O scurtă prezentare a metodelor de detectare a emoţiilor din semnalul vocal (Mai 2018)

Iordache-Stoicescu Laurenţiu-Iulian

***Abstract*— Aplicaţiile de clasificare a emoţiilor sunt folosite pentru a îmbunătăţi percepţia maşinii asupra semanticii mesajului transmis de utilizator. În aplicaţiile de clasificare a emoţiilor pe baza semnalului audio, caracteristicile utilizate cu precădere sunt cele ce ţin de statistica frecvenţei fundamentale, conturul energiei, durata porţiunilor de linişte şi calitatea vocii. Performanţa acestor sisteme scade atunci când sunt prezente mai multe emoţii în semnalul audio. În acest articol s-a realizat o prezentare a caracteristicilor prezente în semnalul vocal pe baza cărora se pot determina emoţiile, a factorilor care influenţează aceste caracteristici precum şi o analiză a câtorva metode de deducere a unor emoţii.**

***Index Terms*— PAD - Pleasure-Arousal-Dominance HMM – Hidden Markov Model SVM – Support vector machine LPC - Linear Predictor Coefficients MFCC - Mel-frequency Cepstrum Coefficients LFPC - Log-frequency power coefficients**

I. Introducere

Semnalul vocal reprezintă cea mai rapidă metodă naturală de comunicare între oameni. Acest fapt a determinat apariţia multor interfeţe de comunicare om-maşină pentru a eficientiza metoda de interacţiune dintre om şi maşină. În ciuda tuturor progreselor realizate în recunoaşterea vorbirii, este încă grea realizarea unei interacţiuni naturale între om şi maşină deoarece maşina nu pricepe emoţiile vorbitorului. Datorită acestei necesităţi a apărut ramura de recunoaştere a emoţiilor din vorbire, aceasta se ocupă cu analizarea stării emoţionale a vorbitorului pentru a oferi diverse semantici mesajului transmis de către acesta.

A. Necesitate

Recunoaşterea emoţiilor din vorbire este folositoare în aplicaţii care necesită o interacţiune naturală între utilizator şi maşină cu ar fi filme web şi aplicaţii de învăţare pentru care sunt necesare ca răspuns emoţiile utilizatorului. De asemenea, recunoaşterea emoţiilor mai sunt importante şi în aplicaţii de bord ale autovehiculelor în care informaţia legată de starea mentală a şoferului poate fi utilizată în favoare siguranţei acestuia. O altă aplicaţie în care se poate utiliza recunoaşterea emoţiilor o reprezintă sistemele de traducere automată în care simpla stare a utilizatorului poate varia înţelesul traducerii.

B. Generalităţi

Sarcina de detecţie a emoţiilor este o sarcină complicat de realizat. Nu este clar ce caracteristici ale vorbirii sunt importante pentru distingerea emoţiilor. Pe lângă această incertitudine se mai adaugă şi variabilitatea semnalului introdusă de diversitatea propoziţiilor, vorbitorilor, ciclurilor de vorbire şi a vitezelor de vorbire. Această sarcină mai este îngreunată şi de faptul că de cele mai multe ori nu este prezentă o singură emoţie la un vorbitor, acestea există cumulativ în enunţurile rostite deoarece. Încă un motiv este acela că manifestarea emoţiilor variază de la de la un vorbitor la altul şi pe baza culturii şi a mediului acestuia.

Emoţiile nu au o definiţie teoretică ce este acceptată de toată lumea. Este foarte întâlnit ca acestea să fie caracterizate în două tipuri:

1) De activare: Activarea se referă la cantitatea de energie necesară pentru a exprima o anumită emoţie.

Anumite emoţii cum ar fi bucuria, furia şi frica provoacă o creştere a pulsului, a tensiunii, schimbări ale ritmului respiraţiei, presiune subglotală mărită, uscare a gurii şi ocazional tremurări musculare. Toate aceste modificări fiziologice provoacă schimbări în caracteristicile acustice ale mesajului transmis. Mesajul este enunţat rapid cu sonor ridicat şi o frecvenţă audio ridicată.

Alte emoţii cum ar fi tristeţea induc efecte inverse faţă de cele anterior menţionate şi anume scăderea pulsului şi a tensiunii, creşterea gradului de umiditate a gurii. Aceste emoţii determină un mesaj audio ce este rostit la intensitate scăzută, cu mai puţine frecvenţe înalte.

