Universitatea Politehnica din București

Facultatea de Automatică și Calculatoare

Ingineria și Managementul Sistemelor de Afaceri

LUCRARE DE DIZERTAȚIE

SISTEM DE RECUNOAȘTERE A AMPRENTEI VOCALE

Coordonator,

Prof. Dr. Ing. Mihai Caramihai

Masterand,

Theodor Adrian Macovei

București 2020

Cuprins

1. Prezentarea scopului lucrării și a etapelor de lucru

Vocea a reprezentat un importat factor evolutiv în parcursul omului, pornind, la fel ca la alte specii, ca mijloc de avertizare și ajungând să devină principalul mod de comunicare, înlesnind formarea legăturilor sociale. Deși, așa cum am menționat anterior, există numeroase specii capabile să producă sunete, această abilitate este mult mai pronunțată la oameni, aceștia fiind capabili de o gamă mult mai largă de modulații. Datorită diferețelor biologice, organele implicate în acest proces au forme și caracteristici care variază de la individ la individ, fapt ce face ca produsul finit, vocea, să fie unic, iar amprenta vocală una dintre cele mai de încredere metode de identificare biometrică.

Pe baza analizei unei mostre de semnal sonor, prin diverse tehnici se pot obține numeroase informații: mesajul și limba (i.e. informație lingvistică), precum și persoana, sexul, vârsta și emoția acestei persoane (i.e. informație paralingvistică). În cazul recunoașterii persoanei, se disting două abordări: identificare și verificare. Verificarea constă într-un proces decizional ce stabilește daca semnalul investigat aparține, într-adevăr, persoanei care face cererea, în timp ce identificarea, care va fi tratată în lucrarea de față, constă în recunoașterea unei persoane dintr-un set de date existent. Cu alte cuvinte, asignarea vocii acelei persoane din baza de date a cărei probabilitate în urma clasificării este cea mai mare.

Scopul acestei lucrări constă în dezvoltarea unei aplicații informatice capabile să recunoască și să clasifice o amprentă vocală (să recunoască vorbitorul dintr-un set de date existent), precum și emoția acestuia.

Prima parte a lucrării are rolul de a descrie elementele fundamentale ce privesc semnalul vocal. Astfel, vor fi prezentate fenomenele de producere și de percepție ale vorbirii, precum și metode de modelare și de procesare ale semnalului sonor. A doua parte a lucrării va fi dedicată analizei vocale și a metodelor de modelare a caracteristicilor vocii. Această parte va detalia și modul în care pot fi recunoscute emoții pe baza vorbirii. A treia parte va consta într-o analiză comparativă a soluțiilor informatice existente ce privesc un sistem de recunoaștere vocală, cât și metricile folosite în evaluarea acestor soluții. Ultima parte a lucrării va fi dedicată implementării aplicației și a prezentării rezultatelor obținute.

2. Elemente fundamentale privind semnalul vocal

2.1. Percepția vorbirii

Pentru a putea înțelege mai bine soluțiile informatice ce implementează identificarea vocală este utilă înțelegerea felului în care omul procesează și analizează aceste semnale.

Sistemul auditiv uman este acel sistem responsabil cu conversia sunetelor din variații ale presiunii aerului în semnale pentru nervul acustic. Sunetul este captat de pavilionul urechii, direcționat către urechea medie unde este filtrat, iar apoi către urechea internă unde este convertit în semnale ce vor ajunge la creier [1]. Urechea internă are structura unei cochilii de melc, cu celule (denumite cili) ce rezonează la frecvențe diferite (de la frecvențe înalte la exterior către frecvențe din ce în ce mai joase către interior). Acest fenomen este ilustrat în figura de mai jos, figură în care cochilia este desfășurată:

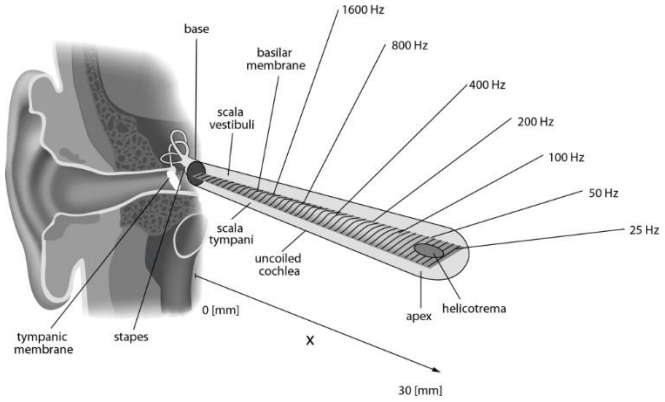


Figura 1 - Urechea internă. Recepționarea frecvențelor [2]

În studiul de față ne interesează cu precădere percepția auditivă și mai puțin componenta biologică a sistemului auditiv, deci modelarea cantitativă a percepției auditive umane. Cu studiul acestui domeniu se ocupă psihoacustica [3], iar unele dintre caracteristicile importante ale acestui domeniu, precum și teorii existente vor fi detaliate mai jos.

Conform [4], procesarea temporală poate fi definită ca fiind "percepția sunetelor ori a alterării acestora (i.e. a stimultilor sonori) într-un cadru de timp definit". Acest concept se consideră ca având o importanță ridicată în majoritatea abilităților auditive, întrucât cele mai multe caracteristici ale semnalului sonor sunt dependente mai mult sau mai puțin de timp (spre

exemplu înțelegerea vorbirii atât în condiții ideale, cât și de zgomot). Sistemul auditiv uman are o rezoluție temporală de 1-2 milisecunde, ce poate fi extinsă până la 0.25 ms [5]. Dincolo de această limită, înțelegerea stimulilor continui nu mai poate fi făcută.

Numeroase teorii ale percepției sonore au fost dezvoltate de-a lungul anilor și continuă să se dezvolte pe măsură ce noi informații apar în domeniu. Aceste teorii pot fi clasificate [6] în: active versus pasive, abordate de sus în jos versus abordate de jos în sus și autonome versus interactive.

Teoriile active pun accentul pe legăturile dintre fenomenul de producere al vocii și cel de percepție (pe proprietățile comune ale acestora). În aceste teorii, cunoașterea de către individ a felului în care se produc sunetele este un factor esențial în recunoașterea vorbirii. Teoriile pasive susțin exact contrariul, faptul că aceste cunoașteri joacă un rol minor, iar accentul cade pe aspectele senzoriale.

Abordarea bottom-up se bazează pe premisa că toată informația necesară recunoașterii sunetelor este conținută în semnalul sonor, iar ascultătorul nu este implicat în procese de cogniție și lingvistice în decodificarea mesajului. Abordarea top-down susține faptul că o importanță foarte mare o au procesele lingvistice și de cogniție.

Teoriile active susțin faptul că mesajul este decodificat într-o ordine (fonetic, lexical, sintactic, semantic etc.), în timp ce teoriile interactive consideră că informația se obține la fiecare dintre aceste etape.

2.2. Metode de modelare a semnalului sonor

În literatura de specialitate, vorbirea este adesea descrisă ca fiind un semnal sonor ce poartă informație lingvistică [7]. Deși această lucrare nu tratează și domeniul psiholingvisticii, trebuie totuși menționat faptul că vorbirea nu este un fenomen care pornește din plămâni, ci din creier. Astfel, mesajul, precum și structura lexico-gramaticală a acestuia se formează în creier, de unde pornesc și comenzile către organele implicate în acest proces [8]. În continuare, ne vom axa doar pe structura și caracteristicile aparatului de vorbire, de la generarea fluxului de aer din plămâni și până la emiterea acestuia prin cele două cavități: orală și nazală.

Principalele elemente componente ale aparatului sunt plamânii, laringele, traheea, faringele, cavitatea bucală și cavitatea nazală. Diafragma împinge aerul din plămâni către laringe, iar prin vibrația corzilor vocale se obține un semnal periodic, ce va fi în continuare variat de restul elementelor: limbă, dinți, bolta palatină, buze etc. Cavitatea orală și cavitatea nazală au rolul de a amplifica și de a emite sunetul format. Elementele enumerate anterior sunt ilustrate în figura de mai jos.

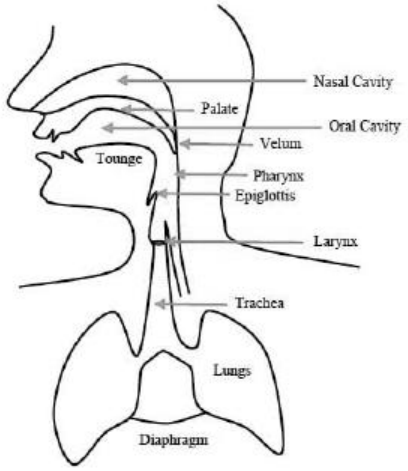


Figura 2 - Aparatul fonator [9]

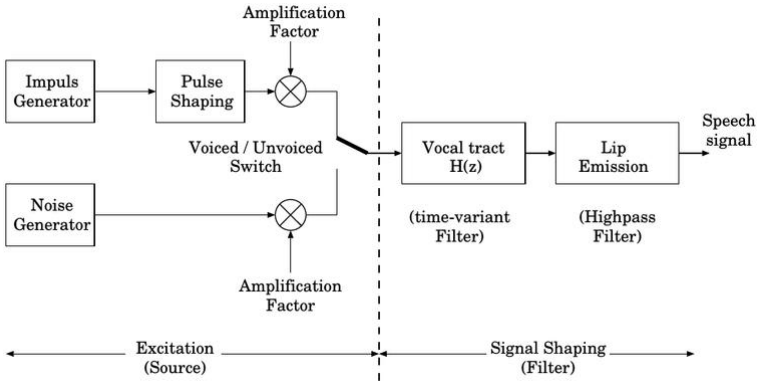


Figura 3 - Abordare sistemică a vorbirii [11]

Datorită diferențelor biologice, organele și caracteristicile enumerate anterior variază de la individ la individ, fapt ce face ca produsul finit, vocea, să fie unic. Astfel, vocea are mai mult de o sută de caracteristici distincte care fac ca amprenta vocală să fie una dintre cele mai de încredere metode biometrice [10].

Din punct de vedere sistemic, putem privi aparatul vocal așa cum este descris în figura 3.

Schema anterioară împarte procesul în două subprocese: cel de excitație (în care semnalul este generat) și cel de filtrare. Întrucât sunetele umane pot consta ori în vorbire, ori în zgomot, există două ramuri pentru producerea sunetelor. Aceste semnale produse sunt ulterior amplificate. Comutatorul din figură modeleaza alegerea sunetului ce va urma a fi scos. Partea de filtrare este compusă din tubul vocal (care din punct de vedere sistemic poate fi modelat ca un filtru variant în timp) și din partea de emitere a semnalului, buzele, care pot fi modelate ca un filtru trece-sus.

2.3. Procesarea vorbirii: spectrograme

Așa cum am menționat de-a lungul acestei lucrări, putem privi sunetul ca fiind un semnal continuu. Cu toate acestea, un calculator va obține acest semnal doar prin conversia de la analog la digital. În implementarea metodelor de identificare a amprentei vom avea nevoie cu lucrul cu eșantioanele semnalului, de aceea fiind utilă analiza semnalelor discrete.

Pentru analiza conținutului în frecvență al unui semnal discret x[n], de durată finită, ce conține N eșantioane se aplică Transformata Fourier Discretă (DFT):

$$\hat{x}[k] = \sum_{n=0}^{N-1} x[n] e^{-i\, n \frac{2\pi k}{N}} \quad , \qquad k = \overline{0, N-1}$$

Această transformată se aplică pentru a ajunge din domeniul timp ăn domeniul frecvență. Pentru a ajunge înapoi în domeniul timp se aplică Transformata Fourier Discretă Inversă:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{x}[k] e^{-n\, i \frac{2\pi k}{N}}, \quad n = \overline{0, N-1}$$

În formă matriceală, cele două ecuații de mai sus pot fi rescrise sub forma [12]:

$$x = \frac{1}{N} F \hat{x}$$

$$\hat{x} = \overline{F} x$$

, unde matricea F este matricea Fourier dată mai jos, iar $\overline{F}$ conjugata complxă a acesteia.

$$F = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{i\frac{2\pi}{N}} & \dots & e^{i2\pi\frac{N-1}{N}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{i2\pi\frac{N-1}{N}} & \dots & e^{i2\pi\frac{(N-1)^2}{N}} \end{bmatrix}$$

Pentru semnalul x de lungime N considerăm partiționarea acestuia în segmente de lungime *m,* unde *m<<n*. Fie $X \in R^{m*(N-m+1)}$ matricea ce conține pe coloane segmentele (prima coloană va fi *[x[0], x[1], ..., x[m-1]]*^{T}, cea de a doua *[x[1], x[2], ..., x[m]]*^{T} etc.). Se poate observa că matricea X este o reprezentare cu grad ridicat de redundanță a lui x.

Spectrograma lui x va fi matricea $\hat{X}$, ale cărei coloane sunt transformatele Fourier discrete ale coloanelor lui X. În lucrarea "Acoustics of Speech and hearing"[13], spectrograma este definită ca fiind "graficul conținutului de energie al unui semnal, reprezentat ca funcție de frecvență și timp", aplitudinea fiind reprezentată prin intensitatea culorii (o scală de nuanțe de griuri, 0 corespunde culorii negre, în timp pe 255 corespunde culorii alb).

