a extrage reprezentări profunde ale hărților de trăsături ale datelor de intrare. Toate reprezentările sunt apoi concatenate într-un singur vector care este ulterior trimis unui clasificator DNN. Abordarea omogenă bazată pe TL, spre deosebire de sistemul eterogen, valorifică strategiile de clasificare de ansamblu (OvO și OvR), dar cu modele TL-DNN. Proprietatea de omogenitate se referă la faptul că, pentru fiecare combinație de clase, se utilizează același model în cadrul unui sistem.

Cele trei abordări bazate pe TL-DNN pentru sistemele SER sunt denumite a patra, a cincea și a șasea abordare globală pentru SER. Pentru abordările bazate pe TL-DNN, datele de intrare furnizate rețelelor trebuie să fie sub formă de spectrograme. Acestea au fost extrase la scară liniară sau logaritmică, utilizând ferestre Hamming cu durata de 25 ms (suprapunere de 15 ms), cu scalare liniară sau Mel și 3 canale de culoare (RGB).

6.4. Metodologie și rezultate experimentale

Baza de date Berlin Database of Emotional Speech (EMODB) [Bur05] este un set de date în limba germană care cuprinde 535 propoziții scurte înregistrate de 10 actori (5 femei, 5 bărbați) special aleși de un juriu având ca principale criterii naturalețea și

inteligibilitatea. Enunțurile au o durată medie de 2,5 s și o durată maximă de 8 s. Cele 7 clase de emoții luate în considerare sunt următoarele Furie (ANG), Dezgust (DIS), Frică (FEA), Tristețe (SAD), Plictiseală (BOR), Fericire (HAP) și Neutru (NEU).

Deoarece această lucrare se concentrează pe expertiză criminalistică, în special monitorizarea comportamentului suspect, clasele de emoții negative sunt mai relevante și mai important de detectat (individual, pentru aplicații care necesită mai multe detalii și nuanțe), precum și manifestările afective negative în general (considerate împreună ca un singur grup). În afară de setul complet de 7 clase, au fost luate în considerare trei subseturi suplimentare. Cele 4 sunt:

- **EMODB-7**: 7 clase: ANG, DIS, FEA, SAD, BOR, HAP și NEU;
- **EMODB-5N**: 5 clase: ANG, DIS, FEA, SAD și NEU;
- **EMODB-4**: 4 clase: ANG, SAD, HAP și NEU;
- **EMODB-2N**: 2 clase: Negative (NEG; grupând ANG, DIS, FEA și SAD) vs. NEU.

Setul de date Crowd-sourced Emotional Multimodal Actors (CREMAD) [Cao14] cuprinde 7.442 de înregistrări în limba engleză de conținut afectiv manifestat în propoziții rostite de 91 de actori (43 de femei, 48 de bărbați). Cele 6 clase de emoții au fost următoarele: Furie (ANG), Dezgust (DIS), Frică (FEA), Tristețe (SAD), Fericire (HAP) și Neutru (NEU). Durata medie a înregistrărilor este de 2,5 s. Cele 4 subseturi utilizate în această lucrare sunt:

- **CREMAD-6**: 6 clase: ANG, DIS, FEA, SAD, HAP și NEU;
- **CREMAD-5N**: 5 clase: ANG, DIS, FEA, SAD și NEU;
- **CREMAD-4**: 4 clase: ANG, SAD, HAP și NEU;
- **CREMAD-2N**: 2 clase: Negative (NEG; grupând ANG, DIS, FEA și SAD) vs. NEU.

Setul de date IEMOCAP [Bus08] include 10 actori (5 femei, 5 bărbați) care lucrează în perechi pentru a rezolva sarcini de vorbire în limba engleză, scrise și improvizate, cu un număr total de 10.039 de înregistrări audio-vizuale. În total sunt 10 clase de emoții discrete (Furie, Frică, Dezgust, Tristețe, Fericire, Frustrare, Entuziasm, Surpriză, Neutru și Altele), dar multe dintre ele subreprezentate. Rezultă necesitatea de a grupa doar un subset mai mic dintre acestea în 4 clase noi, și anume: Neutru (NEU); Tristețe (SAD); Mânie + Frustrare (ANG); și Fericire + Excitare (HAP). Dimensiunile continue sunt *intensitatea* și *valența*. Cele 2 subseturi luate în considerare sunt:

- **IEMOCAP-4**: 4 clase: ANG, HAP, SAD și NEU;
- **IEMOCAP-2N**: 2 clase: Negative (NEG; grupând ANG și SAD) vs. NEU.

Pentru *Abordarea 1* (modele DNN de sine stătătoare), numărul de straturi ascunse a fost ales între 1 și 4, numărul de neuroni pentru primul strat ascuns fiind ales din setul {8, 16, 32, 64, 128, 256, 512, 1024}, iar rata de antrenare selectivă a neuronilor a variat între 0,1 și 0,5. Alți hiperparametri aleși au inclus: funcția de activare ReLU pentru straturile ascunse și fie softmax (pentru clasificare), fie funcția liniară (pentru regresie) drept funcție de activare pentru stratul de ieșire. Aceleași configurații au fost testate ulterior pentru clasificatorii OvO și OvR DNN (*Abordarea 2*), clasificatorul DNN final având o adâncime fixă de 1, 2 sau 3 straturi. Configurațiile au fost de asemenea

Tabelul 6.13a – Comparație între cele mai bune rezultate pentru clasificare (SER) obținute în această lucrare și alte rezultate relevante publicate în literatură.