2) De valenţă: De valenţă: Emoţiile nu pot fi caracterizate doar având în vedere nivelul de energie necesar. De exemplu, bucuria şi furia sunt ambele emoţii cu un nivel mare de energie dar afectează diferit. Această diferenţă este caracterizată de dimensiunea valenţei. Din nefericire nu se ştie precis cum această valenţă afectează caracteristicile acustice pentru alte emoţii. De aceea, clasificarea între emoţiile de excitare înaltă şi cele de excitare joasă se poate realiza cu precizie ridicată iar clasificarea dintre diferite alte emoţii este încă greu de realizat.

O altă chestiune importantă în recunoaşterea emoţiilor constă în alegerea unui set definit de emoţii. Lingviştii au definit un set de aproximativ 300 de emoţii pe care le întâlnim pe parcursul vieţii. Acest pachet de emoţii este însă prea mare pentru a putea fi utilizat momentan, de aceea se recurge la utilizarea unui set restrâns de emoţii de bază. O idee foarte interesantă este aceea că emoţiile pot fi descompuse în aceste emoţii de bază precum culorile pot fi descompuse în culori primare. Emoţiile primare sunt definite de psihologul Paul Ekman ca fiind: tristeţe, agonie, furie, surprindere, frică, dezgust şi bucurie [6].

C. Baza de Date

O chestiune importantă de luat în calcul în evaluarea unui sistem de recunoaştere a emoţiilor din vorbire este gradul de naturaleţe al bazei de date utilizată în evaluarea performanţelor. Se pot trage concluzii incorecte despre sistem dacă datele utilizate nu sunt de calitate.

Din păcate, majoritatea bazelor de date cu date audio ce conţin înregistrări preluate de la vorbitori cu anumite emoţii nu sunt disponibile public. Acest fapt determină o îngreunare a procesului de evoluţie în această ramură. Cercetătorii nu sunt capabili sa se coordoneze din punctul de vedere al greşelilor şi le repetă pentru baze de date diferite.

D. Fapt

În carlinga avioanelor s-a descoperit că sistemele de recunoaştere a vorbirii antrenate cu mesaje de la un subiect aflat sub stres au performanţe mult mai bune decât sistemele antrenate cu mesaje de la subiecţi aflaţi în stare normală.

II. Emoţiile

Cuvântul emoţie este adese considerat ca fiind o stare afectivă a minţii. În conversaţiile umane normale, oamenii sunt capabili să

identifice emoțiile altora pe baza vocii, a posturii, a gesturilor și a expresiei faciale. Emoțiile sunt adesea neglijate în interacțiunea dintre om și mașină. Pentru a putea detecția emoțiilor trebuie determinate trei aspecte:

- Ce reprezintă o stare afectivă?
- Ce semnale din comunicația verbală exprimă informația despre starea afectiva?
- Cum ar trebui combinate diversele caracteristici pentru a optimiza percepția emoțiile de către mașină?

În literatură, emoțiile enumerate de Ekman (furie, dezgust, frică, fericire, tristețe și surprindere) sunt folosite cu precădere. Au fost propuse mai multe modele de ilustrare a relațiilor dintre valențe, excitare și emoții. Figura 1 prezintă modelul PAD al stărilor emoționale o analiză a emoțiilor propuse de alt psiholog, Russell [5].

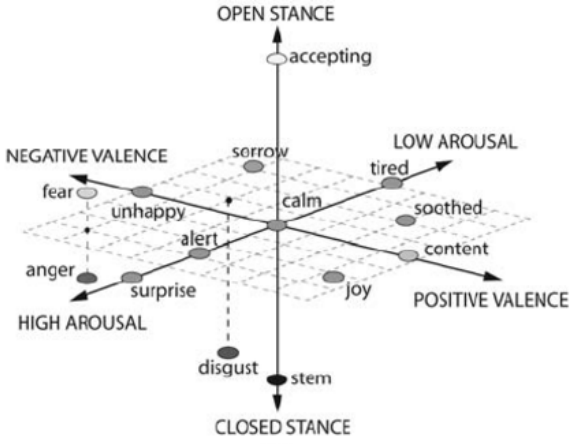

Fig. 1. Modelul PAD 3D de clasificare a emoțiilor [7]