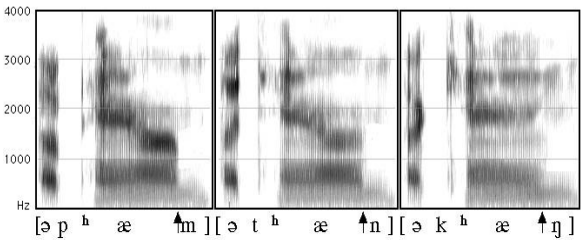


Figura 4 - Spectrograme ale diferitelor sunete [14]

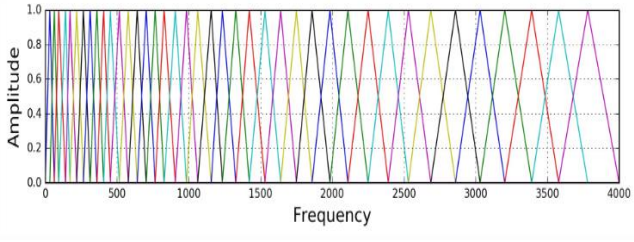


Figura 5 - Filtre Mel [16]

Cea mai cunoscută și mai aplicată metodă de recunoaștere vocală se bazează pe indicii Mel Cepstrali. Algoritmul constă în partiționarea semnalului în scurte segmente și obținerea spectrului pentru fiecare astfel de segment, filtrarea cu un filtru Mel și aplicarea analizei cepstrale pentru a obține coleficienții Mel cepstrali de frecvență. Aceștia se folosesc pentru recunoașterea vorbirii. Analiza Mel se bazează pe teoriile percepției și pe experimente ce au demonstrat faptul că urechea acționează ca un filtru ce se concentrează doar pe anumite componente de frecvență. Aceste filtre nu sunt uniform distribuite pe axa frecvenței, fiind mai dese in zona frecvențelor joase [15]. Acest aspect poate fi mai ușor înțeles dacă ne întoarcem la figura 1 și observăm felul în care urechea internă percepe frecvențele. Filtrele Mel sunt o serie de filtre triunghiulare trece-bandă ce se suprapun, uniform distribuite pe scara Mel. În practică se folosesc uzual între 13 și 40 de astfel de filtre ce acoperă banda de frecvență de 0-8000 Hz. Aceste filtre pot fi observate în figura 5.

3. Analiza vocală. Pattern recognition

3.1. Preprocesarea semnalului vocal

Procesul de identificare bazat pe amprenta vocală constă în achiziția unei mostre sonore, extragerea caracteristicilor acelei mostre și găsirea, într-o bază de date, a persoanei cu caracteristicile cele mai apropiate de cele ale mostrei.

În cadrul etapei de achiziție nu se poate obține niciodată un semnal curat, acesta fiind întotdeauna afectat de zgomot. Zgomotul care se compune cu semnalul vocal are mai multe cauze și nu poate fi evitat. De aceea, pentru ca analiza semnalului să nu fie influențată de acesta se impune o etapă de preprocesare după achiziție, etapă în care, printre altele se realizează și filtrarea zgomotului.

Dacă privim zgomotul ca fiind un semnal aditiv ce se adaugă semnalului sonor, atunci putem privi mostra achiziționată sub forma:

$$y(t) = x(t) + d(t)$$

, unde y(t) reprezintă semnalul achiziționat, format din semnalul vocal pur x(t) și zgomotul d(t).

Merita aici menționată și o altă categorie de zgomote, în afară de cele aditive: există situații în care zgomotul nu este un semnal care se suprapune peste semnalul original, ci este un filtru liniar, caz în care semnalul final, mostra sonoră este o convoluție a zgomotului cu semnalul vocal.

În lucrarea “Preprocessing Technique in Automatic Speech Recognition for Human Computer Interaction: An Overview” [17] sunt prezentate un număr de tehnici de procesare dintre care le am prezentat mai jos pe cele mai importante.

Legat de zgomotul ce afectează semnalul sonor putem menționa o metrică folosită în inginerie: SNR (Signal to Noise Ratio). Această metrică este un raport, adesea exprimat în decibeli, între puterea semnalului vocal și puterea zgomotului:

$$SNR = 20 \log \frac{Vsignal}{Vnoise}$$

Cu cât SNR este mai mic, cu atât mostra achiziționată este mai afectată de zgomot și nu poate fi procesată ulterior. De aceea, se recomandă utilizarea unor semnale cu un SNR cât mai mare.
ZCR (Zero Crossing Rate) este o mărime fizică ce exprimă numărul de variații (de câte ori semnalul variază de la pozitiv la negativ) ale semnalului într-un cadru de timp și este dată de ecuația:

$$Z(i) = \frac{1}{2N}\sum_{n=1}^{N} |sgn[x_i(n)] - sgn[x_i(n-1)]|$$

, unde *sgn ()* reprezintă signatura:

$$sgn(x) = \begin{cases} 1, dacă\ x \geq 0 \\ -1, dacă\ x < 0 \end{cases}$$

Filtrele Wiener merită amintite aici, fiind, conform "Advanced Digital Signal Processing and Reduction" [18], filtre atât cu răspuns finit la impuls (FIR), cât și cu răspuns infinit (IIR) folosite pentru reducerea zgomotului, cu relația intrare-ieșire dată în formula de mai jos:

$$\hat{x}(m) = \sum_{k=0}^{P-1} w_k\, y(m-k) = \boldsymbol{w}^T \boldsymbol{y}$$

unde m este indexul de timp discret, y vectorul de intrări ale filtrului, iar vectorul w este vectorul ce conține coeficienții filtrului Wiener. Eroarea *e* se definește ca fiind diferența dintre semnalul dorit x și ieșirea obținută $\hat{x}$:

$$e(m) = x(m) - \hat{x}(m) = x(m) - \boldsymbol{w}^T \boldsymbol{y}$$

Aceste se bazează pe minimizarea erorii pătratice, eroare definită anterior.
După ce semnalul este împărțit în segmente de dimensiuni reduse, fiecare astfel de cadru este înmulțit cu o fereastră w(n) de lungime N. Această operație are rolul de a evidenția caracteristicile semnalului.
Din lucrarea "The Handbook of Formulas and Tables for Signal Processing" [19] am extras câteva ferestre, acestea fiind definite mai jos:

- Fereastră dreptunghiulară

$$w(n) = 1, 0 \leq n \leq M - 1$$

- Fereastră triunghiulară

$$w(n) = 1 - [1 - \frac{2n}{M-1}], 0 \leq n \leq M - 1$$

- Fereastră Hanning

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{M-1}\right), 0 \leq n \leq M - 1$$

- Fereastră Hamming

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right), 0 \leq n \leq M - 1$$

3.2. Tehnici de clasificare vocală. Rețele neuronale artificiale

3.2.1. Generalităţi

Una dintre soluțiile principale folosite la momentul de față în ceea ce privește recunoașterea amprentei vocale o constituie rețelele neuronale. Rețelele neuronale artificiale fac parte din categoria de algoritmi de învățare automata, mai precis învățare supervizată.

Învățarea automata (în engleză cunoscută drept Machine Learning- ML) reprezintă, conform [20] "știința prin care calculatoarele sunt programate astfel încât să poată învăța din date". Astfel, spre deosebire de algoritmi clasici, în care programatorul specifică explicit regulile după care programul să funcționeze, în cazul învățării automate algoritmul învață singur regulile după care trebuie să funcționeze. Principalele categorii de învățare automată sunt [21]: învățarea supravegheată (algoritmul deduce o funcție pe baza unor date existente de forma perechi intrări/ieșiri), învățarea nesupravegheată (algoritmul deduce un tipar al datelor, acestea nefiind etichetate) și învățarea prin consolidare(algoritmul este optimizat cu ajutorul recompenselor în cazul rezolvării corecte a problemei).

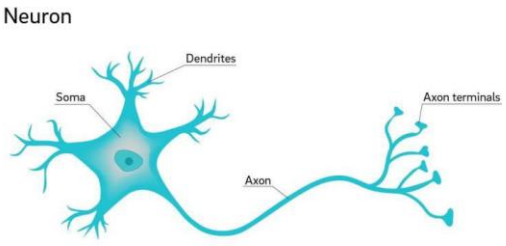
În cazul de față, problema pe care dorim să o rezolvăm este una de învățare suprvizată (algoritmul este antrenat cu un set de date, acestea fiind însoțite și de etichete ale valorilor

corecte), mai exact una de clasificare, întrucât dorim ca algoritmul să clasifice un semnal de voce umană într-una dintre categorii. Această problemă va fi rezolvată cu ajutorul rețelelor neuronale, a căror funcționare va fi detaliată mai jos.

Rețelele neuronale sunt definite, conform [22], ca fiind "un ansamblu de elemente simple de procesare interconectate, numite unități sau noduri, a căror funcționare se bazează pe funcționarea neuronilor ființelor vii".

Pentru o mai bună înțelegere a rețelelor neuronale artificiale și a diverselor arhitecturi ale acestora poate fi util studiul creierului din punct de vedere neurologic, dar în cazul acestei lucrări ceea ce merită dezvoltat este doar modelarea neuronului. Acest model este util pentru a putea înțelege funcționarea neuronilor din rețelele artificiale.

Neuronii (sau celulele nervoase) "consituie elementul de bază al creierului uman, fiind formați din dendrite, axoni, sinapse și corp celular". Funcționarea neuronului se bazează pe receptarea informației de la alte celule (prin dendrite), combinarea acestor informații într-o manieră neliniară și transmiterea rezultatului (prin terminațiile axonilor) către alte celule [23]. Aceste părți componente pot fi observate în figura 1.



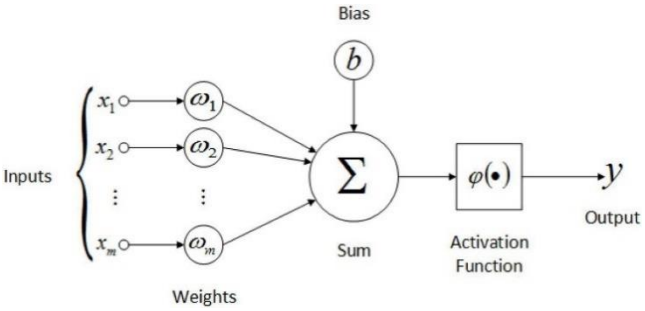


Figura 6- Structura unui neuron natural [24] *Figura 7- Structura unui neuron artificial[25]*

Neuronii artificiali sunt modelați astfel încât să replice structura și comportamentul neuronilor naturali. Astfel, fiecare neuron are un număr *m* de intrări ($x_1, x_2,...,x_m$), fiecare intrare fiind ponderată cu o anumită constantă ($w_1, w_2,...,w_m$). Acestea sunt însumate și trec printr-o funcție de activare al cărei scop este acela de a introduce neliniarități [26].

Această structură este prezentată în figura 3, alături de cea a neuronului natural. Se observă, astfel, structura similară a acestora.

Astfel, pe baza figurii 3 ne putem da seama că rezultatul pe care un astfel de neuron îl va produce va fi de forma:

$$y = f(\sum_{i=1}^{n} w_i x_i + b)$$

, unde y reprezintă ieșirea, w_i reprezintă ponderea, x_i intrarea, b bias-ul și f funcția de activare.

Așa cum am enunțat anterior, rolul funcțiilor de activare este acela de a introduce neliniarități în cadrul modelului. Acestea sunt folosite pentru o mai bună modelare (se permite, astfel, modelarea neliniarităților, în natură acestea fiind mult mai des întâlnite decât liniaritățile). Fără astfel de funcții de activare, rețeaua neuronală nu ar fi decât o regresie liniară.

Principalele funcții de activare folosite sunt Tanh, ReLU, Sigmoid si Liniar, acestea, alături de formula lor matematică și graficul asociat putând fi observate în figura de mai jos:

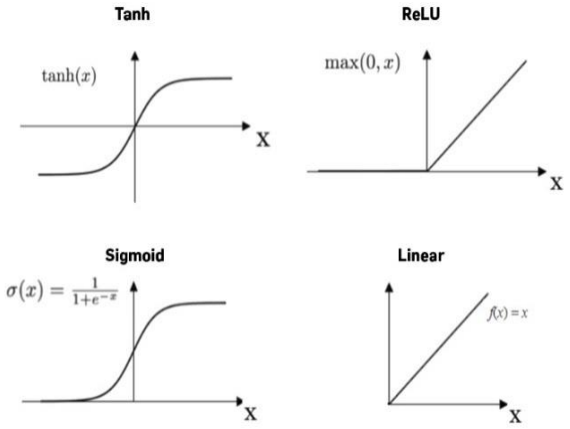


Figura 8- Principalele funcții de activare [27]

O rețea neuronală este alcătuită din mai multe niveluri conectate între ele, fiecare dintre acestea având un număr de neuroni. Figura 4 prezintă structura unei rețele neuronale artificiale. Se poate observa, astfel, faptul că rețeaua este compusă dintr-un nivel de intrare (cu un număr de neuroni egal cu numărul de intrări ale problemei), unul de ieșire (cu un număr de

neuroni egal cu numărul de ieșiri ce se doresc a fi obținute) și un număr de niveluri intermediare. Numărul de nivele intermediare, precum și numărul de neuroni din fiecare nivel intermediar sunt alese de cel care proiectează rețeaua în funcție de problema ce se dorește a fi rezolvată și complexitatea acesteia.

Astfel, la fel cum creierul uman are capacitatea de a își aminti și aplica experiențele anterioare în cadrul unora noi (datorită numărului mare de neuroni și de conexiuni ale acestora), același comportament se dorește a fi replicat și în cadrul rețelelor neuronale artificiale. Figura numărul 4 prezintă arhitectura unei astfel de rețele:

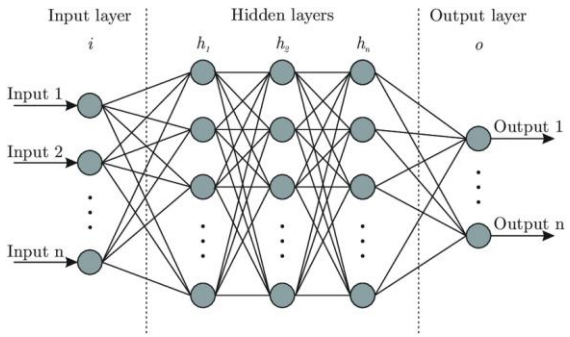


Figura 9 – Structura unei rețele neuronale artificiale [28]

3.2.2. Rețele neuronale convoluționale

Există numeroase tipuri de rețele neuronale artificiale, dar mai jos nu vor fi descrise decât cele convoluționale, întrucât sunt singurele care au fost folosite în cadrul acestei soluții.