Subsetul de date	**Sistem**	**Perf.**	
		UA [%]	**WA [%]**
EMODB-7	[Ker19] – SVM + eliminarea recursivă a trăsăturilor	–	**86,2**
	[Che18] – CRNN	–	82,8
	[Lot17] – SNN + LSM + banc de filtre Gammatone	–	82,4
	[Bis13] – SVM + recunoașterea genului	–	81,5
	[Cha14] – GMM	79,8	–
	[Yil21] – SVM + selecția trăsăturilor	78,6	79,1
	[Kan21] – GA + grupare	77,5	–
	[Cha14] – SVM	77,0	–
	Această lucrare: *Abordarea 2* – clasificare în ansamblu (OvR) folosind mai multe DNN-uri (FCNN).	**82,6**	**82,9**
EMODB-4	[Vas15] – GMM + SVM	–	84,3
	Această lucrare: *Abordarea 2* – clasificare în ansamblu (OvR) folosind mai multe DNN-uri (FCNN).	**88,9**	**89,1**
EMODB-5N	[He15] – MLP + propagare inversă bazată pe GA	–	80,4
	Această lucrare: *Abordarea 2* – clasificare în ansamblu (OvR) folosind mai multe DNN-uri (FCNN).	**91,2**	**91,4**
EMODB-2N	[Cas08] – SVM	–	95,8
	[Vas15] – GMM + SVM	–	94,9
	Această lucrare: *Abordarea 1* – DNN (FCNN) de sine stătător.	**95,1**	**98,3**
CREMAD-6	[Gha20] – SVM	–	**57,2**
	[Gha19] – LSTM	–	**57,0**
	[Bea18] – LSTM	–	41,5
	Această lucrare: *Abordarea 6* – clasificare în ansamblu (OvO) omogenă folosind modele TL-DNN (EfficientNetB1).	**51,8**	**54,6**
CREMAD-4	**Această lucrare**: *Abordarea 6* – clasificare în ansamblu (OvO) omogenă folosind modele TL-DNN (EfficientNetB1).	**65,8**	**70,3**
CREMAD-5N	**Această lucrare**: *Abordarea 6* – clasificare în ansamblu (OvO) omogenă folosind modele TL-DNN (EfficientNetB0).	**54,7**	**58,7**
CREMAD-2N	**Această lucrare**: *Abordarea 1* – DNN (FCNN) de sine stătător.	**72,8**	**72,6**
IEMOCAP-4	[Yi22] – DNN + augmentare adversarială a datelor	**63,7**	**63,2**
	[Lat20a] – MLP + date sintetizate prin GAN	**61,0**	–
	[Yi22] – SVM + augmentare adversarială a datelor	**60,0**	**64,7**
	[Yil21] – SVM + selecția trăsăturilor	**59,4**	59,5
	[Mir17] – MLP + LSTM + atenție	58,7	**63,5**
	[Pan20] – LSTM	48,7	57,1
	[Rao17] – MLP + i-vectori	–	48,8
	Această lucrare: *Abordarea 2* – clasificare în ansamblu (OvR) folosind mai multe DNN-uri (FCNN).	**58,7**	**61,6**
IEMOCAP-2N	[Rah12] – SVM + adaptarea trăsăturilor	–	69,8
	Această lucrare: *Abordarea 1* – DNN (FCNN) de sine stătător.	**69,0**	**71,2**

Tabelul 6.13b – Comparație între cele mai bune rezultate pentru regresie (SER) obținute în această lucrare și alte rezultate relevante publicate în literatură. Rezultatele sunt separate pe fiecare dimensiune afectivă: A = intensitate, V = valență.

Set de date	Sistem	Performanță					
		MSE (f. cost)		ρ		ρ_c	
		A	V	A	V	A	V
IEMOCAP	[Atm20] – MLP	–	–	–	–	0,611	0,301
	[Zha18b] – DNN	–	–	–	–	0,392	**0,715**
	Această lucrare: *Abordarea 1* – DNN (FCNN) de sine stătător.	**0,073**	**0,180**	**0,677**	**0,408**	**0,621**	**0,343**

utilizate pentru mai multe tipuri de sisteme multidomeniu (*Abordarea 3)*, precum și pentru secțiunile de clasificare complet conectate în experimentele TL-DNN.

Pentru toate experimentele, s-a utilizat ca metodologie de testare validarea încrucișată cu 10 repetări, cu o împărțire de 80% / 20%, asigurându-se cât mai bine ca fiecare clasă de emoții și fiecare gen să fie reprezentate proporțional în fiecare subset de antrenare și validare. Separarea vorbitorilor a fost asigurată pentru toate experimentele.

În Tabelul 6.13a și Tabelul 6.13b sunt făcute comparații de performanță între sistemele propuse și altele raportate în literatura de specialitate.

6.5. Concluziile capitolului

În acest capitol, a fost oferită o introducere detaliată a sarcinii SER și a provocărilor acesteia, stabilind cele două paradigme fundamentale de modelare a emoțiilor. Au fost propuse, dezvoltate și testate sisteme SER bazate pe rețele neuronale profunde (DNN) care acoperă șase niveluri de complexitate: DNN-uri de sine stătătoare, clasificare în ansamblu (unul-vs.-unul, OvO, și unul-vs.-restul, OvR) folosind mai multe DNN-uri conectate între ele, și sisteme care valorifică învățarea prin transfer (TL) pentru modelele moderne de vârf de învățare profundă pentru recunoașterea imaginilor, fie drept modele TL-DNN independente, fie drept clasificatori în ansamblu eterogeni sau omogeni. Sistemele au fost testate pe cele mai relevante seturi de date SER disponibile: EMODB, CREMAD și IEMOCAP, pentru setul complet (standard) de clase, precum și pentru subseturi suplimentare de emoții negative relevante pentru monitorizarea comportamentului suspect și alte aplicații care intră în aria de aplicare a acestei lucrări.

Sistemele propuse au obținut rezultate superioare (până la **83%** acuratețe) pentru subsetul de toate clasele EMODB, în timp ce performanța pentru subseturile corespunzătoare CREMAD și IEMOCAP a fost mai mică (până la **55%** acuratețe pentru CREMAD și **62%** acuratețe pentru IEMOCAP), dar totuși comparabilă cu alte cercetări publicate. În plus, pentru toate subseturile conținând doar emoții negative, soluțiile propuse au oferit cele mai bune performanțe raportate până în momentul de față.

Capitolul 7

Remanența emoțiilor în vorbire

Acest capitol acoperă studiul *remanenței emoțiilor în vorbire* în contextul expertizei criminalistice și modul în care recunoașterea emoțiilor (SER) poate fi aplicată. Părți din conținutul de față au fost publicate ca articol de revistă de către candidat [Mih22b].

7.1. Context și sinteza cercetării existente în domeniu

Dincolo de performanța sistemelor SER în sine, una dintre provocările acestei sarcini constă în a discerne evoluții temporale ale conținutului afectiv care ar putea indica un comportament suspect. În acest scop, conținutul afectiv al mostrelor de vorbire a fost analizat, utilizând un set de date nou, pe intervale de timp scurt (până în 1 oră) și pe intervale mai lungi (peste 5 zile) [Mih22b].

Cercetările anterioare [Liu21, Su21, Zha21a, Zha22] au arătat că modelele ML și DL încă au performanțe relativ slabe aplicate pe seturi de date nefolosite la antrenare, chiar și atunci când se utilizează tehnici avansate și costisitoare pentru adaptarea datelor de intrare. Această putere de generalizare redusă poate fi cauzată, cel puțin în parte, de diferențele de exprimare emoțională în raport cu cultura, mediul, vârsta etc. vorbitorului, dar nu există dovezi concludente în favoarea sau împotriva acestei idei.

7.2. Studiu asupra remanenței emoțiilor în vorbire

Cele două ipoteze principale ale acestui studiu au fost următoarele [Mih22b]:

1) Dacă o interacțiune umană este solicitantă emoțional pentru subiect, atunci răspunsul său afectiv nu se va diminua instantaneu după ce interacțiunea se încheie, ci pe o perioadă mai lungă de timp, iar interacțiunile ulterioare neutre emoțional vor fi încă însoțite de o stare afectivă negativă.

2) În contextul existenței în viitor a unui eveniment cu încărcătură emoțională pentru subiect (și de care acesta este conștient), pe măsură ce evenimentul se apropie, subiectul va experimenta emoții de intensitate mai mare și va prezenta un răspuns afectiv corespondent crescut.