Din acest model se disting următoarele:

- Excitarea(Arousal) - emoțiile sunt caracterizate pe baza nivelului lor de excitare.
- Valență - emoțiile sunt caracterizate în funcție de cât sunt de pozitive sau de negative. Emoțiile pozitive sporesc atenția și alte funcții cognitive ale unei persoane

III. Caracteristici utilizate pentru recunoașterea emoțiilor

Pentru realizarea unui sistem de recunoaștere a emoțiilor din vorbirea unui subiect este necesară identificarea caracteristicilor ce caracterizează diversele emoții. O alegerea caracteristicilor va influența performanțele sistemului.

Pentru o analiză eficientă a caracteristicilor trebuie să se țină seama de regiunea de analiză utilizată pentru extragerea caracteristicilor, să se aleagă caracteristica potrivită pentru identificarea emoției, să se ia în calcul efectele proceselor uzuale de procesare a semnalului vocal cum ar fi filtrarea și eliminarea zonelor de liniște dar și de asemenea să se ia în calcul faptul că trăsăturile acustice s-ar putea să nu fie suficiente pentru modelarea emoțiilor și că ar putea să se folosească și alte tipuri de caracteristici cum ar fi informația discursului sau trăsăturile faciale. În continuare vor fi detaliate aceste aspecte.

A. Caracteristici locale și caracteristici globale

Primul aspect se referă la selectarea regiunii de analiză folosită pentru extragerea caracteristicilor. Aici există două abordări. Divizarea semnalului în cadre de durată scurtă pentru care se calculează vectorul caracteristic sau extragerea de mărimi statistice globale pentru întreaga propoziție.

Deoarece semnalul vocal este nestaționar, se practică divizarea acestuia în ferestre mici de 20 până la 40 de ms. O caracteristică importantă a semnalului vocal, care facilitează analiza acestuia este că acesta este aproximativ staționar pe perioade scurte de timp. Din fiecare astfel de fereastră sunt extrase caracteristici prozodice cum ar fi frecvența fundamentală și energia și denumite caracteristici locale.

Pe de altă parte, caracteristicile globale reprezintă statistici ale tuturor caracteristicilor locale extrase din mesaj. Majoritatea cercetătorilor au agreat faptul că aceste caracteristici globale sunt superioare caracteristicilor locale din punct de vedere al preciziei și a timpului de clasificare. Caracteristicile globale mai au încă un avantaj față de cele locale și anume acela că sunt mult mai puține la număr, ceea ce facilitează o execuție mult mai rapidă a unor algoritmi cum ar fi cel de ”validare în cruce”. Algoritmul de validare în cruce reprezintă o tehnică de evaluare a modelelor predictive petiționând mostra originală într-un set de antrenare pentru antrenarea modelului și un set de evaluare [2]. Utilizarea acestora este eficientă doar în cazul distingerii emoțiilor de excitație înaltă (ex. furia, frica și bucuria) de cele de excitare joasă(ex. tristețea). Caracteristicile globale nu reușesc să realizeze o clasificare precisă a emoțiilor cu excitare similară de exemplu identificarea furiei față de bucurie. De asemenea, mai prezintă și dezavantajul că informația temporală prezentă în mesaj este pierdută. Și încă un dezavantaj îl reprezintă chiar faptul că sunt într-un număr mai mic decât caracteristicile locale deoarece s-ar putea să nu se poată antrena modele de tipul HMM sau SVM.

O altă abordare pentru extragerea de caracteristici o reprezintă segmentarea semnalului pe durate scurte pentru analiza fonemelor. Această abordare se bazează pe faptul că forma spectrului fonemelor variază în funcție de emoții. Această afirmație este valabilă doar pentru vocale.

B. Clasificarea caracteristicilor

Un aspect important în recunoașterea emoțiilor îl reprezintă extragerea caracteristicilor semnalului vocal care caracterizează eficient emoția din mesaj dar în același timp este independentă de vorbitor sau de conținutul lexical.

Caracteristicile vorbirii pot fi grupate în următoarele categorii: Caracteristici continue, caracteristici calitative, caracteristici spectrale și caracteristici bazate pe operatorul TEO. În figura 1 se pot observa exemple de caracteristici aparținând fiecărei categorii.