Rețele Neuronale Convoluționale (în engleză Convolutional Neural Networks - CNN) reprezintă o categorie de rețele neuronale, folosite preponderent în acele aplicații în care se dorește extragerea de trăsături specifice (denumite în engleză *features*). Astfel, aceste rețele se pretează aplicațiilor ce lucrează cu imagini (ex. detecție și recunoaștere de obiecte, de fețe etc.).

CNN sunt compuse din cel puțin un strat de convoluție, acestea fiind de fapt seturi de filtre de mici dimensiuni, aplicabile local, straturi de pooling, al căror rol este acela de reducere a dimensionalității și straturi de activare, în care se determină impactul avut de filtre[29].

Deși numărul și ordinea de straturi de convoluție, de pool și de activare diferă de la arhitectură la arhitectură, fiecare astfel de rețea trebuie să conțină cel puțin un strat de convoluție, iar ultimul strat trebuie să fie unul unidimensional (un vector). În cazul aplicațiilor de clasificare, acesta este urmat de o funcție de activare care dă ieșirea y (mai exact probabilitatea asociată fiecărei ieșiri).

Stratul de convoluție reprezintă "înmultirea intrării cu o matrice de mici dimensiuni (adesea 2x2, 3x3 etc), prin glisarea matricei de-a lungul imaginii inițiale, cu scopul de a amplifica trăsăturile specifice, indiferent de poziția lor în imagine". [30]

În figura de mai jos se poate observa o arhitectură de rețea neuronală convoluțională, folosită pentru recunoașterea cifrelor scrise de mână.

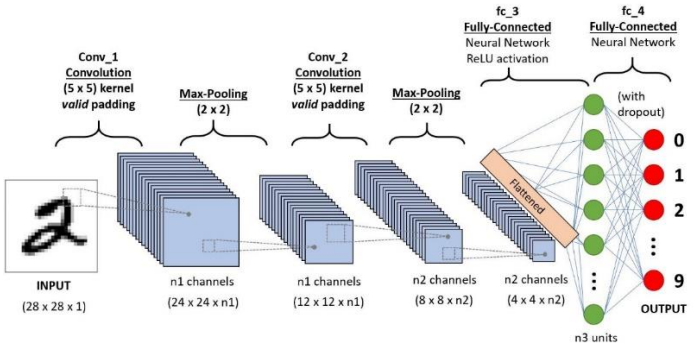


Figura 10- Structura unei rețele neuronale convoluționale[31]

Motivul pentru care acest tip de rețele neuronale pot fi aplicate în recunoașterea unei persoane ține de felul în care pot fi privite atât spectrogramele, cât și indicii Mel Cepstrali, i.e. matrice de două dimensiuni. În cazul indicilor MFCC, aceștia sunt matrice bidimensionale de forma N*M, unde N reprezintă numărul de ferestre de timp alese, iar M reprezintă numărul de coeficienți Mel aleși.

3.3. Recunoașterea emoțională

Până acum s-a discutat problema identificării unei persoane pe baza amprentei sale vocale. Această informație care se obține din semnalul sonor este una de natură paralingvistică. Emoțiile umane reprezintă, de asemenea, o altă informație ce poate fi extrasă din semnalul vocal, fără a fi nevoie de analiza mesajului propriu-zis. Putem menționa faptul că în prezent recunoașterea emoțiilor unui subiect se face în două moduri: prin analiza vocii și prin analiza expresiilor faciale, dar aceasta din urmă nu face obiectul acestui studiu. Dicționarul online al Asociației Americane de Psihologie (American Psychological Association) definește emoțiile ca fiind un model complex de reacții ce implică elemente de ține de experiențe, psihologie și comportament prin care un individ încearcă să facă față unui eveniment semnificativ personal, iar calitatea acelei emoții este dată de importanța acelui eveniment [32].

În psihologie și filosofie există numeroase abordări ce tratează emoțiile, cauzele acestora, diferențele dintre acestea și motivații sau sentimente, dar aceste abordări nu fac obiectul acestui studiu, ceea ce ne interesează fiind clasificările emoțiilor și câteva emoții care pot fi identificate ulterior. În lucrarea “Emotion: A Psychoevolutionary Synthesis” a lui Plutchik [33], acesta formulează zece postulate ale teoriei psihoevoluționiste a emoțiilor primare. Mare parte dintre aceste postulate țin de evoluționism și pot fi omise, dar acele postulate ce au un rol ulterior în cadrul proiectului afirmă că fiecare emoție poate exista în diverse grade de intensitate și nivele de excitație, iar toate emoțiile pot fi împărțite într-un număr mic de emoții primare (stări ideale, conceptualizate sub formă de perechi opuse), iar toate celelalte se obțin prin combinarea acestora. Plutchick consideră că emoțiile primare sunt în număr de opt și acestea sunt frica, furia, veselia, dezgustul, acceptarea, surpriza, prevederea și tristețea.

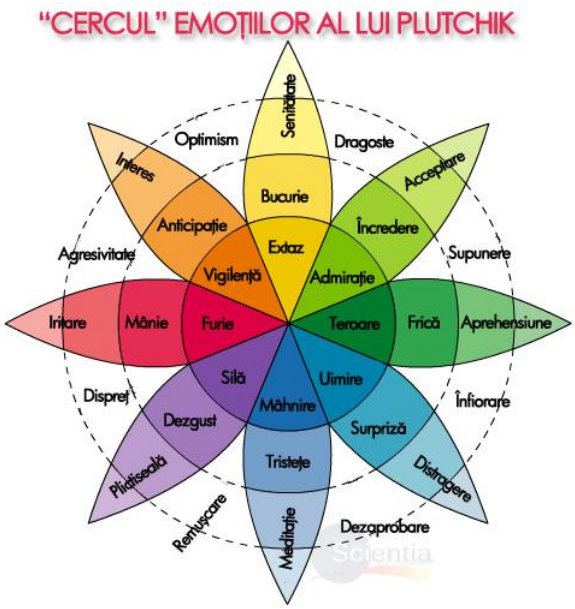


Figura 11 – Cercul emoțiilor al lui Plutchik [34]

În lucrarea "What's Basic about Basic Emotions?" [35] sunt enumerate diversele teorii privind emoțiile de bază, printre care și cea a lui Plutchik enumerată anterior. Vom enumera în continuare alți psihologi, precum și emoțiile pe care aceștia le consideră a fi primare: Weiner și Graham (1984) consideră doar două emoții fundamentale - fericirea și tristețea, Mowrer (1960) consideră durerea și plăcerea, Frijda (1986) consideră șase emoții fundamentale (dorința, fericirea, interesul, surpriza, mirarea și tristețea), iar Tomkins (1984) nouă (furia, interesul, disprețul, dezgustul, primejdia, frica, bucuria, rușinea și surpriza).

Aceste teorii vor fi necesare ulterior, atunci când pe baza analizei unui semnal vocal sistemul să poată recunoaște emoția (sistemul va avea un număr de emoții primare, tratate mai sus, iar mostra va fi comparată cu acestea și dată o categorie cât mai apropiată).

Pentru identificarea emoțiilor unei persoane, procesul este similar celui de identificare a persoanei.

În lucrarea "A Study of Speech Emotions Recognition Methods" [36] sunt menționate etapele ce trebuie parcurse pentru identificarea emoțiilor (arhitectura unui astfel de sistem): achiziția unei mostre sonore, ulterior sistemul de procesare al vocii extrage din eșantioane valori precum tonul sau energia, aceste marimi fizice sunt reduse la un set de caracteristici, iar un clasificator învață supervizat, cu exemple, cum să asocieze aceste caracteristici cu emoțiile și oferă o ieșire

cu emoția corespunzătoare. Această arhitectură este ilustrată și în figura de mai jos, preluată din aceeași sursă:

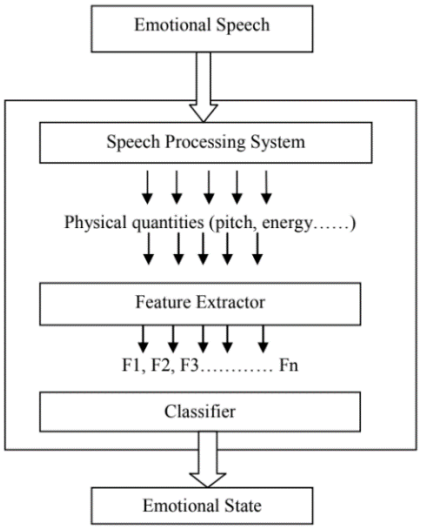


Figura 12 – Funcționarea unui sistem de recunoaștere a emoțiilor[36]

Caracteristicile ce pot fi obținute din semnalul vocal în cazul emoțiilor sunt aceleași ca în cazul identității persoanei și sunt caracteristici obținute din energia acelui semnal sau din tonul acestuia, prin analiza statistică a acestuia (și obținerea de mărimi precum media, mediana, covarianța, minimul, maximul etc). De asemenea și clasibicatorii ce pot fi folosiți sunt aceiași (SVM, HMM, KNN etc), deci se poate observa faptul că în cadrul implementării problema diferențierii identificării persoanei cu cea a emoțiilor se pune abia în faza de clasificare. Întrucât această lucrare este dedicată atât extragerii identității persoanei, cât și a emoțiilor acesteia, să considerăm câteva situații în care aceste informații pot fi aplicate împreună. Spre exemplu, în cazul în care recunoașterea unei persoane dintr-o bază de date ar fi înlocuită cu confirmarea identității acelei persoane, atunci amprenta vocală ar putea fi folosită ca tehnică biometrică pentru deblocarea unei case inteligente. În continuare, pe baza emoțiilor acelei persoane (și a persoanei din familie care a deblocat casa), ambientul poate fi modificat pentru a se potrivi cât mai bine stării acelei persoane (și aici ne referim la gradul de luminozitate, culoarea luminii, temperatura, muzica etc.). O altă situație ar putea fi aceea a unui autovehicul inteligent, de familie (deci cu mai mulți șoferi înregistrați în baza de date), care să asiste șoferul și să adjusteze stilul de condus în funcție de acesta și de emoțiile pe care acesta le are în acel moment.

4. Analiza comparativă

4.1. Metrici de evaluare

Matricea de confuzie reprezintă "o matrice pătratică, de dimensiune n x n, asociată unui clasificator ce ilustrează valorile prezise și cele corecte" [37].

Figura de mai jos prezintă o matrice de confuzie folosită pentru clasificarea binară, aceasta permițând o mai bună înțelegere a metricilor folosite. Pe baza acesteia se pot observa patru mărimi: numărul de predicții negative corecte (*Tn- True Negative*), numărul de predicții positive incorecte (*Fp- False Positive*), numărul de predicții negative incorecte (*Fn- False Negative*) și numărul de predicții positive corecte (*Tp- True Pozitive*).

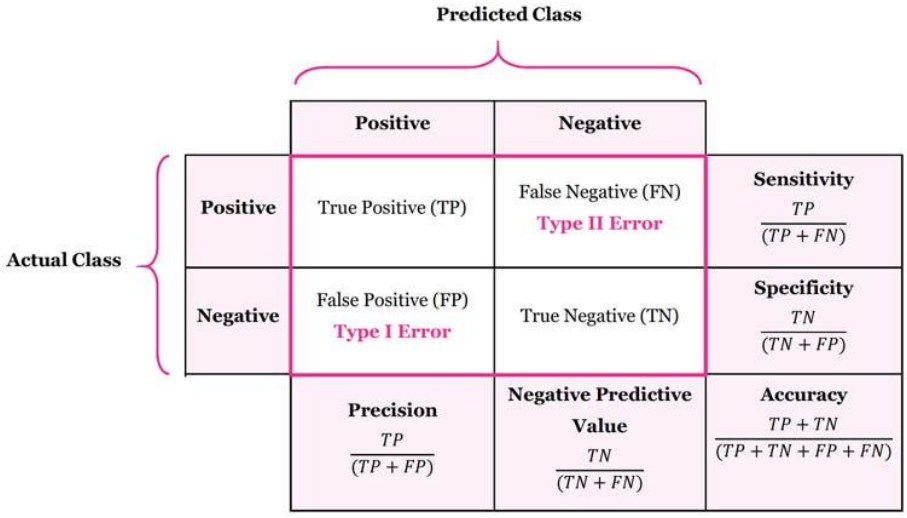


Figura 13- Matrice de confuzie pentru clasificarea binară [38]

Asfel, pe baza matricei de confizie se pot obține în mod direct următoarele metrici, sintetizate din lucrarea "A systematic analysis of performance measures for classification tasks" [39]:

- Senzitivitatea: $\frac{Tp}{Tp+Fn}$ Senzitivitatea oferă informații asupra eficienței cu care un clasificator identifică predicțiile pozitive corecte.
- Specificitatea: $\frac{Tn}{Tn+Fp}$ Specificitatea reprezintă analogul senzitivității, fiind eficiența cu care clasificatorul identifică predicțiile negative corecte.