Pentru a testa ambele ipoteze, a fost dezvoltat un set de date folosind înregistrări vocale ale interacțiunilor recurente cu un număr de studenți având examene restante și care învățau pentru a le susține pentru a doua sau a treia oară. Au fost implicați în acest proces 18 studenți (4 femei, 14 bărbați), cu vârste cuprinse între 19,7 ani și 23,3 ani. Numărul total de rostiri înregistrate a fost de 270, iar durata totală a conținutului setului de date este de 1 oră și 8 minute.

Pentru a valida aplicabilitatea recunoașterii automate a emoțiilor din vorbire în acest context, au fost dezvoltate sisteme bazate pe rețele neurale complet conectate (FCNN), urmând aceeași abordare descrisă în Capitolul 6. Au fost luate în considerare două structuri pentru straturile ascunse: arhitectura „constant", care constă în același număr de noduri pentru fiecare strat ascuns; și arhitectura „log2dec", care constă într-un număr progresiv mai mic de noduri per strat, urmând o lege logaritmică. Adâncimea a variat între 2 și 4, cu un număr inițial de noduri de 256, 128, 64 sau 32. Pentru fiecare strat ascuns, s-a utilizat antrenarea selectivă a neuronilor, cu o rată cuprinsă între 20% și 50%. Printre ceilalți hiperparametri aleși se numără: funcția de activare ReLU pentru straturile ascunse și funcția de activare softmax (pentru clasificare) sau identitate (pentru regresie) pentru stratul de ieșire.

Datele sunt prezentate față de fiecare moment de timp în Figura 7.2, etichetele referindu-se la răspunsul afectiv inițial și la răspunsul afectiv după 15 și, respectiv, 30 de minute de conversație neutră. Pentru problema de regresie, valorile pentru intensitate și valență vs. fiecare zi sunt reprezentate în Figura 7.3 pentru fiecare moment de timp: pentru fiecare vorbitor în parte (linii subțiri) și media tuturor vorbitorilor (linii groase).

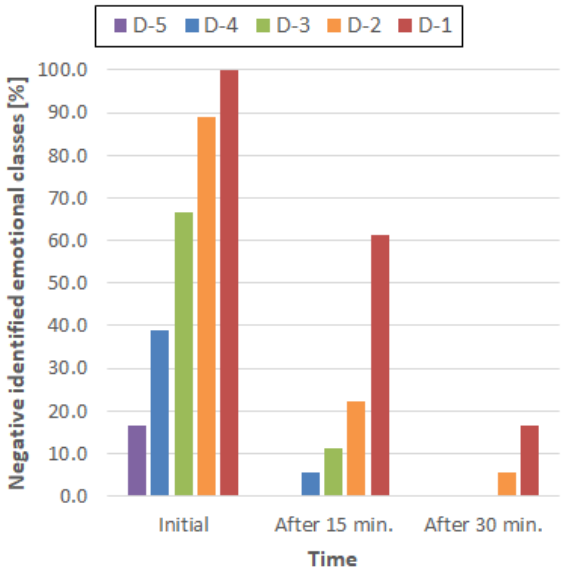

Figura 7.2 – Procentul vorbitorilor identificați ca exprimând emoții *negative*.

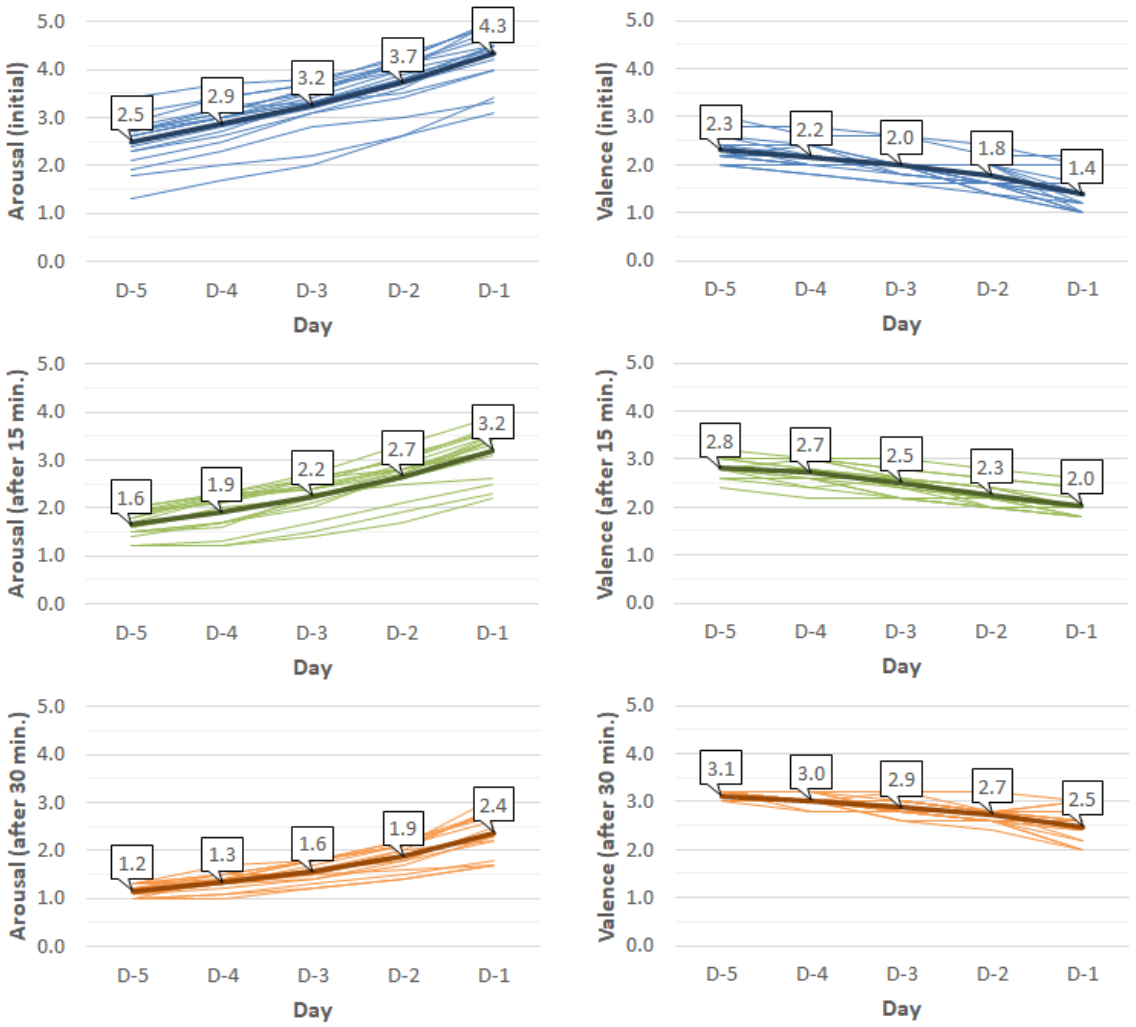

Figura 7.3 – Evoluția intensității și a valenței în funcție de zi, pentru fiecare moment de timp. Liniile subțiri reprezintă evoluțiile vorbitorilor individuali, în timp ce liniile groase etichetate reprezintă valorile medii pentru toți vorbitorii.