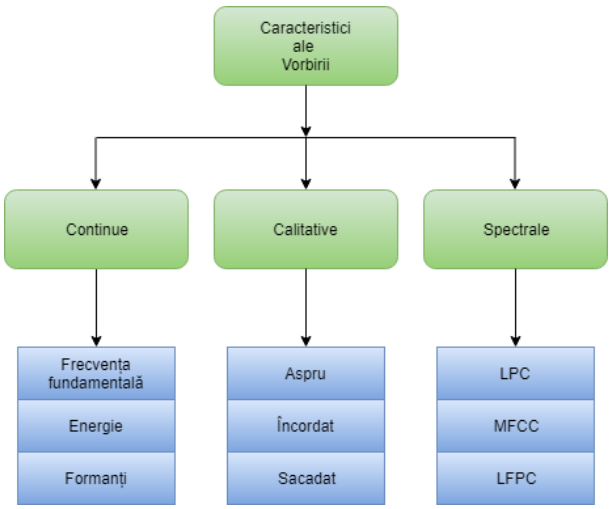

Fig. 2. Categorii de caracteristici ale semnalelor vocale

1) Caracteristici de vorbire continue: Caracteristicile prozodice continue cum ar fi frecvența fundamentală și energia exprimă o mare parte din sentimentele existente într-un mesaj. S-a observat că starea de excitare a subiectului influențează distribuția spectrală a energiei precum și frecvența și durata pauzelor din vorbire. Conform studiilor, aceste caracteristici acustice pot fi grupate în [1]:

- Caracteristici asociate frecvenței fundamentale;
- Caracteristici bazate pe formanți
- Caracteristici asociate energiei semnalului vocal
- Caracteristici ale articulării

2) Caracteristici de calitate a vocii: Calitatea vocii este strâns legată de tipurile de emoții, aceasta este descrisă de emoții care direcționează puternic subiecții către un șir de acțiuni. Acestea sunt opuse emoțiilor fundamentale care influențează pozitiv sau negativ acțiunile și gândurile unei persoane. O gamă largă de variabile fonetice contribuie la impresia subiectivă de calitate a vocii. Corelațiile acustice sunt grupate în următoarele categorii [1]:

- Nivelul vocii, amplitudinea semnalului, energia și durata s-au dovedit a fi metrici bune pentru nivelul vocii.
- Frecvența fundamentală a vocii
- Fraze, foneme, cuvinte
- Structurile temporale

3) Caracteristici bazate pe spectru: Acestea sunt adesea reprezentări ale semnalului pe perioade scurte de timp. Emoțiile dintr-un mesaj influențează distribuția spectrală de energie a mesajului. De exemplu, mesajele care sunt rostite când subiectul este fericit au un nivel ridicat al energiei pentru frecvențele înalte pe când cele rostite de subiecți care sunt triști emoțional au nivele scăzute ale energiei pentru aceleași frecvențe. Pentru o mai bună exploatare a gamei de frecvențe, spectrul este trecut printr-o serie de filtre trece-bandă. Caracteristicile spectrale sunt extrase pe baza rezultatelor obținute de la fiecare filtru. Deoarece percepția umană a frecvenței fundamentale nu este modelată liniar, filtrele trece-bandă sunt distribuite uniform cu privire la o metodă neliniară potrivită cum ar fi scala frecvențelor Mel.

C. Procesarea vorbirii

Înaintea extragerii caracteristicilor este necesară o preprocesare a semnalului audio. De exemplu, datorită diferențelor din mediul de înregistrare este necesară o normalizare a energiei pentru toate mesajele. Alt exemplu îl reprezintă netezirea contururilor extrase, pentru aceasta se folosește metoda suprapunerii cadrelor. Pentru eliminarea ondulațiilor din spectrul de frecvențe sunt utilizate ferestre de tip Hamming.

Deoarece intervalele de liniște pot oferi informații importante asupra stării emoționale, acestea sunt păstrate intacte, față de procesele normale de analiză a vorbirii.

Odată extrase caracteristicile se poate să fie necesară o postprocesare înainte de a antrena clasificatorul. De exemplu, este posibil ca vectorii extrași să aibă unități diferite și prin urmare, valorile lor numerice pot avea ordine de mărime diferite sau pot lipsi cu desăvârșire.