- Precizia: $\frac{Tp}{Tp+Fp}$ Precizia este raportul dintre predicțiile pozitive corecte și numărul total de pre dicții pozitive. Principala problemă a preciziei o constituie faptul că aceasta nu ia în calcul predicțiile negative.
- Acuratețea: $\frac{Tp+Tn}{Tp+Tn+Fp+Fn}$ Acuratețea consitituie una dintre cele mai folosite metrici folosite pentru evaluarea unui model de clasificare, datorită simplității ei și a faptului că abordează toate cele patru elemente. Aceasta reprezintă procentul de predicții corecte obținute de model din numărul total de observații. O problemă ce apare în lucrul cu acuratețea o constituie faptul că două sisteme cu comportamente diferite, unul cu o rată slabă a detecției *Tp* și o rată puternică a *Tn*, iar celălalt cu o rată puternică de *Tp* și una slabă de *Tn* vor avea acuratețea similară (deși comportamentul lor este unul diferit).

După cum s-a putut observa anterior, fiecare dintre aceste metrici are limitările ei și nu poate să trateze toate problemele ce apar în cadrul unei clasificări. De aceea, pe baza acestora s-au definit următoarele metrici:

- Valoarea F (*F-Score*): $F = 2 * \frac{precizie*senzitivitate}{precizie+senzitivitate}$. Una dintre metricile folosite pentru a preveni problemele descrise anterior este valoarea F ce ia în calcul atât senzitivitatea, cât și precizia, aceasta fiind media armonică a celor două [40]. Probleme identificate în lucrul cu aceasta sunt ignorarea *Tn* și acordarea de ponderi similare preciziei și senzitivității.
- Valoarea F ponderată $F = (1 + \beta^2) * \frac{precizie*senzitivitate}{\beta^2*precizie+senzitivitate}$. Pentru a rezolva problema descrisă anterior (ponderile egale ale preciziei și senzitivității) a fost introdusă valoarea F ponderată, aceasta fiind o generalizare a valorii F. Constanta β este folosită pentru a putea varia ponderea preciziei în raport cu senzitivitatea.
- Curba ROC (*Receiver Operating Characteristic Curve*). Conform [41], "graficele ROC sunt grafice bidimensionale, în care rata *Tp* este plotată pe axa Y iar rata I pe axa Ox". Astfel, acest grafic permite vizualizarea compromisului făcut între beneficii (*Tp*) și costuri (*Fp*).
- AUC: $\frac{1}{2}(\frac{Tp}{Tp+Fn} + \frac{Tn}{Tn+Fp})$. Problema în lucrul cu ROC apare atunci când dorim să comparăm două modele pe baza curbelor acestora. Pentru aceasta se folosește aria de

sub curba ROC (*Area Under the Curve*), acesta fiind un număr cuprins între 0 și 1. Conform [42] "această metrică începe să capete popularitate în aplicațiile de clasificare, fiind mult mai robustă decât acuratețea, aceasta luând în considerare distribuția claselor".

După cum am menționat la începutul acestei secțiuni, matricea de confuzie de la care am pornit și metricile derivate pe baza acesteia erau folosite pentru clasificarea binară. În cadrul proiectului propus, clasificarea va fi una multi-clasă (sistemul trebuie să prezică un utilizator dintr-un set de mai mulți utilizatori, o emoție din mai multe clase de emoții). Astfel, abordarea de clasificare binară nu poate fi făcută. Totuși, abordarea și metricile pentru clasificatorii cu mai multe clase se pot deriva din cele binare, având aceeași semnificație. Aceștia se obțin prin abordarea de tip unul-versus-restul (în engleză cunoscută drept *one-versus-all*). Conform [43] există două abordări ce se pot aplica pentru a obține oricare dintre metricile descrise anterior: media micro și media macro. Pentru a exemplifica acești algoritmi am folosit un clasificator cu patru clase, denumite generic A, B, C, D.

Abordarea micro este ilustrată în figura de mai jos și presupune crearea unei matrice de confuzie pentru fiecare clasă prin abordarea *one-versus-all*. Aceste matrice vor fi însumate, obținându-se o matrice de dimensiune 2x2, căreia i se va aplica formula de obținere a metricii dorite.

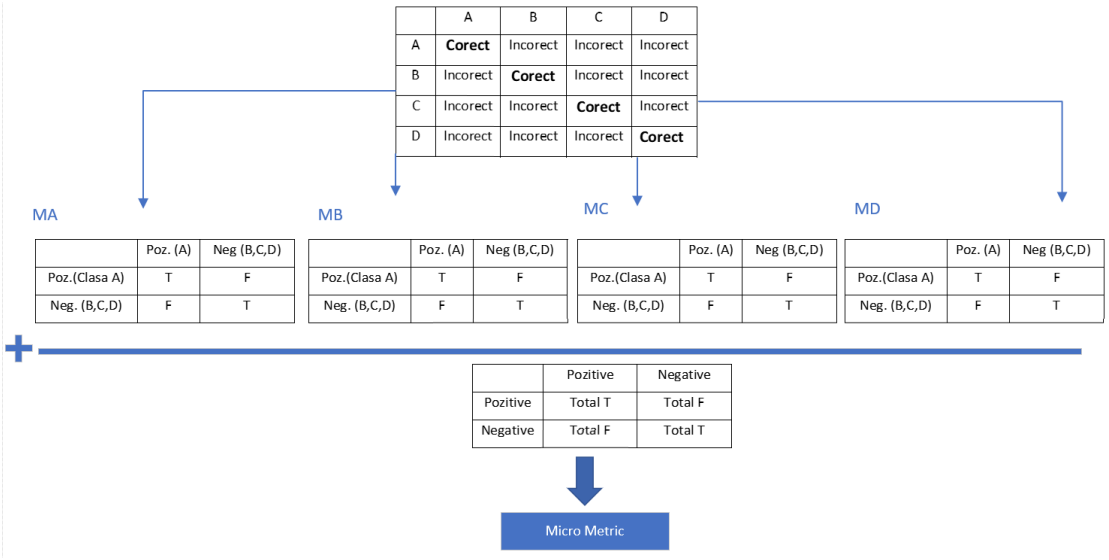


Figura 14 – Obținerea unei metrici de tip micro

Obținerea unei metrici de tip macro este ilustrată în figura de mai jos, calculul acesteia pornind, la fel ca în cazul anterior, de la crearea matricelor de confuzie parțiale (matricele fiecărei clase). Diferența constă în faptul că mai apoi metrici parțiale se obțin din aceste matrice, iar o medie (ce poate fi și ponderată dacă acest lucru se dorește) generează metrica de tip macro.

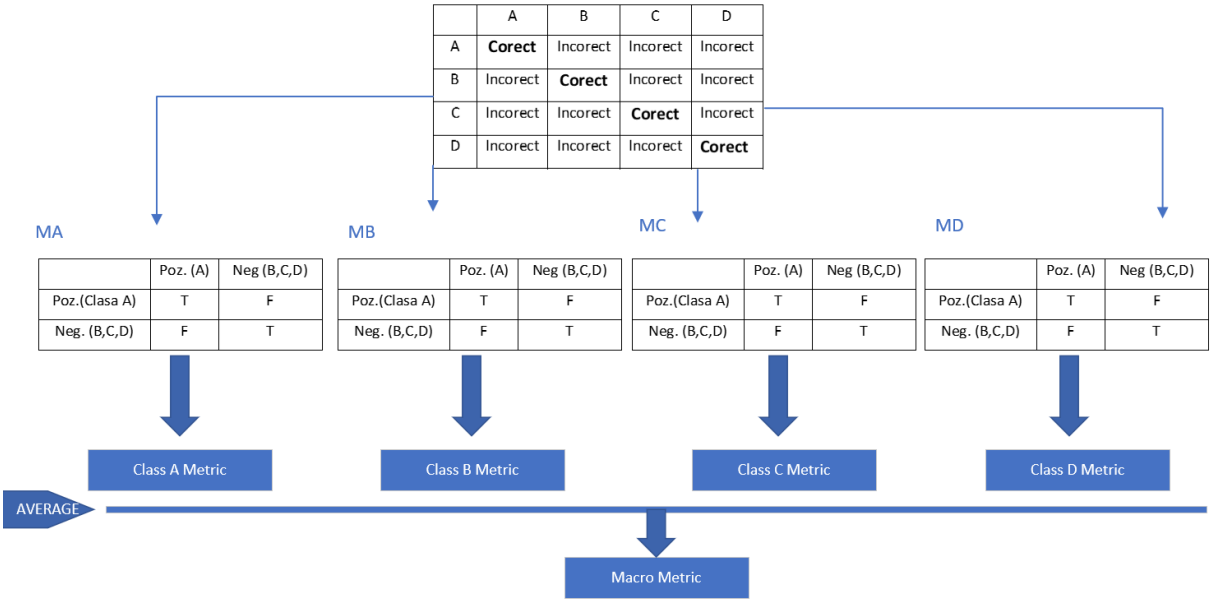


Figura 15 – Obținerea unei metrici de tip macro

Ambele metrici sunt utile pentru analiza unui clasificator, întrucât prima nu este sensibilă la prevalența datelor în cadrul setului, putând fi utilizată pentru compararea mai multor seturi de date, în timp ce cea de-a doua permite observarea comportamentului fiecărei clase și acordarea de ponderi diferite dacă este cazul.

4.2. Soluții informatice

4.2.1. M. Kwon, S. Kwon, 2019

O primă soluție informatică existentă o constituie cea propusă de M. Kwon și S. Kwon în 2019 [44]. Aceasta este o soluție folosită pentru recunoașterea emoțiilor, acestea fiind extrase din spectrograma semnalului vocal. Emoțiile folosite de către aceștia, emoții pe baza cărora sunt realizate clasificările sunt în număr de cinci și anume neutru, vesel, trist, furios și entuziasmat. Din această lucrare vom aborda doar arhitectura rețelei neuronale, formată în acest caz din 7 straturi de convoluție. Dimensiunile straturilor de convoluție nu vor fi descrie explicit, întrucât acestea sunt ilustrate în figura de mai jos. Astfel, spectrograma obținută constituie întrarea în rețeaua de convoluție, iar ieșirea acesteia o constituie un strat de tip *Flatten,* ce transforma matricele în vectori coloană (în acest caz de dimensiune 521x1). Ieșirea sistemului va fi dată de un clasificator de tip SoftMax ce va asigna probabilitatea aferentă fiecărei emoții.

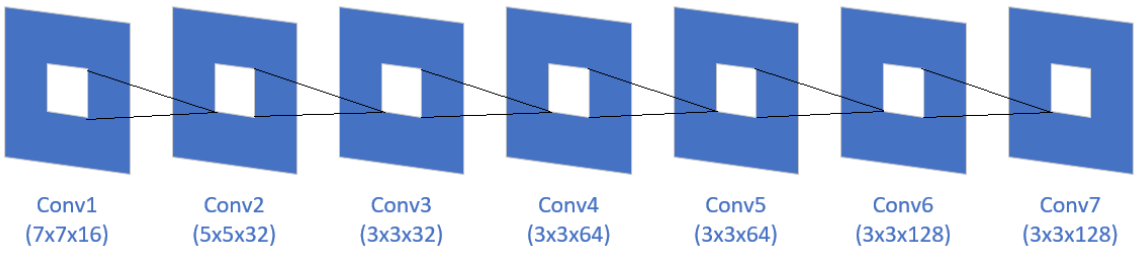


Figura 16 – Rețeaua de convoluție a soluției descrise

În cazul lucrării de față autorii au folosit două seturi de date. În cazul primului set de testare acuratețea de testare a fost de 81.75%, în timp ce pentru al doilea a fost de 79.50%. Precizia, senzitivitatea și valoarea F pe primul set de date sunt făcute la nivel general și se pot observa în tabelul de mai jos.

Emoție	Precizie	Senzitivitate	Valoare F
Furios	96%	87%	91%
Vesel	58%	85%	69%
Neutru	100%	76%	86%
Trist	69%	92%	79%

Tabel 1 – Performanța soluției descrise

4.2.2. P. Dhakal, P. Damacharla, 2019

În lucrarea “A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface” [45] autorii realizează o analiză comparativă a trei metode standard de clasificare din zona învățării automate (SVM, RF și DNN). Tabelul de mai jos sintetizează rezultatele obținute de către aceștia pentru unul dintre seturile de date de testare.

Clasificator	Acuratețe	Specificitate	Senzitivitate
SVM	91.66	93.06	90.42
RF	94.87	97.85	92.23
DNN	93.52	94.17	92.92

Tabel 2 – Comparație a clasificatorilor descriși

Pe lângă această comparație, un alt lucru ce merită extras din analiza făcută de aceștia îl reprezintă arhitectura rețelei neuronale de convoluție. Aceasta conține două straturi de convoluție, două de pooling și unul de ieșire. Peste straturile de convoluție se aplică un filtru de dimensiunea 5x5, iar peste cele de pooling de dimensiune 2x2. Ieșirea rețelei neuronale este un vector coloană, de dimensiune 256x1, acesta fiind intrarea în clasificatorul de tip SoftMax ce stabilește clasa intrării. Pentru a înțelege mai ușor topologia acestei rețele am creat figura de mai jos ce o sintetizează.

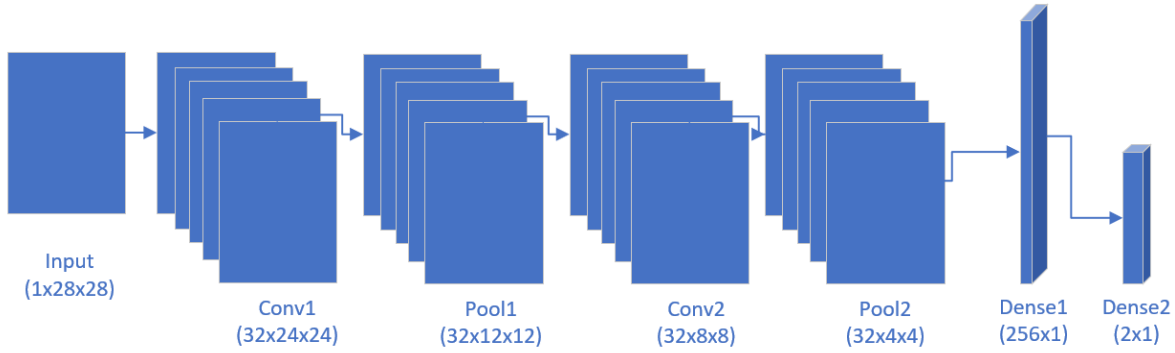


Figura 17 – Topologia rețelei neuronale a soluției descrise

4.2.3. J. Nordby, 2019

O altă soluție informatică este cea propusă de Jon Nordby în anul 2019 [46]. Acesta folosește indicii Mel Cepstrali și rețelele neuronale convoluționale pentru a clasifica diversele sunete înregistrate din mediul înconjurător (claxon, câine, sirenă etc.). În total, numărul de clase folosit de acesta este de 10 clase.