7.3. Concluziile capitolului

În acest capitol, s-a oferit o perspectivă asupra remanenței emoțiilor în vorbire prin investigarea pe termen scurt (sub 1 oră) și lung (5 zile), diferite de cele utilizate în alte cercetări privind recunoașterea emoțiilor, mai relevante pentru aplicațiile vizate.

S-a dovedit că: (1) dacă o interacțiune umană este solicitantă emoțional pentru subiect, atunci răspunsul său afectiv nu va scădea instantaneu, ci pe o perioadă mai lungă de timp, iar interacțiunile ulterioare neutre emoțional vor fi încă însoțite de o stare afectivă negativă; și (2) dacă un eveniment cu încărcătură emoțională este iminent pentru subiect, pe măsură ce evenimentul se apropie, subiectul va experimenta emoții de intensitate mai mare și va prezenta un răspuns afectiv corespondent crescut.

Capitolul 8

Concluzii

8.1. Dezvoltări și rezultate obținute

În Capitolul 1, au fost definite scopul și obiectivele, acoperind dezvoltarea de sisteme de inteligență artificială pentru recunoașterea automată a elementelor paralingvistice utilizând doar date audio, cu accent pe emoții negative, nivel ridicat de stres și detecția minciunilor, principalul domeniu de aplicare fiind expertiza criminalistică a vorbirii.

Capitolul 2 a prezentat un rezumat al principalelor concepte teoretice utilizate în dezvoltarea acestei lucrări în ceea ce privește trăsăturile de semnal vocal extrase algoritmic pentru analiză și procesare, modelele adoptate în dezvoltarea sistemelor, precum și metodologiile de antrenare și testare, tehnicile și măsurile de performanță.

În Capitolul 3, au fost propuse și dezvoltate sisteme de detecție a stresului din vorbire (SSD), utilizând ansambluri de rețele neurale complet conectate (DNN). S-au obținut îmbunătățiri pe setul de date SUSAS față de rezultatele de ultimă generație raportate anterior, cu o acuratețe (neponderată / ponderată) de **68,8% / 65,5%** pentru cazul de stres *real* cu 4 clase, **59,2% / 62,4%** pentru cazul de stres *real* cu 3 clase, **66,7% / 81,4%** pentru cazul de stres *real* cu 2 clase, **75,5% / 75,5%** pentru cazul de stres *simulat* cu 4 clase, și **76,1% / 78,4%** pentru cazul de stres *simulat* cu 4 clase.

Capitolul 4 s-a axat pe prezentarea bazei de date Romanian Deva Criminal Investigation Audio Recordings (RODeCAR): un set de date de rostiri adevărate și neadevărate, dezvoltat de candidat prin analiza, procesarea și examinarea înregistrărilor originale arhivate din 9 cazuri criminale. Setul de date conține **3 h 28 min** de rostiri înregistrate de la 20 de vorbitori (4 femei, 16 bărbați): 2 h 6 min (60,5%) conținut adevărat, 1 h 22 min (39,5%) conținut neadevărat, adnotat în mod obiectiv.

În Capitolul 5, sarcina principală este dezvoltarea de sisteme pentru detecția minciunilor (DSD). În arhitectura propusă, a fost necesar și un subsistem de detecție a vorbirii, care a fost mai întâi dezvoltat și discutat. După ce s-a demonstrat că abordarea

la nivel de rostire este mai potrivită pentru aplicațiile criminalistice decât alte abordări la nivel de vorbitor sau la nivel de înregistrare, au fost propuse, implementate și validate patru sisteme DSD bazate pe rețele neurale. Cea mai performantă arhitectură a fost o nouă rețea hibridă. La nivel de rostire, sistemul atinge o acuratețe de **63,7%** pe setul de date RLDD și **62,4%** pe setul de date RODeCAR. Sistemul propus a fost testat pentru a valida performanța la nivel de vorbitor și la nivel de înregistrare. Pentru RLDD, performanța la nivel de vorbitor a fost de **85,6%**, reprezentând o creștere de 20,22% față de alte sisteme comparabile raportate în literatură; iar la nivel de înregistrare a fost de **88,6%**, o creștere de 8,71%. Abordarea la nivel de înregistrare nu este compatibilă cu RODeCAR; abordarea la nivel de vorbitor a condus la o acuratețe de **83,5%**.

Capitolul 6 a oferit o introducere detaliată în sarcina de recunoaștere a emoțiilor din vorbire (SER). Au fost propuse, dezvoltate și testate sisteme bazate pe DNN-uri cuprinzând șase niveluri de complexitate, inclusiv DNN-uri, DNN-uri multiple conectate în ansamblu, precum și sisteme care valorifică învățarea prin transfer (TL) pentru modelele cele mai performante de învățare profundă pentru recunoașterea imaginilor, fie drept modele TL-DNN independente, fie prin clasificare în ansamblu eterogenă sau omogenă. Sistemele au fost testate pe cele mai relevante seturi de date SER: EMODB, CREMAD și IEMOCAP. Sistemele propuse au obținut rezultate bune (până la **83%** acuratețe) pentru subsetul cu toate clasele EMODB, în timp ce performanța pe subseturile corespondente CREMAD și IEMOCAP a fost mai mică (până la **55%** pentru CREMAD și **62%** pentru IEMOCAP), dar totuși comparabilă cu alte cercetări publicate. În plus, pentru toate subseturile care cuprind doar conținut afectiv negativ, soluțiile propuse au oferit cea mai bună performanță.

În sfârșit, Capitolul 7 a servit ca o continuare a problemei SER, oferind o discuție detaliată asupra remanenței emoțiilor în vorbire, investigând efectele pe termen scurt (sub 1 h) și lung (5 zile), care sunt mai relevante pentru aplicațiile avute în vedere în cadrul acestei lucrări. S-a dovedit că: (1) dacă o interacțiune umană determină o reacție emoțională pentru subiect, atunci răspunsul afectiv al acestuia nu se va diminua instantaneu, ci pe o perioadă mai lungă de timp, iar interacțiunile ulterioare neutre din emoțional vor fi în continuare însoțite de o stare afectivă negativă mai intensă (remanență emoțională); și (2) dacă urmează un eveniment cu încărcătură emoțională pentru subiect, pe măsură ce evenimentul se apropie, subiectul va experimenta emoții de o intensitate mai mare și va prezenta un răspuns afectiv crescut.

8.2. Contribuții originale

Contribuții generale și globale

- Strategiile de clasificare utilizând ansambluri de rețele neurale în configurațiile OvO și OvR au fost dezvoltate și valorificate cu succes pentru SSD, DSD și SER. Descrierea strategiilor a fost prezentată în Capitolul 3, iar rezultatele obținute prin utilizarea lor au fost detaliate în Capitolul 3 și Capitolul 5 și publicate în [Mih21b, Mih22a].