D. Combinarea caracteristicilor acustice cu alte surse de informații

În multe situații, nu este suficientă doar informația provenită din semnalul vocal. Se pot utiliza informațiile referitoare de expresiile faciale sau informația mesajului pentru a putea îmbunătăți performanțele de recunoaștere

1) Utilizarea informațiilor acustice și lingvistice: Semnificația mesajului vorbit este un bun indicator al emoției transmise. Pentru a putea utiliza informațiile lingvistice este necesară recunoașterea secvenței de cuvinte din propoziție. Pentru aceasta este necesar un model de limbă. Acesta descrie constrângerile posibilelor secvențe de cuvinte ale unei limbi. Un astfel de model de limbă este modelul N-gram, acesta atribuie probabilități secvențelor de cuvinte ce pot avea loc. În figura 2 este prezentată arhitectura de bază a unui sistem de recunoaștere a vorbirii ce combină rolurile modelului acustic cu cel al modelului lingvistic pentru identificarea celei mai bune secvențe în care am putea detecta o anumită emoție.

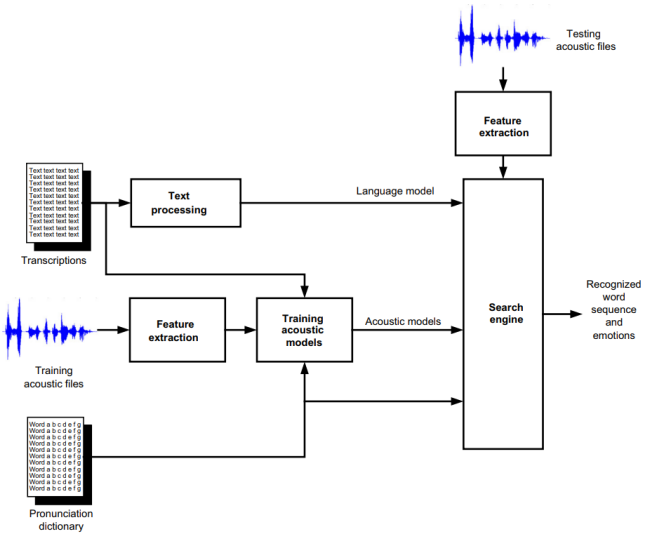

Fig. 3. Categorii de caracteristici ale semnalelor vocale

Modelul de limbă este realizat pe baza transcrierii cuvintelor, în paralel are loc extragerea de vectori de caracteristici din semnalul vocal. Vectorul rezultat împreună cu dicționarul de limbă sunt utilizate pentru a antrena modelele acustice de foneme. La recunoaștere sunt utilizate atât modelul acustic cât și cel de limbă pentru a putea recunoaște secvența după următoarea regulă:

$$\begin{aligned}\hat{W} &= arg\max_{w} P(W|Y) = arg\max_{w} \frac{P(W)P(Y|W)}{P(Y)} \\ &= arg\max_{w} P(W)P(Y|W)\end{aligned} \tag{1}$$

Formula utilizează funcția argmax, care selectează argumentul ce maximizează probabilitatea secvenței de cuvinte. Dezvoltarea din ecuația 1 are la bază regula Bayes și s-a făcut ținând cont de faptul că probabilitatea mesajului vorbit p(X) este independentă de secvența de cuvinte W. Ultimul rezultat evidențiază doi factori care pot fi estimați direct. Problema inițială (găsirea secvenței de cuvinte pe baza mesajului vorbit) a fost împărțită în două sub-probleme mai simple: a) estimarea probabilității apriori a secvenței de cuvinte p(W) și b) estimarea probabilității mesajului vorbit dată fiind secvența de cuvinte pronunțată p(X|W). Primul factor poate fi estimat utilizând exclusiv un model de limbă, iar cel de-al doilea poate fi estimat cu ajutorul unui model acustic. Cele două modele pot fi construite independent așa cum se va vedea în secțiunea următoare, dar vor fi folosite împreună pentru a decoda un mesaj vorbit, așa cum arată ecuația 1 [3].