Deși aplicația acestuia este centrată pe implementarea algoritmului pe microcontrollere pentru a o folosi în zona de IoT, merită extrase din această lucrare arhitectura propusă și rezultatele obținute.

Acesta analizează comparative două arhitecturi de rețele neuronale convoluționale. Prima are următoarea topologie: Conv2D, Pool, Conv1D, Pool, Conv1D, după acest ultim strat de convoluție urmând unul de tip Flatten și două de tip dense ce realizează clasificarea cu ajutorul algoritmului SoftMax. Cea de-a doua topologie este derivată din prima, cu excepția faptului că nu conține cele două straturi de MaxPool2D. Fiind o aplicație preponderent IoT, autorul este interesat mai degrabă de metrici ce privesc utilizarea procesorului și a altor resurse, astfel că singura metrică ce poate fi de folos în cadrul acestei lucrări este acuratețea. În cazul primei topologii acuratețea generală obținută este de 72.3% ±4.6, în timp ce pentru a doua este de 68.3% ±5.2.

Autorul creează matricea de confuzie și observă că cele mai bune valori de acurateț obținute pentru fiecare clasă sunt de 96%, 86% și 82.7%, în timp ce cele mai slabe sunt de 60% și 47%. Aceste rezultate duc la ideea, la fel ca în cazul anterior, că sistemul per total nu are un comportament dorit datorită variațiilor mari în ceea ce privește diferitele clase.

4.2.4. SincNet

SincNet este o aplicație de recunoaștere a vorbitorului, aplicație propusă de M. Ravanelli și Y. Bengio în lucrarea “Speaker Recognition from Raw Waveform with SincNet” [47]. Aplicația se bazează pe rețelele neuronale de convoluție pentru a clasifica vorbitorul, rețele implementate cu ajutorul bibliotecii Keras, dar spre deosebire de alte aplicații similare, aceasta nu analizează indicii MFCC, ci doar semnalul sonor. Astfel, particularitățile vocale se extrag direct din semnal, pe baza unor filtre de mici dimensiuni ce extrag variațiile în domeniul timp. Arhitectura sistemului este descrisă în figura de mai jos:

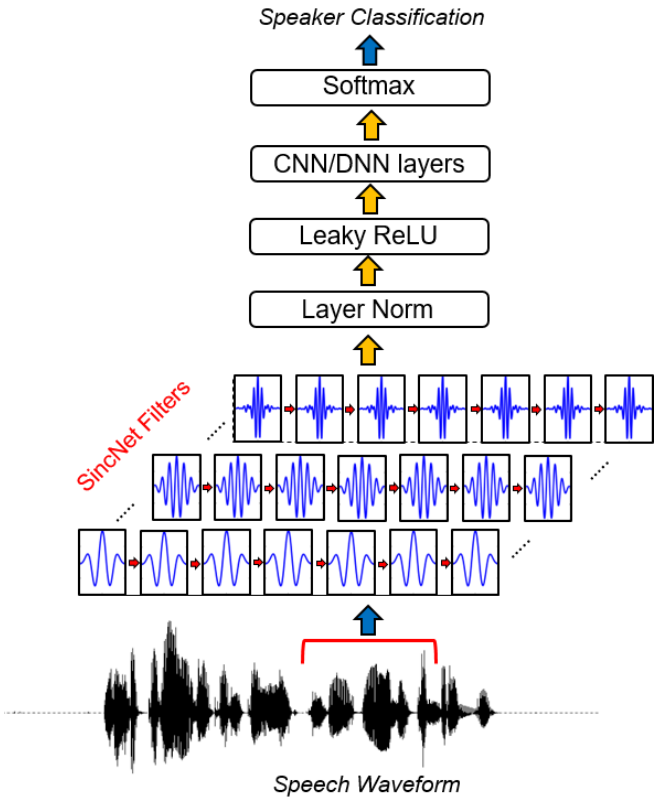


Figura 18- Arhitectura SincNet [48]

Aplicația este modoficată pentru a putea fi folosită atât pentru identificare de persoane, cât și pentru verificare. Rezultatele pe care le voi prezenta țin strict de partea de identificare, întrucât aceasta este cea care ne interesează. Autorii verifică sistemul cu ajutorul a două seturi de date și obțin următoarele valori pentru rata de eroare a clasificării, de 0.96 respectiv 0.85.

5. Implementare

5.1. Limbaj de programare. Biblioteci

Codul sursă al acestui proiect este scris în limbajul Python 3. Acest limbaj a fost ales datorită popularității sale la momentul actual, fiind unul dintre principalele limbaje de programare cerute și folosite în aplicații de Machine Learning, atât în producție cât și în cercetare [49].

Pentru ușurință în lucru și obținerea unor rezultate mai bune, în completarea limbajului de programare Python 3 am decis să folosesc biblioteca Tensorflow.

Conform documentației oficiale [50], acesta este "o platformă open-source dedicată învățării automate". Această platformă este dezvoltată de către Google Brain, fiind adesea folosită de Google în cercetare și producție.

Conform unei lucrări științifice a celor de la Google din anul 2015 (atunci când de altfel l-au și lansat pe piață) [51], un avantaj semnificativ al bibliotecii îl constituie "flexibilitatea, acesta putând să fie folosit la soluții de vedere artificială, robotică, NLP (Natural Language Processing), extracție de informații etc., pe o gamă foarte largă de sisteme (de la telefoane și tablete până la sisteme de plăci grafice distribuite)".

Cu timpul, popularitatea TensorFlow a crescut, ajungând să fie folosit de firme mari de pe piață, precum Intel, Coca-Cola, Twitter etc.[50], creându-se o numeroasă comunitate de dezvoltatori în jurul său care să ofere publicului

Popularitatea bibliotecii, comunitatea exitinsă din jurul său, precum și numeroasele publicații și programe open-source oferite pentru a facilita învățarea au fost principalele motive care m-au condus spre a alege Tensorflow în detrimentul altor biblioteci precum PyToch, Theano sau Caffe.

5.2. Seturi de date utilizate

În realizarea acestui proiect am folosit două seturi de date: unul propriu și setul de date RAVDESS. Aceste seturi vor fi detaliate mai jos, realizându-se și o scurtă analiză comparativă a acestora.

5.2.1. Setul de date propriu

Un set de date pe care l-am folosit pentru acest proiect este unul propriu, realizat prin înregistrarea unui număr de cinci persoane apropiate (două persoane de sex masculin și trei persoane de sex feminin). Pentru anonimizare, numele persoanelor au fost înlocuite cu indecși de la 0 la 4. Personanele cu indecșii 0 și 3 sunt de sex masculin, în timp ce persoanele cu indecșii 1, 2 și 4 sunt de sex feminin.

În cazul acestui set de date toate persoanele au fost rugate să repete aceeași propoziție “Ana are mere”, neexistând, astfel, diferite mesaje pentru identificare ca în cazul altor seturi. Această abordare se numește recunoaștere dependentă de text. Pe baza observațiilor făcute pe performanțelor obținute la setul RAVDESS am decis să nu înregistrez decât trei emoții: neutru, vesel și nervos.

5.2.2. Setul de date RAVDESS

Al doilea set de date folosit pentru acest proiect este setul RAVDESS (The Ryerson Audio-Visual database of Emotional Speech and Song), descărcat de pe platforma Kaggle [52].

Setul RAVDESS este destinat aplicațiilor ce recunosc persoana și emoțiile atât după voce, cât și după față, dar întrucât soluția noastră nu se baza pe recunoașterea facială, am eliminat din setul de date fișierele de tip audio-video sau video.

Setul conține un număr de 24 de actori profesioniști (12 bărbați și 12 femei) ce recită două fraze diferite pe diferite intonații (cu diferite emoții). Fiecărui actor îi corespunde un subset de 60 de înregistrări. Spre deosebire de setul propriu, unde nu aveam decât o singură propoziție, acest

set de date are două(“Kids are talking by the door” și “Dogs are sitting by the door”) [53]. Astfel, acest set de date permite atât o abordare de tip dependentă de text (în cazul în care recunoașterile se fac în cadrul unui subset cu același mesaj), cât și una independentă de text (în cazul în care se folosesc ambele propoziții). Emoțiile sunt în număr de opt și anume: neutru, calm, vesel, trist, furios, temător, dezgustat și surprins.

În tabelul de mai jos se regăsesc principalele caracteristici ale celor două seturi de date.

Set de date	**Propriu**	**RAVDESS**
Nr. persoane	5	24
Nr. persoane de sex M	2	12
Nr. persoane de sex F	3	12
Nr. emoții	3	8
Nr. înregistrări/persoană	45	60
Nr. fraze	1	2
Nr. total înregistrări	**225**	**1440**

Tabel 3 - Comparație între cele două seturi de date

5.3. Recunoașterea persoanei

5.3.1. Arhitectura soluției

Figura de mai jos prezintă arhitectura generală a soluției propuse de recunoaștere a vorbitorului. Astfel, intrarea în sistem o constituie un fișier de tip audio. În cazul de față am folosit fișiere de tip mono, deși o astfel de soluție poate fi aplicată, cu mici modificări și fișierelor de tip stereo. Aceste fișiere audio sunt transformate în coeficienți Mel-Cepstrali, aceștia constituind intrarea în rețeaua neuronală de convoluție. Ieșirea acesteia constă într-un vector de probabilități asociate fiecărei categorii de ieșire. Fiecare element al arhitecturii va fi detaliat în paragrafele de mai jos.

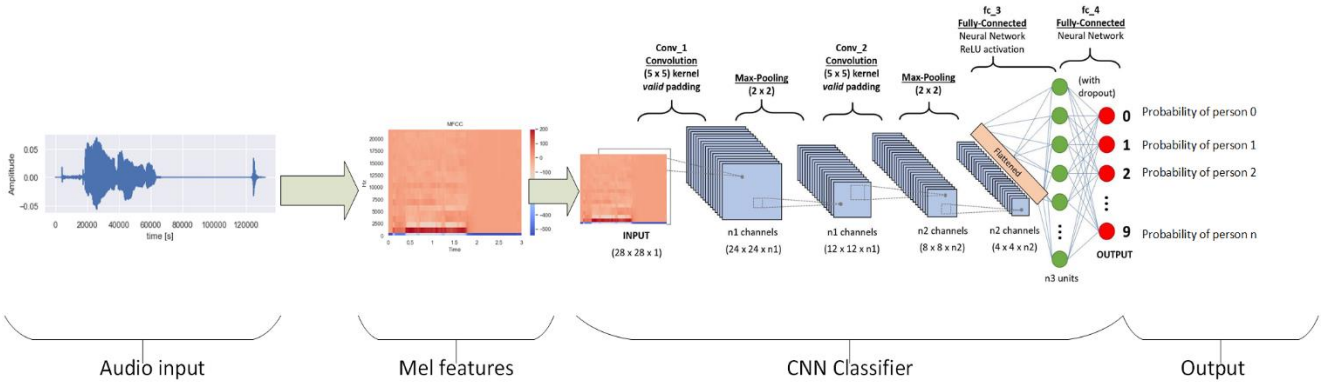


Figura 19- Schema generală a sistemului

5.3.2. Extragerea particularităţilor sonore

Așa cum am precizat anterior, pentru extragerea amprentei, în loc de a lucra direct cu semnalul sonor se lucrează cu mărimi ce reduc dimensionalitatea și evidențiază particularitățile mai ușor (ex. spectrograme, spectrograme Mel, coeficienți Mel). Figura de mai jos ilustrează aceste trei tipuri de reprezentare a semnalului sonor. Datorită rezultatelor bune și a celei mai mici mărimi, în cadrul proiectului am ales să lucrez cu coeficienții Mel, extrași din semnalul sonor cu ajutorul bibliotecii Python *librosa* [54].

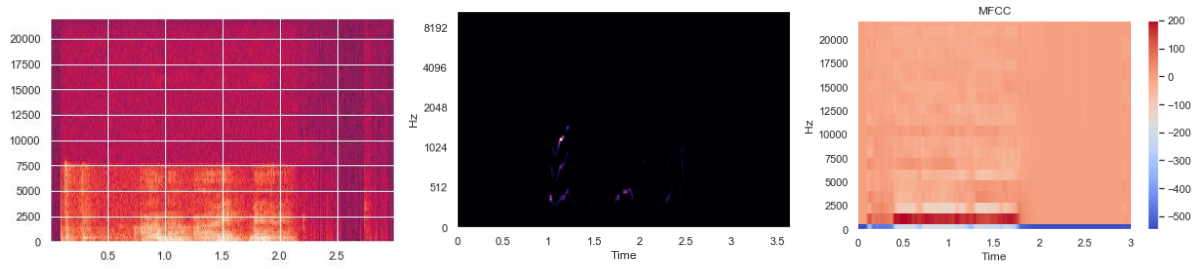


Figura 20- Spectrograma, spectrograma Mel și coeficienții Mel

5.3.3. Rețeaua Neuronală Convoluțională

Așa cum am detaliat în capitolele anterioare, o rețea neuronală de convoluție constituie cea mai bună clasă de rețele neuronale pentru astfel de aplicații. Rețeaua a fost implementată cu ajutorul bibliotecii Tensorflow *Keras* [55]. Rețeaua este formată din două straturi de convoluție

bidimensională. Ambele straturi realizează convoluția cu ajutorul nucleelor de dimensiune 2x2 și folosesc activare de tip ReLu, fiind urmate de un nivel de pooling de dimensiune de 2x2 de tip Max (extrage valoarea maximă din matricea 2x2).