- A fost dezvoltat un set de trăsături extrase algoritmic pentru sarcinile de recunoaștere automată a elementelor paralingvistice, extinzând cel mai utilizat set de trăsături de referință disponibil în literatura de specialitate cu mai multe mărimi suplimentare. În toate sarcinile (SSD, DSD, SER), setul de trăsături propus a fost utilizat cu succes. Descris în Capitolul 2, rezultatele obținute cu rețele neurale antrenate pe acesta au fost discutate în Capitolul 3, Capitolul 5 și Capitolul 6 și publicate în [Mih21b, Mih22a].

- Un algoritm de selecție a trăsăturilor pentru probleme de clasificare binară bazat pe testul Kolmogorov-Smirnov a fost propus și aplicat pentru DSD. Descrierea algoritmului a fost prezentată în Capitolul 5, iar rezultatele obținute prin încorporarea acestuia în sistemele DSD au fost publicate în [Mih22a].

- Un subsistem de detecție a vorbirii (VAD) a fost dezvoltat și utilizat pentru a extrage trăsăturile prozodice utilizate pentru DSD. Descrierea sistemului a fost prezentată în Capitolul 5, iar rezultatele au fost publicate în [Mih21a].

Contribuții la detecția stresului din vorbire (SSD)

- Au fost utilizate abordări noi pentru SSD în contextul aplicațiilor de expertiză criminalistică, folosind grupări de clase și analize specifice domeniului de aplicare al acestei lucrări. Rezultatele obținute au demonstrat performanțe îmbunătățite față de majoritatea literaturii publicate anterior. Rezultatele au fost prezentate în Capitolul 3 și publicate în [Mih21b].

Contribuții la detecția minciunilor din vorbire (DSD)

- Setul de date Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) a fost dezvoltat pentru sarcinile DSD. Descrierea sa completă a fost prezentată în Capitolul 4 și publicată în [Mih19b]. Acesta este:
  i) unul dintre foarte puținele seturi de date disponibile public care oferă date realiste și comentate obiectiv pentru DSD, cuprinzând înregistrări de comportament nesimulat în scenarii realiste cu mize mari;
  ii) singurul set de date de acest tip disponibil pentru limba română;
  iii) o bază de date consistentă pentru aplicații paralingvistice, în special DSD, care cuprinde aproximativ 3,5 ore de vorbire.

- Pentru DSD, a fost utilizată o abordare nouă, mai dificilă, mai detaliată și mai potrivită pentru expertiza criminalistică, prin antrenarea sistemelor propuse pentru a discerne între vorbirea adevărată și neadevărată la nivelul rostirilor (pe termen scurt) în loc de al înregistrărilor (pe termen lung) sau al vorbitorului (determinarea atitudinii generale a unei persoane):

i) raționamentul a fost explicat în Capitolul 5 și publicat în [Mih22a], reprezentând primele rezultate publicate folosind abordarea la nivel de rostire propusă;
ii) rezultatele obținute la nivel de înregistrare și la nivel de vorbitor au fost, de asemenea, prezentate în Capitolul 5 și reprezintă o creștere semnificativă a performanțelor față de literatura publicată anterior.

- Arhitecturile hibride de rețele neurale profunde au fost dezvoltate pentru DSD, combinând extragerea automată a trăsăturilor specifică rețelelor neurale convoluționale cu relevanța unor trăsături bine alese, extrase algoritmic. Descrierea detaliată a arhitecturilor hibride a fost realizată în Capitolul 5, iar rezultatele obținute cu această abordare au fost publicate în [Mih21a, Mih22a].

Contribuții la recunoașterea emoțiilor din vorbire (SER)

- O investigație teoretică și experimentală a remanenței emoțiilor în vorbire a fost realizată pentru a valida două ipoteze importante pentru expertiza criminalistică: (1) răspunsurile afective la evenimente cu încărcătură emoțională se diminuează pe perioade lungi de timp, interacțiunile neutre ulterioare fiind însoțite de stări afective negative; și (2) evenimentele iminente cu încărcătură emoțională determină emoții de intensitate mai mare și manifestări de răspunsuri afective corespondent intensificate. Acestea au fost prezentate în Capitolul 7 și publicate în [Mih22b]. Studiul aplicat:
  i) este unul dintre puținele efectuate pe această temă și singurul realizat pentru intervalele de timp alese (până în o oră, zilnic, 5 zile), care au fost justificate ca fiind cele mai relevante pentru aplicațiile vizate;
  ii) a inclus validarea experimentală a utilizării sistemelor SER pentru monitorizarea manifestărilor emoționale și a evoluției temporale a acestora, relevante pentru aplicațiile avute în vedere.

- Pentru SER au fost dezvoltate modele dimensionale îmbunătățite, rafinând legătura dintre paradigma claselor emoționale discrete și cea de modelare continuă a spațiului afectiv pentru analiza și recunoașterea emoțiilor. Acestea au fost discutate în Capitolul 6 și publicate în [Mih21c]. Modelele:
  i) au fost elaborate prin corelarea datelor din unul dintre puținele seturi de date SER cu adnotare dublă, de asemenea unul dintre cele mai des citate și utilizate în cercetările recente din domeniu;
  ii) au fost îmbunătățite prin multimodalitate, prin rafinarea versiunilor inițiale obținute pe baza datelor audio cu date textuale relevante claselor emoționale dintr-un corpus de dimensiune mare.

- Au fost utilizate abordări noi pentru SER în contextul aplicațiilor de expertiză criminalistică, cu accent pe emoții negative. O parte din rezultatele

obținute au demonstrat performanțe îmbunătățite față de majoritatea literaturii publicate anterior în acest domeniu. Rezultatele au fost prezentate în Capitolul 6 și publicate parțial în [Mih19a].

- Sisteme bazate pe învățarea prin transfer au fost dezvoltate pentru SER folosind cele mai performante rețele neurale moderne de recunoaștere a imaginilor (VGG16, VGG19, Inception, Xception și mai multe versiuni ale ResNet50, NASNet și EfficientNet), de sine stătătoare și prin clasificare în ansamblu. Metodologia și rezultatele au fost prezentate în Capitolul 6.

Mențiune asupra activității de cercetare

Părți din această lucrare au fost susținute prin participarea candidatului între 2017 și 2020 ca asistent de cercetare în cadrul proiectului PN-III-P2-2-2.1-SOL-2016-02-0002, acord 2SOL/2017, finanțat de Guvernul României prin UEFISCDI: *Sisteme inteligente pentru analiză video și audio – Tehnologii și sisteme video inovatoare pentru reidentificarea persoanei și analiza comportamentului disimulat* (SPIA-VA) [Mih20].

8.3. Lista lucrărilor originale

În cadrul acestui doctorat, au fost publicate 3 articole de jurnal (unul clasat Q1, unul clasat Q2) și 4 lucrări de conferință, pentru care candidatul a fost primul autor.