2) Utilizarea informațiilor acustice, lingvistice și de discurs: Indicatorii de discurs reprezintă expresii lingvistice care transmit informații explicite despre structura discursului sau au o contribuție semantică la aceasta. În contextul de recunoaștere a emoțiilor,

informațiile legate de discurs mai prezintă și modul în care utilizatorul interacționează cu sistemul. Se întâmplă adesea ca utilizatorul să manifeste emoții in timpul interacțiunii cu sistemul, emoții cum ar fi frustrarea. Informațiile de discurs sunt combinate cu informațiile acustice pentru a îmbunătăți performanțele de recunoaștere.

IV. Metode de recunoaștere a emoțiilor

În continuare se a prezenta o comparație între rezultatele obținute de două metode de recunoaștere a emoțiilor. Prima metodă se bazează pe utilizarea modele de mixturi Gaussiene antrenate pe baza caracteristicilor globale cum ar fi frecvența fundamentală și conturul energiei semnalului audio. A doua metodă utilizează modele Markov ascunse antrenate pe baza caracteristicilor locale ale semnalului audio.

A. Modele de mixturi Gaussiene

Modelele de mixturi Gaussiene oferă o bună aproximare a caracteristicilor extrase din semnalul audio prin amestecul ponderat de densități Gaussiene. Coeficienții mixturilor s-au calculat cu ajutorul unui algoritm de maximizare a așteptării. Fiecare emoție este modelata într-un GMM. În experiment s-a remarcat un rezultat bun utilizând un număr de 16 mixturi Gaussiene.

	Recognized emotion						
Labeled emotion	ang	dis	fea	sur	joy	ntl	sad
ang	**91.1**	1.2	0.6	6.3	0.4	0.1	0.3
dis	5.4	**76.8**	6.7	0.1	6.8	3.2	1.0
fea	0.2	6.4	**82.8**	0.6	3.0	0.3	6.7
sur	2.4	2.2	3.0	**87.2**	4.6	0.1	0.5
joy	3.0	0.7	0.8	0.0	**93.2**	0.2	2.1
ntl	0.2	3.4	0.4	0.5	2.7	**89.6**	3.2
sad	0.2	0.1	5.8	3.8	0.4	2.2	**86.6**

Fig. 4. Matricea de confuzie a analizei utilizând modele GMM [8]

În tabel sunt următoarele abrevieri: sur - surprise, joy - joy, ang - anger, fea - fear, dis - disgust, sad - sadness și ntl - neutral [8].

B. Modele Markov ascunse

Modelul HMM reprezintă un automat cu stări finite alcătuit dintr-un set de stări conectate între care se tranziționează. Secvența de stări a modelului HMM este ascunsă, cu toate Matricea de confuzie a fost realizată utilizând 64 de stări și patru mixturi Gaussiene.

	Recognized emotion						
Labeled emotion	ang	dis	fea	sur	joy	ntl	sad
ang	**68.5**	12.7	2.6	1.8	2.7	8.4	3.3
dis	12.8	**84.7**	2.1	0.3	0.0	0.1	0.0
fea	1.8	0.1	**95.4**	0.2	2.0	0.4	0.1
sur	6.3	6.7	6.3	**73.5**	6.1	0.9	0.2
joy	10.1	11.8	7.9	1.2	**68.0**	0.5	0.5
ntl	10.4	0.9	1.0	0.1	1.9	**79.6**	6.1
sad	5.9	10.1	2.8	2.1	2.2	1.8	**75.1**

Fig. 5. Matricea de confuzie a analizei utilizând modele HMM [8]

References

[1] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray : “Survey on speech emotion recognition: Features, classification schemes, and databases”
[2] validare în vruce: https://www.openml.org/a/estimation-procedures/1
[3] Horia Cucu: “Proiect de cercetare-dezvoltare în Tehnologia Vorbirii”
[4] Senaka Amarakeerthi, Rasika Ranaweera, and Michael Cohen: Speech-based Emotion Characterization using Postures and Gestures in CVEs
[5] Vicki R. LeBlanc • Meghan M. McConnell • Sandra D. Monteiro: “Predictable chaos: a review of the effects of emotions on attention, memory and decision making”
[6] Paul Ekman: “Emotions Revealed”
[7] Jonghwa Kim, Elisabeth Andre: “Emotion Recognition Based on Physiological Changes in Music Listening”
[8] Björn Schuller, Gerhard Rigoll, and Manfred Lang: “HIDDEN MARKOV MODEL-BASED SPEECH EMOTION RECOGNITION”