Singura diferență între cele două straturi este că primul extrage 32 de filtre, în timp ce al doilea 48. După cele două nivele de convoluție am folosit un strat de tip Flatten, strat ce modifică dimensionalitatea obținută, aplatizând tensorii în vectori unidimensionali. În cadrul acestui ultim strat funcția de activare este de tip softmax, întrucât rezultatul acesteia este o distribuție de probabilități (distribuția de probabilități a ieșirilor). Figura de mai jos ilustrează un sumar al arhitecturii rețelei.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 19, 299, 32)       160
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 9, 149, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 9, 149, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 8, 148, 48)        6192
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 4, 74, 48)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 4, 74, 48)         0
_________________________________________________________________
flatten (Flatten)            (None, 14208)             0
_________________________________________________________________
dense (Dense)                (None, 24)                341016
=================================================================
Total params: 347,368
Trainable params: 347,368
Non-trainable params: 0
```

Figura 21 - Topologia rețelei neuronale convoluționale

Funcția de optimizare folosită este ADAM, antrenarea realizându-se cu o rată de 0.0001 pentru un număr de 45 de epoci. Datele de antrenare și validare au fost împărțite în raport de 4:1.

5.3.4. Antrenarea rețelei

Figura de jos ilustrează evoluția funcțiilor de acuratețe (graficul din stânga) și de cost (graficul din dreapta) în cadrul procesului de antrenare al rețelei pentru toți cei 24 de indivizi din setul RAVDESS. Din figură se poate observa că nu este necesară antrenarea rețelei pentru mai mult

de 45 de epoci, neexistând îmbunătățiri semnificative. Se previne, astfel și fenomenul de overfitting.

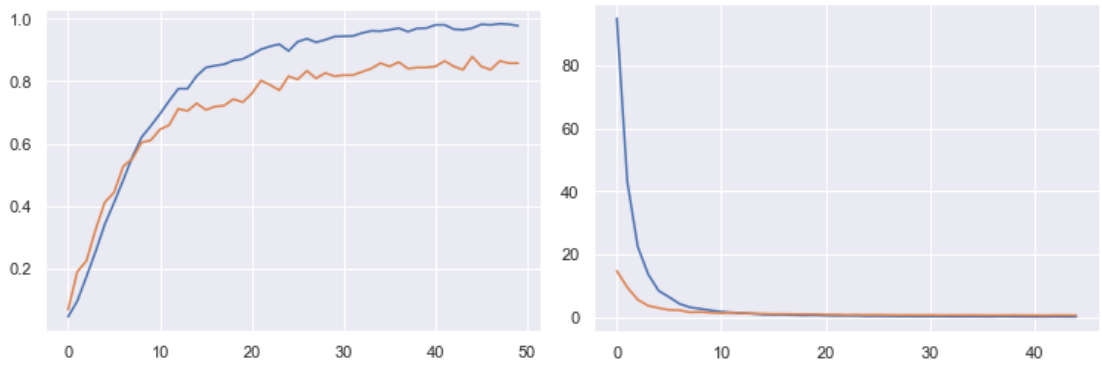


Figura 22 – Acuratețea si funcția de cost

Rezultatul rețelei va fi un vector unidimensional, de lungime egală cu numărul de clase, fiecare element din vector reprezentând probabilitatea acelei persoane. Elementul cu cea mai mare probabilitate va fi ales drept soluție, acesta reprezentând persoana pe care sistemul o consideră a fi cea mai probabilă. Un exemplu de ieșire este ilustrat în figura de mai jos, atât sub formă numerică cât și grafică. Se poate observa, astfel, că elementul cu indexul 5 constituie cel mai probabil element.

```
Out[178]: array([0.03892376, 0.03891166, 0.03891166, 0.03891166, 0.10459333,
                 0.03891166, 0.03891169, 0.03891171, 0.03891188, 0.03891166,
                 0.03891167, 0.03891166, 0.03891166, 0.03891166, 0.03932307,
                 0.03891166, 0.03891378, 0.03891166, 0.03892421, 0.03891166,
                 0.03891166, 0.03891166, 0.0389117 , 0.03891166], dtype=float32)
```

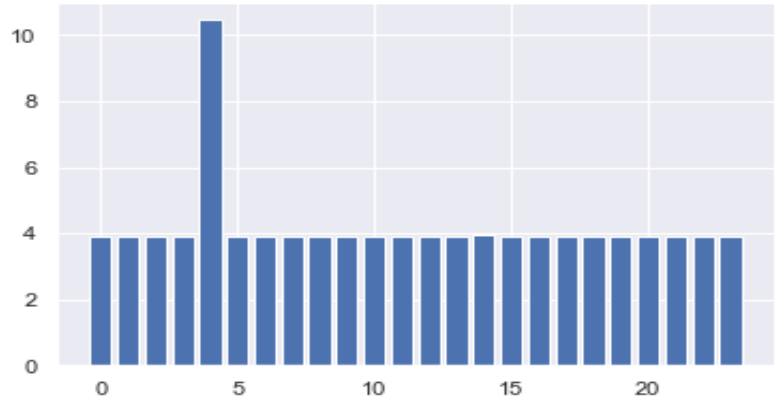


Figura 23- Distribuția probabilităților ieșirilor

5.3.5. Performanțe obținute

Tabelul de mai jos analizează performanțele obținute în recunoașterea persoanei în cazul setului de date RAVDESS. Trebuie menționat faptul că aceste valori reprezintă acuratețea subsetului de validare și nu cele ale subsetului de antrenare.

Nr. Pers.	Acuratețe
2	91.66%
3	91.66%
4	89.58%
5	95.00%
6	87.50%
7	85.71%
8	88.54%
9	86.11%
10	87.50%
11	88.63%
12	87.50%
13	83.33%
14	85.12%
15	93.88%
16	86.45%
17	85.78%
18	90.74%
19	88.59%
20	90.00%
21	89.68%
22	83.71%
23	88.77%
24	83.60%
Min	**83.33%**
Max	**95.00%**
Medie	**88.22%**

Tabel 4 - Performanțele obținute în recunoașterea persoanei

Din acest tabel se poate observa faptul că aceste valori variază între 83.33% și 95%, media valorilor fiind de 88.22%. Alt lucru ce se poate observa din acest tabel este faptul că soluția propusă nu suferă fluctuații mari în funcție de numărul de clase (numărul de persoane din înregistrări).

O comparație cu valorile obținute din setul propriu de date este ilustrată în tabelul de mai jos.

Nr. Pers.	Acuratețe Set Propriu	Acuratețe RAVDESS
2	94.11%	91.66%
3	87.50%	91.66%
4	84.35%	89.58%
5	84.21%	95.00%

Tabel 5- Comparație acuratețe set de date propriu- set RAVDESS

Se poate observa faptul că valorile sunt apropiate iar sistemul reușește să distingă persoane și în cadrul setului de date propriu. Astfel, aplicația poate fi folosită în situații cotidiene, pe date culese din lumea reală, fără a fi nevoie neapărat de lucrul cu date de laborator.

Matricea de confuzie reprezintă o metodă de a evalua performanța rețelei. Aceasta constă într-o matrice pătratică, de dimensiune data de numărul de clase. Astfel, datele de pe orizontală reprezintă datele prezise, în timp ce cele de pe vertical reprezintă valorile corecte. Un sistem este cu atât mai performant cu cât majoritatea datelor se află pe diagonala principal a matricei.

Figura de mai jos ilustrează matricea de confuzie obținută pe datele de validate ale setului RAVDESS. După cum se poate observa, diagonala proeminentă sugerează o bună clasificare a datelor.

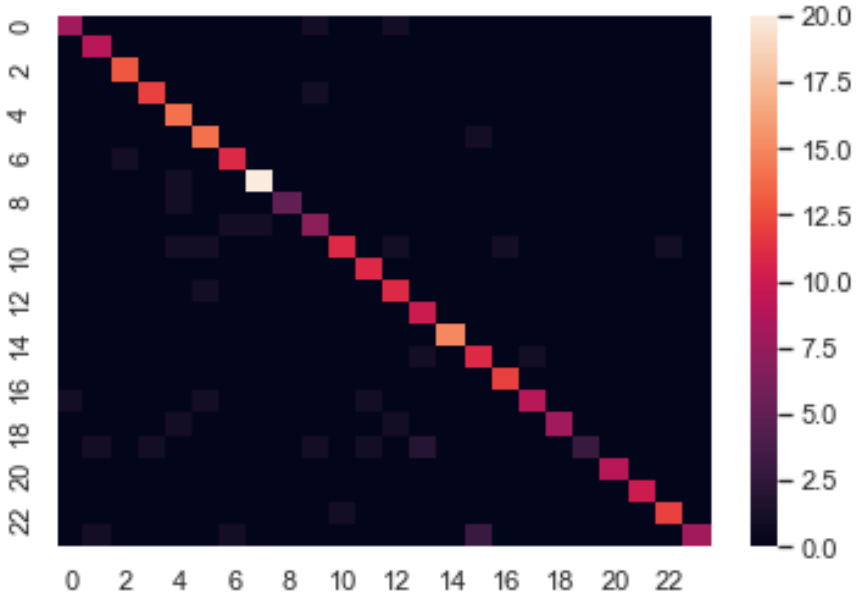


Figura 24- Matricea de confuzie a datelor de validare RAVDESS

Figura următoare ilustrează matricele de confuzie obținute pe setul de date propriu pentru două, trei, patru și cinci clase de persoane. Se poate observa din figură rezultatele satisfăcătoare ale aplicației.

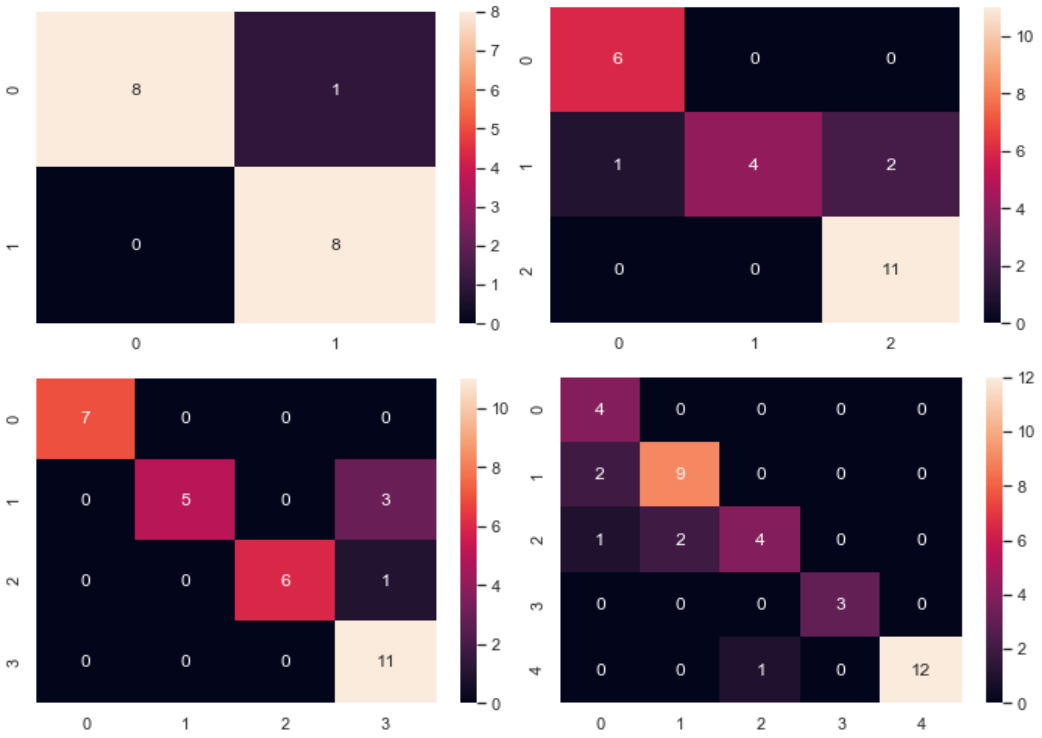


Figura 25- Matrice de confuzie pentru 2, 3, 4 și 5 clase ale setului propriu

Pe baza matricelor de confuzie și cu ajutorul bibliotecii *sklearn.metrics* [56] am obținut metricile descrise în tabelul de mai jos. Acuratețea nu este prezentă în acest tabel deoarece a fost tratată anterior, aceasta intrând în componența optimizatorului pe baza căruia s-a realizat antrenarea rețelei. Ca notație am folosit simbolul µ pentru metricile micro și M pentru cele macro. O altă mențiune ce trebuie făcută este aceea valorile au fost extrase pe baza setului RAVDESS.

Nr. clase	F1µ	F1M	Precizie µ	Precizie M	Senz. µ	Senz. M
2	87.50%	87.47%	87.50%	89.28%	87.50%	88.46%
3	88.88%	87.84%	88.88%	91.11%	88.88%	87.77%
4	83.33%	80.92%	83.33%	85.00%	83.33%	79.70%
5	90.00%	90.99%	90.00%	91.82%	90.00%	90.74%
24	83.97%	84.08%	83.97%	85.36%	83.97%	85.23%

Tabel 6 - Metrici de evaluare pentru modelul de recunoaștere a persoanei

5.4. Recunoașterea emoției

5.4.1. Arhitectura soluței

Pentru recunoașterea emoțiilor soluția este relativ similară celei de recunoaștere a persoanei. Cu toate acestea există mici particularități ce au fost necesare datorită unor performanțe mai slabe ce s-au observat în lucrul cu arhitectura anterioară. Acestea vor fi detaliate în paragrafele următoare.

O altă particularitate a acestei abordări constă în faptul că rețeaua poate fi antrenată anterior cu un set de date precum RAVDESS pentru a învăța să distingă emoțiile și apoi folosită pe un set propriu, având performanțe mai bune decât dacă ar fi fost antrenată direct pe acesta din urmă.

5.4.2. Extragerea particularităţilor sonore

Spre deosebire de recunoașterea persoanei, unde nu era nevoie decât de un număr de 20 de coeficienți Mel per înregistrare pentru a obține rezultate satisfăcătoare, în cadrul recunoașterii de emoții am observat faptul că sistemul obține performanțe mai bune dacă se lucrează cu 40 de coeficienți. De asemenea am modificat și redimensionarea suportului de timp, de la 3 secunde la 4 secunde. În cadrul setului propriu de date această modificare nu are niciun fel de consecință, întrucât toate fișierele sunt înregistrate cu o lungime fixă de 3 secunde. Cu toate acestea, în cadrul setului RAVDESS duratele sunt variabile. Redimensionarea la 3 secunde aducea rezultate satisfăcătoare în cadrul recunoașterii de persoană, dar în cazul emoțiilor am observant rezultate mai bune pentru o dimensiune temporală mai mare, ce nu pierde atât de multă informație.

5.4.3. Rețeaua neuronală convoluțională

Rețeaua neuronală folosită este derivată din cea folosită anterior. Cu toate acestea, în cadrul implementării am observat rezultate mai satisfăcătoare daca adaug după stratul tip Flatten încă

alte două straturi de tip Dense, primul cu 400 de noduri și cel de-al doilea cu 200. La fel ca în cazul anterior, ultimul strat îl constituie unul de tip Dense cu număr de noduri egal cu numărul de clase ales. Topologia rețelei este prezentată în figura de mai jos.