Articole de jurnal

1) **Ș. Mihalache** și D. Burileanu, „Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition," în *University Politehnica of Bucharest Scientific Bulletin, Series C – Electrical Engineering and Computer Science*, vol. 83, nr. 4, Editura Politehnica Press, București, pp. 137-148, Dec. 2021. ISSN: 2286-3540. [Mih21c]
ISI WOS: 000741473700013 (Q4, IF: 0.3)

2) **Ș. Mihalache** și D. Burileanu, „Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," în *Sensors*, vol. 22, nr. 3, art. nr. 1228, MDPI, Basel, Elveția, pp. 1-21, Feb. 2022. ISSN: 1424-8220. DOI: 10.3390/s22031228. [Mih22a]
ISI WOS: 000759983300001 (Q1, IF: 3.576 – Feb. 2022)

3) **Ș. Mihalache**, D. Burileanu, E. Franți, M. Dascălu și C.A. Brătan, „Lasting emotions – An investigation of short- and long-term affective content remanence in speech," în *Romanian Journal of Information Science and Technology*, vol. 25, nr. 1, Editura Academiei Române, București, pp. 20-35, Mar. 2022. ISSN: 1453-8245. [Mih22b]
ISI WOS: 000775912300002 (Q2, IF: 3.5)

Lucrări de conferință

4) **Ș. Mihalache**, D. Burileanu, G. Pop și C. Burileanu, „Modulation-based speech emotion recognition with reconstruction error feature expansion,” în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, România, pp. 1-6, Oct. 2019, IEEE NY. ISBN: 978-1-7281-0983-1. DOI: 10.1109/SPED.2019.8906537. [Mih19a]
ISI WOS: 000571718700004

5) **Ș. Mihalache**, G. Pop și D. Burileanu, „Introducing the RODeCAR database for deceptive speech detection,” în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, România, pp. 1-6, Oct. 2019, IEEE NY. ISBN: 978-1-7281-0983-1. DOI: 10.1109/SPED.2019.8906542. [Mih19b]
ISI WOS: 000571718700006

6) **Ș. Mihalache**, I.A. Ivanov și D. Burileanu, „Deep neural networks for voice activity detection,” în *Proc. International Conference on Telecommunications and Signal Processing* (*TSP*), Brno, Czech Republic, pp. 191-194, Iul. 2021, IEEE NY. ISBN: 978-1-6654-2933-7. DOI: 10.1109/TSP52935.2021.9522670. [Mih21a]
ISI WOS: 000701604600041

7) **Ș. Mihalache**, D. Burileanu și C. Burileanu, „Detecting psychological stress from speech using deep neural networks and ensemble classifiers,” în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), București, România, pp. 74-79, Oct. 2021, IEEE NY. ISBN: 978-1-6654-2786-9. DOI: 10.1109/SpeD53181.2021.9587430. [Mih21b]
ISI WOS: 000786794700014

8.4. Perspective pentru dezvoltări ulterioare

În ceea ce privește viitoarele cercetări și activități ale candidatului în domeniul învățării automate, al învățării profunde, al analizei și prelucrării vorbirii, al recunoașterii automate a elementelor paralingvistice și al expertizei criminalistice ale vorbirii, există mai multe căi promițătoare disponibile și de mare interes și relevanță în continuare.

Pentru fiecare dintre sarcinile de recunoaștere a elementelor paralingvistice din cadrul acestei lucrări, anume SSD, DSD și SER, se pot obține îmbunătățiri suplimentare prin utilizarea unor modele convenționale diferite, cum ar fi autoencodere suprapuse (SAE) sau mașini extreme de învățare (ELM). Dincolo de investigarea modelelor alternative, o idee atractivă suplimentară este aceea de a adapta tehnici specifice altor tipuri de rețele neurale profunde, de exemplu, mecanismele de atenție utilizate pentru rețele neurale recurente (RNN) sau transformatoare. În special pentru sarcina DSD, o abordare multimodală care să implice atât date audio, cât și date textuale, probabil ar conduce la sisteme mai performante. Pentru SER, sunt necesare cercetări suplimentare pentru dezvoltarea unor sisteme de recunoaștere automată independente de limbă.

Bibliografie

[Atm20] B.T. Atmaja și M. Akagi, "Deep multilayer perceptrons for dimensional speech emotion recognition," în *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Auckland, Noua Zeelandă, pp. 325-331, Dec. 2020.

[Avi19] A.R. Avila et al., "Speech-based stress classification based on modulation spectral features and convolutional neural networks," în *Proc. European Signal Processing Conference* (*EUSIPCO*), A Coruna, Spania, pp. 1-5, Sep. 2019.

[Bac95] J.A. Bachorowski și M.J. Owren, "Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context," în *Psychological Science*, vol. 6, nr. 4, pp. 219-224, Iul. 1995.

[Bac99] J.A. Bachorowski, "Vocal expression and perception of emotion," în *Current Directions in Psychological Science*, vol. 8, nr. 2, pp. 53-57, Apr. 1999.

[Bea18] R. Beard et al., "Multi-modal sequence fusion via recursive attention for emotion recognition," în *Proc. Conference on Computational Natural Language Learning* (*CoNLL*), Bruxelles, Belgia, pp. 251-259, Oct. 2018.

[Bes16] S. Besbes și Z. Lachiri, "Multi-class SVM for stressed speech recognition," în *Proc. International Conference on Advanced Technologies for Signal and Image Processing* (*ATSIP*), Monastir, Tunisia, pp. 782-787, Mar. 2016.

[Bis06] C. Bishop, *Pattern Recognition and Machine Learning*, Ed. I, Springer, New York, New York, Statele Unite ale Americii, 2006.

[Bis13] I. Bisio et al., "Gender-driven emotion recognition through speech signals for ambient intelligence applications," în *IEEE Transactions on Emerging Topics in Computing*, vol. 1, nr. 2, pp. 244-257, Dec. 2013.

[Bur05] F. Burkhardt et al., "A database of German emotional speech," în *Proc. INTERSPEECH*, Lisabona, Portugalia, pp. 1517-1520, Sep. 2005.

[Bus08] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," în *Language Resources & Evaluation*, vol. 42, nr. 4, art. nr. 335, Noi. 2008.

[Cao14] H. Cao et al., “CREMA-D: crowd-sourced emotional multimodal actors dataset,” în *IEEE Transactions on Affective Computing*, vol. 5, nr. 4, pp. 377-390, Dec. 2014.

[Cas06] S. Casale, A. Russo, și S. Serrano, “Classification of speech under stress using features selected by genetic algorithms,” în *Proc. European Signal Processing Conference* (*EUSIPCO*), Florența, Italia, pp. 1-5, Sep. 2006.

[Cas08] S. Casale et al., “Speech emotion classification using machine learning algorithms,” în *Proc. IEEE International Conference on Semantic Computing*, Santa Monica, California, Statele Unite ale Americii, pp. 158-165, Aug. 2008.