```
Model: "sequential_54"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_56 (Conv2D)           (None, 39, 399, 32)       160
_________________________________________________________________
max_pooling2d_56 (MaxPooling (None, 19, 199, 32)       0
_________________________________________________________________
dropout_56 (Dropout)         (None, 19, 199, 32)       0
_________________________________________________________________
conv2d_57 (Conv2D)           (None, 18, 198, 48)       6192
_________________________________________________________________
max_pooling2d_57 (MaxPooling (None, 9, 99, 48)         0
_________________________________________________________________
dropout_57 (Dropout)         (None, 9, 99, 48)         0
_________________________________________________________________
flatten_28 (Flatten)         (None, 42768)             0
_________________________________________________________________
dense_82 (Dense)             (None, 400)               17107600
_________________________________________________________________
dense_83 (Dense)             (None, 200)               80200
_________________________________________________________________
dense_84 (Dense)             (None, 3)                 603
=================================================================
Total params: 17,194,755
Trainable params: 17,194,755
Non-trainable params: 0
```

Figura 26 - Topologia rețelei

5.4.4. Antrenarea rețelei

Antrenarea rețelei se va face pentru un număr de 25 de epoci pentru a preveni fenomenul de overfitting ce s-a observat la un număr mai mare de 25 de epoci. Rata de antrenare rămâne aceeași, 0.0001. Rezultatele antrenării se pot observa în figura de mai jos.

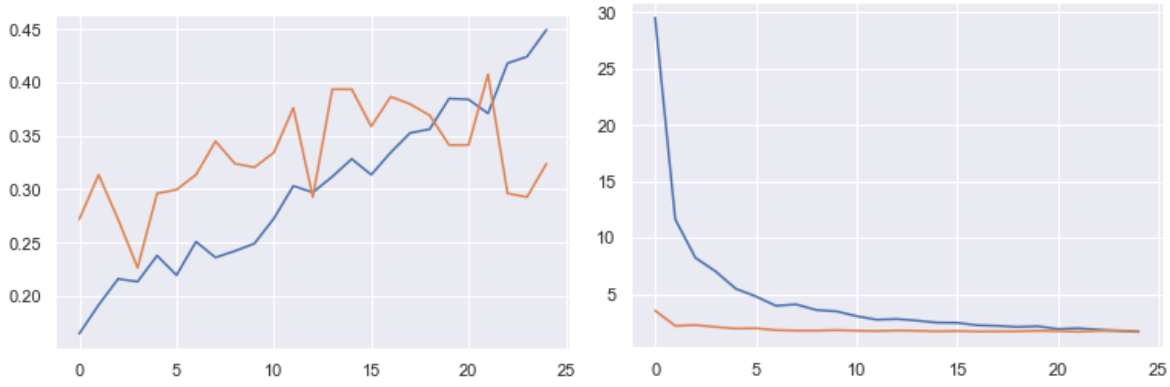


Figura 27- Graficele pentru acuratețe si funcție cost

5.4.5. Performanțe obținute

Inițial am aplicat procesul de antrenare pentru întreg setul de date (pentru toate cele opt emoții ale setului RAVDESS). Figura de mai jos prezintă matricea de confuzie obținută.

Figura 28 - Matricea de confuzie pentru cele 8 emoții

Rezultatele slabe observate în matrice m-au condus să restrâng numărul de clase ale emoțiilor.

Figurile de mi jos reprezintă rezultate obținute pentru două clase de emoții (Vesel-Trist, Furios-Speriat, Neutru-Trist). În cadrul primei clasificări acuratețea obținută a fost de 83.11%, pentru a doua clasificare a fost de 80.71% , iar pentru ultima de 75.86%.

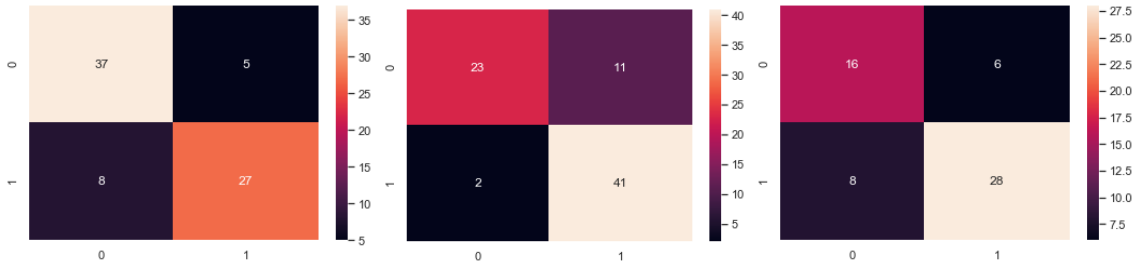


Figura 29 – CM pt. perechile Vesel-Trist, Furios-Speriat, Neutru-Trist

În figura de mai jos se poate observa rezultatul obținut prin utilizarea a trei clase de emoții (Neutru-Vesel-Furios), caz în care acuratețea obținută a fost de 66.67%. Pe baza matricei se poate observa faptul că această clasificare permite distinctia celor trei emoții.

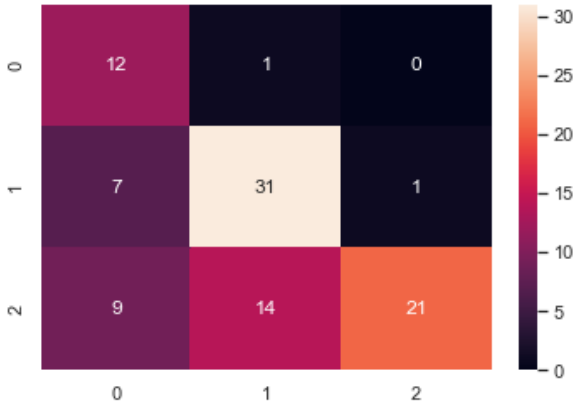


Figura 30 - Matricea de confuzie pentru Neutru-Vesel-Furios

Tabelul de mai jos prezintă metricile obținute pentru clasele de mai sus. Structura și modul de obținere al acestora sunt similar celor de la recunoașterea persoanei. Pentru a simplifica tabelul am folosit următoarea notație: N pentru clasa neutru, V pentru vesel, F furios, T trist și S speriat. Metricile sunt obținute pe baza lucrului cu setul RAVDESS.

Nr. clase	F1µ	F1M	Precizie µ	Precizie M	Senz. µ	Senz. M
2 (V,T)	77.92%	77.78%	77.92%	83.00%	77.92%	80.68%
2 (F,S)	87.01%	87.01%	87.01%	87.09%	87.01%	87.04%
2 (N,T)	77.58%	76.74%	77.58%	76.64%	77.58%	79.34%
3 (N,V,F)	72.91%	73.79%	72.91%	74.73%	72.91%	73.61%
8 (toate)	41.11%	38.98%	44.11%	44.68%	41.11%	44.25%

Tabel 7 - Metrici de evaluare pentru modelul de recunoaștere a emoției

5.5. Interfața grafică

Interfața grafică a fost realizată în Python, folosind biblioteca *tkinter* [57]. După deschiderea aplicației utilizatorului îi este afișat meniul principal al programului, ilustrat în figura 31. Acesta conține un număr de patru butoane ce trimit utilizatorul către fereastra de test, fereastra de adăugare și ștergere a utilizatorilor, fereastra de Help, ultimul buton fiind cel de Exit ce realizează închiderea aplicației. Asupra ferestrei de Help nu vom insista, aceasta fiind ilustrată în figura 32 și fiind fereastra ce conține indicații oferite utilizatorului pentru a îl ajuta în lucru cu programul.

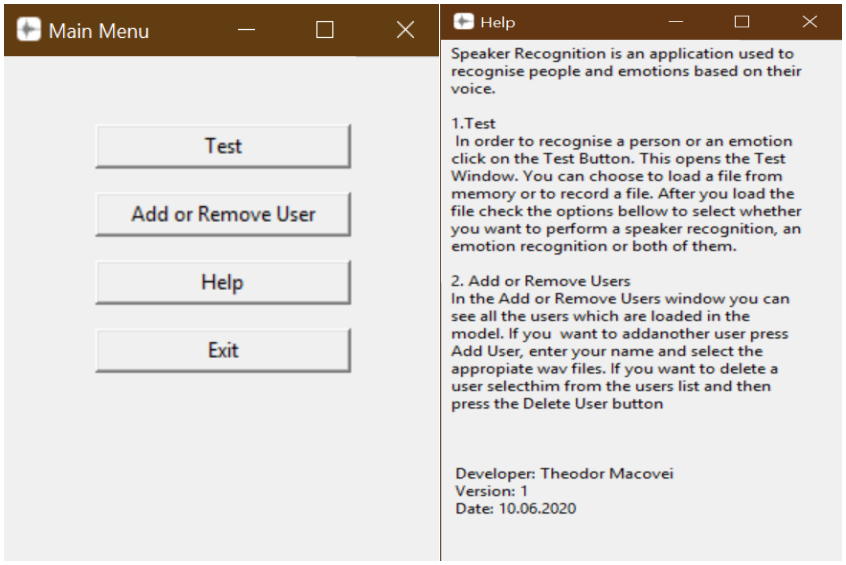
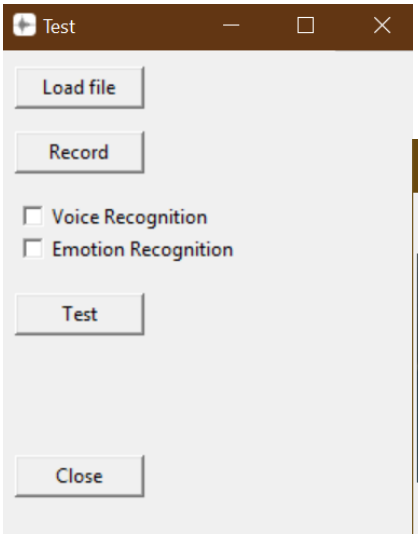


Figura 31 – Meniul principal al aplicației *Figura 32 – Fereastra de Help*

5.5.1. Fereastra de testare

Prin apăsarea butonului *Test* din meniul principal se va deschide fereastra de testare, ilustrată în figura 33. Pentru obținerea vocii vorbitorului există două opțiuni: încărcarea unui fișier de tip *.wav* ori înregistrarea acestuia prin apăsarea butonului *Record*. Tipul de înregistrare este predefinit ca fiind de 3 secunde. După ce programul obține ori prin încărcare ori prin înregistrare un fișier audio, utilizatorul trebuie să aleagă ce rezultat dorește să obțină

(recunoașterea persoanei, recunoașterea emoției ori recunoașterea atât a persoanei cât și a emoției acesteia), bifând corespunzător cele două checkbox-uri. După apăsarea butonului *Test*, utilizatorului îi va fi oferit rezultatul obținut. Atât în cazul în care utilizatorul nu selectează un fișier sau nu înregistrează, cât și în cazul în care utilizatorul nu selectează ce rezultat dorește să obțină, acesta va fi atenționat prin intermediul unei ferestre de tip pop-up. Alături de figura 33 este ilustrat un exemplu de rezultat obținut în urma testării prin înregistrare și a selecției ambelor opțiuni, atât cea de recunoaștere a persoanei cât și cea de recunoaștere a emoției (figura 34).



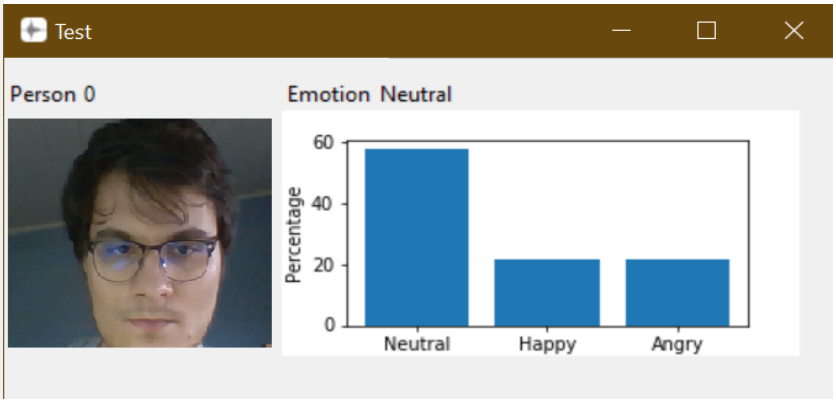


Figura 33–Fereastra de testare *Figura 34–Rezultat obținut pt. testare de persoană și emoție*

5.5.2. Fereastra de adăugare și stergere utilizatori

Prin apăsarea din meniul principal al butonului *Add or Remove Users*, utilizatorului i se deschide fereastra de adăugare și ștergere a utilizatorilor, ilustrată în figura 35. Aceasta este formată dintr-un listbox ce conține listați toți utilizatori înregistrați și două butoane, unul de adăugare de utilizatori (*Add User*) și unul de ștergere de utilizatori (*Delete User*). Dacă se dorește ștergerea unui utilizator trebuie selectat din listbox acel utilizator și apăsat butonul *Delete User*. Pentru a preveni ștergerile accidentale, utilizatorului îi este deschisă o fereastră de tip pop-up prin care este întrebat dacă este sigur că dorește ștergerea unei persoane. Dacă se apasă pe

butonul *Add User*, utilizatorului i se deschide o fereastră în care trebuie să adauge numele utilizatorului și fișierele de tip wav ce conțin înregistrări ale vocii acestuia, acestea fiind câmpuri obligatorii. În cazul în care acest lucru se dorește, poate fi adăugată și o poză de profil a utilizatorului ce urmează a fi introdus prin bifarea chechbox-ului *Profile Picture* și selectarea, prin apăsarea butonului *Select File* a fișierului ce conține poza. După ce sunt selectate opțiunile înregistrării, utilizatorul trebuie să apese pe butonul *Add user to model* și să aștepte până când sistemul va fi actualizat. Aceste aspecte sunt ilustrate în figura 36.

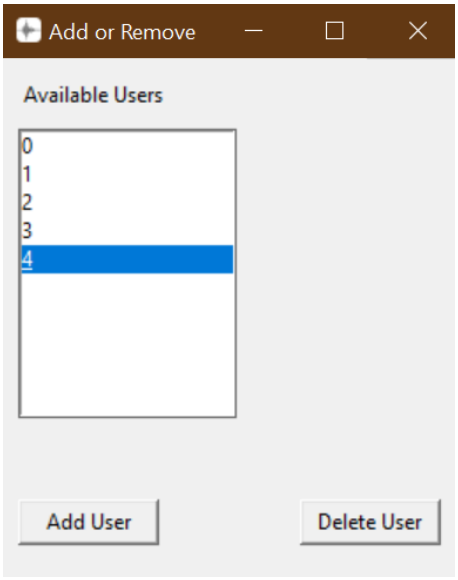


Figura 35 – Fereastra de adăugare și ștergere utilizatori

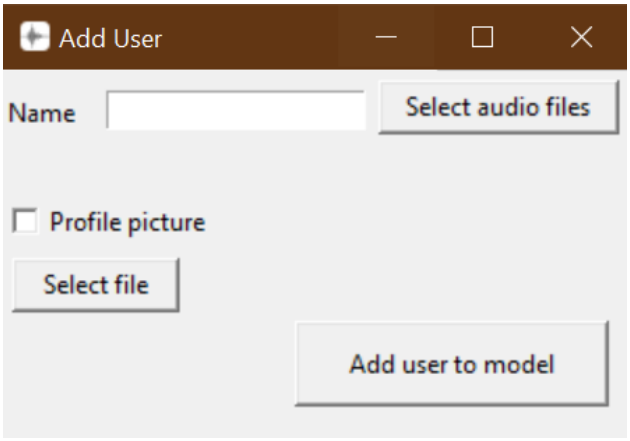


Figura 36 – Adăugarea de utilizatori

6. Concluzii

Această lucrare a tratat problema studiului și a implementării unui sistem de recunoaștere a persoanei și a emoțiilor bazat pe amprenta vocală a acesteia.

Prima parte a lucrării este dedicată studiului elementelor fundamentale ce privesc semnalul vocal, mai exact a felului în care se realizează și este percepută vorbirea de către om și a metodelor de modelare și procesare ale semnalului vocal. A doua parte a lucrării tratează analiza vocală și descrie metodele folosite pentru clasificarea vocală a persoanei și a emoției, în acest caz rețelele neuronale convoluționale. A treia parte a lucrării reprezintă o analiză a soluțiilor informatice existente și a metodelor și metricilor folosite pentru evaluarea unei astfel de aplicații. Ultima parte a lucrării a fost dedicată implementării unui astfel de sistem. Funcționarea acestuia se bazează pe rețelele neuronale convoluționale pentru a putea obține particularități ale vocii fiecărui vorbitor și a realiza clasificarea.