[Cha14] T. Chaspari, D. Dimitriadis, și P. Maragos, “Emotion classification of speech using modulation features,” în *Proc. European Signal Processing Conference* (*EUSIPCO*), Lisabona, Portugalia, pp. 1552-1556, Sep. 2014.

[Che18] M. Chen et al., “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” în *IEEE Signal Processing Letters*, vol. 25, nr. 10, pp. 1440-1444, Oct. 2018.

[Esp11] M. Espi, “Using spectral fluctuation of speech in multi-feature HMM-based voice activity detection,” în *Proc. INTERSPEECH*, Florența, Italia, pp. 2613-2616, Aug. 2011.

[Fah20] S. Fahad et al., “DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features,” în *Circuits, Systems, and Signal Processing*, vol. 40, nr. 1, pp. 466-489, Iul. 2020.

[Fat21a] E.P. Fathima Bareeda, B.S. Shajee Mohan, și K.V. Ahammed Muneer, “Lie detection using speech processing techniques,” în *Journal of Physics: Conference Series*, vol. 1921, pp. 12-28, Mar. 2021.

[Fuj10] M. Fujimoto, S. Watanabe, și T. Nakatani, “Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight normalization,” în *Proc. INTERSPEECH*, Makuhari, Chiba, Japonia, pp. 3102-3105, Sep. 2010.

[Fuj14] H. Fujimura, “Simultaneous gender classification and voice activity detection using deep neural networks,” în *Proc. INTERSPEECH*, Singapore, pp. 1139-1143, Sep. 2014.

[Gha19] E. Ghaleb, M. Popa, și S. Asteriadis, “Multimodal and temporal perception of audio-visual cues for emotion recognition,” în *Proc. International Conference on Affective Computing and Intelligent Interaction* (*ACII*), Cambridge, Marea Britanie, pp. 552-558, Sep. 2019.

[Gha20] E. Ghaleb, M. Popa, și S. Asteriadis, “Metric learning-based multimodal audio-visual emotion recognition,” în *IEEE MultiMedia*, vol. 27, nr. 1, pp. 37-48, Mar. 2020.

[Goo16] I. Goodfellow, Y. Bengio, și A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, Statele Unite ale Americii, 2016.

[Han98] J.H.L. Hansen et al., *Getting started with the SUSAS: Speech Under Simulated and Actual Stress database*, Raport tehnic RSPL-98-10, Robust Speech Processing Laboratory, Universitatea Duke, Durham, Statele Unite ale Americii, Apr. 1998.

[Has06] T. Hastie, R. Tibshirani, și J. Friedman, *The Elements of Statistical Learning*, Ed. a II-a, Springer, New York, New York, Statele Unite ale Americii, 2006.

[He09] L. He et al., "Stress detection using speech spectrograms and sigma-pi neuron units," în *Proc. International Conference on Natural Computation*, Tianjian, China, pp. 260-264, Aug. 2009.

[He15] L. He, Y. Bo, și G. Zhao, "Speech-oriented negative emotion recognition," în *Proc. Chinese Control Conference* (*CCC*), Hangzhou, China, pp. 3553-3558, Iul. 2015.

[Jai16] M. Jaiswal, S. Tabibu, și R. Bajpai, "The truth and nothing but the truth: multimodal analysis for deception detection," în *Proc. IEEE International Conference on Data Mining Workshops* (*ICDMW*), Barcelona, Spania, pp. 938-943, Dec. 2016.

[Jin15] Q. Jin et al., "Speech emotion recognition with acoustic and lexical features," în *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Brisbane, Queensland, Australia, pp. 4749-4753, Apr. 2015.

[Kan21] S. Kanwal și S. Asghar, "Speech emotion recognition using clustering-based GA-optimized feature set," în *IEEE Access*, vol. 9, pp. 125830-125842, Sep. 2021.

[Ker19] L. Kerkeni et al., "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," în *Speech Communication*, vol. 114, pp. 22-35, Noi. 2019.

[Kit07] N. Kitaoka, K. Yamamoto, și T. Kusamizu, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," în *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding* (*ASRU*), Kyoto, Japonia, pp. 607-612, Dec. 2007.

[Kop19] D. Kopev et al., "Detecting deception in political debates using acoustic and textual features," în *Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), Singapore, pp. 652-659, Dec. 2019.

[Lat20a] S. Latif et al., "Augmenting generative adversarial networks for speech emotion recognition," în *Proc. INTERSPEECH*, Shanghai, China, pp. 521-525, Oct. 2020.

[Laz99] R.S. Lazarus, *Stress and Emotion: A new synthesis*, Ed. I, Springer, New York, New York, Statele Unite ale Americii, 1999.

[Li07] X. Li et al., "Stress and emotion classification using jitter and shimmer features," în *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Honolulu, Hawaii, Statele Unite ale Americii, pp. 1081-1084, Apr. 2007.

[Liu20] S. Liu et al., "Hierarchical component-attention based speaker turn embedding for emotion recognition," în *Proc. International Joint Conference on Neural Networks* (*IJCNN*), Glasgow, Marea Britanie, pp. 1-7, Iul. 2020.

[Liu21] N. Liu et al., "Transfer subspace learning for unsupervised cross-corpus speech emotion recognition," în *IEEE Access*, vol. 9, pp. 95925-95937, Iul. 2021.

[Lot17] R. Lotfidereshgi și P. Gournay, "Biologically inspired speech emotion recognition," în *Proc. IEEE International Conference on Acoustics Speech, and Signal Processing* (*ICASSP*), New Orleans, Louisiana, Statele Unite ale Americii, pp. 5135-5139, Mar. 2017.

[Mat09] D. Matsumoto, *The Cambridge Dictionary of Psychology*, Ed. I, Cambridge University Press, Cambridge, Marea Britanie, 2009.

[Men17] G. Mendels et al., "Hybrid acoustic-lexical deep learning approach for deception detection," în *Proc. INTERSPEECH*, Stockholm, Suedia, pp. 1472-1476, Aug. 2017.

[Mih19a] S. Mihalache, D. Burileanu, G. Pop, și C. Burileanu, "Modulation-based speech emotion recognition with reconstruction error feature expansion," în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, România, pp. 1-6, Oct. 2019.

[Mih19b] S. Mihalache, G. Pop, și D. Burileanu, "Introducing the RODeCAR database for deceptive speech detection," în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, România, pp. 1-6, Oct. 2019.

[Mih20] S. Mihalache și D. Burileanu, *Speech emotion recognition for dissimulated behavior monitoring in surveillance applications*, Raport final, Proiect 2SOL/2017 – PN-III-P2-2.1-SOL-2016-02-0002, *Intelligent Systems for Video and Audio Analysis – Technologies and Innovative Video Systems for Person Re-identification and Analysis of Dissimulated Behavior* (*SPIA-VA*), Apr. 2020.

[Mih21a] S. Mihalache, I.A. Ivanov, și D. Burileanu, "Deep neural networks for voice activity detection," în *Proc. International Conference on Telecommunications and Signal Processing* (*TSP*), Brno, Cehia, pp. 191-194, Iul. 2021.