S-a putut observa faptul că în ceea ce privește vorbitorul sistemul este capabil să învețe trăsăturile specifice ale fiecăruia și să distingă între diverși vorbitori. De asemenea, pe baza testelor făcute anterior s-a observat faptul că sistemul nu suferă variații semnificative în funcție de numărul de persoane folosit.

Cu toate acestea, în cazul emoțiilor s-a putut observa faptul ca rezultatele variază atât în funcție de emoțiile alese, cât și de numărul acestora. Acest aspect era de așteptat, fiind observat și în cadrul sistemelor informatice existente descrise. Aceste rezultate se pot explica prin caracterul subiectiv al emoțiilor, ușor de distins pentru oameni, dar nu la fel de clare pentru algoritmi. Întelegerea emoțiilor este un fenomen pur subiectiv, oamenii putând atribui diverse conotații acestora, dar lucrum ai greu de obținut în cadrul rețelelor neuronale. După cum s-a observant în cadrul diverselor testări de mai sus, pentru a putea obține rezultate cât mai bune în cazul recunoașterii de emoții fără a afecta complexitatea sistemului, o soluție ar fi aceea de reducere a numărului de clase (pentru un număr mai mic de emoții sistemul reușeste să le distingă). O altă modificare ce poate fi adusă pentru îmbunătățirea recunoașterii emoțiilor ar fi

introducerea de analiză facială, întrucât cele două analize combinate pot aduce rezultate mai bune.

Astfel, sistemul propus reușește să atingă obiectivele propuse initial, acelea de recunoaștere a vorbitorului și a emoției. Aplicații în viața cotidiană ale acestui sistem sunt diverse, dar o categorie unde acestea pot avea un impact semnificativ îl constituie algoritmii de sugestie. Algoritmul poate fi folosit într-o casă, mașină etc. pentru a genera un playlist audio care să fie în concordanță cu preferințele acelei persoane și a stării acestuia la momentul respectiv. De asemenea, algoritmul poate fi folosit pe roboții telefonici pentru o mai bună experiență cu utilizatorul (robotul modificându-și frazele și tonul pentru a se apropia de necesitățile și starea clientului). În cazul recunoașterii de persoană, pe baza unor mici modificări, algoritmul poate fi folosit pentru verificare și nu identificare, caz în care poate fi folosit pentru aplicații de securitate. Metodele biometrice de securitate devin din ce în ce mai populare, iar vocea este una dintre cea mai sigură metodă biometrică.

7. Bibliografie

[1] “The Human Auditory System”, V. Ruiz, Septembrie 2014

[2]https://upload.wikimedia.org/wikipedia/commons/6/65/Uncoiled_cochlea_with_basilar_membrane.png

[3] “Psychoacoustics: A Brief Historical Overview”, W. Yost, Acoustics Today, 2015, Vol. 11

[4] “Temporal Processing: The Basics”, J. Shinn, The Hearing Journal, Iulie 2003, Vol. 56, pag. 52

[5] “Voice Modeling Methods for Automatic Speaker Recognition”, T. Stadelmann, Philipps Universitat Marburg, Aprilie 2010, pag. 35

[6] “Models and Theories of Speech Production and Perception”, UT Dallas, cap. 14

[7] “Voice based Biometric Security System”, A. Mitra, S. Bisht, V. Ranjan

[8] “The Production of Speech Sounds”, F. Trujillo, English Phonetics and Phonology

[9] https://www.researchgate.net/figure/A-schematic-diagram-of-the-human-speech-production-mechanism_fig1_268817013

[10] “Biometric Authentication. Types of biometric identifiers”, A. Babich, Haaga-Helia, University of Applied Sciences, 2012, pag. 46

[11] https://ei.uni-paderborn.de/en/nt/teaching/veranstaltungen/digital-speech-signal-processing/

[12] https://www.princeton.edu/~cuff/ele201/files/spectrogram.pdf

[13] “Acoustics of Speech Hearing”, UCL, Psychology and Language Science, Lecture 1-10: Spectrograms

[14] http://ec-concord.ied.edu.hk/phonetics_and_phonology/wordpress/learning_website/chapter_3_consonants_new.htm

[15] “Speech Technology: A Practical Introduction. Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis”, K. Prahallad, Carnegie Mellon University & International Institute of Information Technology Hyderabad

[16] https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040, accesat 13 Iunie 2020

[17] “Preprocessing Techniquie in Automatic Speech Recognition for Human Computer Interaction: An Overview”- Y. Ibrahim, J. Odiketa, T. Ibiyemi, Anale. Seria Informatica Vol. XV 1-2017

[18] “Advanced Digital Signal Processing and Reduction”, Second Edition, S. Vaseghi, 2000, John Wiley & Sons Ltd

[19] “The Handbook of Formulas and Tables for Signal Processing” - A.D. Poularikas, CRC Press LLC, 1999

[20] “Hands-On Machine Learning with Sckit-Learn, Keras & Tensorflow”- A. Geron, O’Reilly, 2019

[21] “Machine Learning Algorithms and Applications” - M. Mohammed, M. Khan, E. Bashier, CRC Press, 2017

[22] “An Introduction to Neural Networks”, K. Gurney, University of Sheffield, 1997

[23] http://osp.mans.edu.eg/rehan/ann/Artificial%20Neural%20Networks.htm, 31 Mai 2020

[24] https://ro.sciencewal.com/80966-deep-learning-versus-biological-neurons-floating-point-numbers-spikes-and-neurotransmitters-6eebfa3390e9-22, 31 Mai 2020

[25] https://ro.sciencewal.com/80966-deep-learning-versus-biological-neurons-floating-point-numbers-spikes-and-neurotransmitters-6eebfa3390e9-22, 31 Mai 2020

[26] “An Introduction to Artificial Neural Network”-K. Harsh, N. Bharath , Department of Electrical & Electronics Engineering, Jain University, Bangalore,Vol 1 2016

[27] https://docs.paperspace.com/machine-learning/wiki/activation-function, 1 Iunie 2020

[28] https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051, 31 Mai 2020

[29] “A Practical Approach to Convolutional Neural Networks”- D. Perez, Universidad de Sevilla, 5 Mar. 2019

[30] “Convolutional Neural Networks”- J. Wu, Nanjing University, China, 14 May 2020

[31] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, 31 Mai 2020

[32] https://dictionary.apa.org/emotion

[33] “Emotion: A Psychoevolutionary Synthesis” - R. Plutchik, January 1, 1980, Harper & Row

[34]https://www.scientia.ro/homo-humanus/psihologie/468-ce-sunt-si-care-sunt-emotiile.html

[35] “What’s Basic about Basic Emotions?” - A. Ortony, T. Turner, Psychological Review, 1990, Vol. 97, No. 3, 315-331

[36] “A Study of Speech Emotions Recognition Methods”, A. Joshi, R. Kaur, International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 4, April 2013, pg. 28-31

[37] “Confusion Matrix-based Feature Selection”- S. Visa, B. Ramsay, A. Ralescu, E. Knaap, Midwest Artificial Intelligence and Cognitive Science Conference 2011, Vol 710

[38] https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

[39] “A systematic analysis of performance measures for classification tasks” – M. Sokolova, G. Lapalme, Universite de Montreal, May 2009

[40] “The truth of the F-measure” – Y. Sasaki, University of Manchester, Obtober 2007

[41] “ROC Graphs: Notes and Practical Considerations for Researchers” – T. Fawcett, HP Laboratories, March 2004

[42] “Performance Evaluation for Learning Algorithms” – N. japkowitz, University of Ottawa

[43] “Advisory Panel on Consumer Prices- Technical. Guidelines for selecting metrics to evaluate classification in price statistics production pipelines ”, APCT-T(19)10, UK Statistics Authority

[44] “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition” – M. Kwon, S. Kwon, Sejong University, Seoul, 28 December 2019

[45] “A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface” – P. Dhakal, P. Damacharla, A. Javaid, University of Toledo, March 2019

[46]“Environmental Sound Classification on Microcontrollers using Convolutional Neural Networks” – J. Nordby, Norwegian University of Life Sciences, 2019

[47] “Speaker Recognition from Raw Waveform with SincNet”- M. Ravanelli și Y. Bengio, Mila, Universite de Montreal, CIFAR Fellow

[48] https://github.com/mravanelli/SincNet/blob/master/SincNet.png

[49] https://www.python.org/about/, 31 Mai 2020

[50] https://www.tensorflow.org/, 25 Mai 2020

[51] “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”- M. Abadi, A.Agarwal et al, November 9,2015

[52] https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio, 31 Mai 2020

[53] “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English”, S. Livingstone, F. Russo, PLoS ONE 13(5), 2018

[54] https://librosa.github.io/librosa/feature.html#spectral-features

[55] https://www.tensorflow.org/api_docs/python/tf/keras/Model

[56] https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

[57] https://docs.python.org/3/library/tkinter.html