[Mih21b] S. Mihalache, D. Burileanu, și C. Burileanu, "Detecting psychological stress from speech using deep neural networks and ensemble classifiers," în *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), București, România, pp. 74-79, Oct. 2021.

[Mih21c] S. Mihalache și D. Burileanu, "Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition," în *University Politehnica of Bucharest Scientific Bulletin, Series C*, vol. 83, nr. 4, pp. 137-148, Dec. 2021.

[Mih22a] S. Mihalache și D. Burileanu, "Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," în *Sensors*, vol. 22, nr. 3, art. nr. 1228, pp. 1-21, Feb. 2022.

[Mih22b] S. Mihalache, D. Burileanu, E. Franți, M. Dascălu, și C.A. Brătan, "Lasting emotions – An investigation of short- and long-term affective content remanence in speech," în *Romanian Journal of Information Science and Technology*, vol. 25, nr. 1, pp. 20-35, Mar. 2022.

[Mir17] S. Mirsamadi, E. Barsoum, și C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," în *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), New Orleans, Louisiana, Statele Unite ale Americii, pp. 2227-2231, Mar. 2017.

[Mor12] G. S. Morrison, P. Rose, și C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," în *Australian Journal of Forensic Sciences*, vol. 44, nr. 2, pp. 155-167, Iun. 2012.

[Pan20] Z. Pan et al., "Multi-modal attention for speech emotion recognition," în *Proc. INTERSPEECH*, Shanghai, China, pp. 364-368, Oct. 2020.

[Per15] V. Perez-Rosas et al., "Deception detection using real-life trial data," în *Proc. ACM on International Conference on Multimodal Interaction*, New York, New York, Statele Unite ale Americii, pp. 59-66, Noi. 2015.

[Rah12] T. Rahman și C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," în *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Kyoto, Japonia, pp. 5117-5120, Mar. 2012.

[Rao17] W. Rao et al., "Investigation of fixed-dimensional speech representations for real-time speech emotion recognition system," în *Proc. International Conference on Orange Technologies* (*ICOT*), Singapore, pp. 197-200, Dec. 2017.

[Sch14] B. Schuller et al., "The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load," în *Proc. INTERSPEECH*, Singapore, pp. 427-431, Sep. 2014.

[Sen22] U.M. Sen et al., "Multimodal deception detection using real-life trial data," în *IEEE Transactions on Affective Computing*, vol. 13, nr. 1, pp. 306-319, Mar. 2022.

[Sha23b] D. Sharma et al., "Speech emotion recognition system using SVD algorithm with HMM model," în *Proc. International Conference for Advancement in Technology* (*ICONAT*), Goa, India, pp. 1-5, Ian. 2023.

[Shi20] H.K. Shin et al., "Speaker-invariant psychological stress detection using attention-based network," în *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Auckland, Noua Zeelandă, pp. 308-313, Dec. 2020.

[Su21] B.H. Su și C.C. Lee, "A conditional cycle emotion GAN for cross corpus speech emotion recognition," în *Proc. IEEE Spoken Language Technology Workshop* (*SLT*), Shenzhen, China, pp. 351-357, Ian. 2021.

[Tan21] D. Tang et al., "Adieu recurrence? End-to-end speech emotion recognition using a context stacking dilated convolutional network," în *Proc. European Signal Processing Conference* (*EUSIPCO*), Amsterdam, Netherlands, pp. 1-5, Ian. 2021.

[Vas15] J.C. Vasquez-Correa et al., "Emotion recognition from speech under environmental noise conditions using wavelet decomposition," în *Proc. International Carnahan Conference on Security Technology* (*ICCST*), Taipei, Taiwan, pp. 247-252, Sep. 2015.

[Vel19] A. Velichko et al., “Applying ensemble learning techniques and neural networks to deceptive and truthful information detection task in the flow of speech,” în I. Kotenko et al. (Eds.) *Intelligent Distributed Computing XIII. Studies in Computational Intelligence*, vol. 868, Springer, Cham, Elveția, pp. 477-482, Oct. 2019.

[Ver09] B. Verschuere, V. Prati, și J. De Houwer, “Cheating the lie detector,” în *Journal of Psychological Science*, vol. 20, nr. 4, pp. 410-413, Apr. 2009.

[Vil12] G. Villar, J. Arciuli, și D. Mallard, “Use of ‘um’ in the deceptive speech of a convicted murderer,” în *Applied Psychoacoustics*, vol. 33, nr. 1, pp. 83-95, Ian. 2012.

[War13] A.B. Warriner, V. Kuperman, și M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 English lemmas,” în *Behavior Research Methods*, vol. 45, nr. 4, pp. 1191-1207, Dec. 2013.

[Wat99] D. Watson et al., “The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence,” în *Journal of Personality and Social Psychology*, vol. 76, nr. 5, pp. 820-838, Mai 1999.

[Yi22] L. Yi și M.W. Mak, “Improving speech emotion recognition with adversarial data augmentation network,” în *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, nr. 1, pp. 172-184, Ian. 2022.

[Yil21] S. Yildirim, Y. Kaya, și F. Kilic, “A modified feature selection method based on metaheuristic algorithms for speech emotion recognition,” în *Applied Acoustics*, vol. 173, art. nr. 107721, Feb. 2021.

[Zao14] L. Zao, D. Cavalcante, și R. Coelho, “Time-frequency feature and AMS-GMM mask for acoustic emotion classification,” în *IEEE Signal Processing Letters*, vol. 21, nr. 5, pp. 620-624, Mai 2014.

[Zha18a] Y. Zhang et al., “Attention based fully convolutional network for speech emotion recognition,” în *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Honolulu, Hawaii, Statele Unite ale Americii, pp. 1771-1775, Noi. 2018.

[Zha18b] H. Zhao, N. Ye, și R. Wang, “Transferring age and gender attributes for dimensional emotion prediction from big speech data using hierarchical deep learning,” în *Proc. IEEE International Conference on Big Data Security on Cloud* (*BigDataSecurity*), *IEEE International Conference on High Performance and Smart Computing* (*HSPC*), *and IEEE International Conference on Intelligent Data and Security* (*IDS*), Omaha, Nebraska, Statele Unite ale Americii, pp. 20-24, Mai 2018.

[Zha19] Z. Zhao et al., “Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition,” în *IEEE Access*, vol. 7, pp. 97515-97525, Iul. 2019.

[Zha20] J. Zhang, S.I. Levitan, și J. Hirschberg, “Multimodal deception detection using automatically extracted acoustic, visual, and lexical features,” în *Proc. INTERSPEECH*, Shanghai, China, pp. 359-363, Oct. 2020.

[Zha21a] J. Zhang et al., “Cross-corpus speech emotion recognition using joint distribution adaptive regression,” în *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Toronto, Ontario, Canada, pp. 3790-3794, Iun. 2021.

[Zha22] W. Zhang et al., “Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression,” în *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, nr. 2, pp. 588-598, Iun. 2022.