UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Martí Felip, David; Silva Pereira, Silvana, dir. Clasificación de señales EEG del cerebro humano usando tecnicas de aprendizaje automático. 2024. (Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/317361

under the terms of the BY-NC-ND license

# Classificació de senyals EEG del cervell humà mitjançant tècniques d'aprenentatge automàtic

David Martí Felip

**Resum–** Aquest estudi classifica senyals EEG segons ritmes lingüístics (Mora, Syllable, Stress) utilitzant aprenentatge automàtic. Analitzant dades de 48 participants (24 nadius en llengua anglesa i 24 en castellana), es van preprocessar senyals EEG i extreure característiques com la potència mitjana i complexitat Lempel-Ziv. Malgrat l'èxit en l'agrupament de canals, els models amb kernels tant lineals com no lineals no van diferenciar els ritmes, i van obtenir precisions properes a l'atzar. Els resultats suggereixen que les diferències entre dades amb ritmes lingüístics diferents són massa subtils per a les característiques tradicionals, assenyalant la necessitat d'enfocaments més avançats i més volum de dades.

**Paraules clau–** EEG, ritmes lingüístics, aprenentatge automàtic, processament de senyal, complexitat Lempel-Ziv, agrupació de canals, anàlisi TF, models de classificació, Phase Locking Value, Louvain, Saltanaj

**Abstract–** This study classifies EEG signals according to linguistic rhythms (Mora, Syllable, Stress) using machine learning. Analysing data from 48 participants (24 native English speakers and 24 Spanish speakers), EEG signals were preprocessed and features such as mean power and Lempel-Ziv complexity were extracted. Despite optimal channel clustering, models with both linear and non-linear kernels failed to distinguish between rhythms, achieving near-chance accuracy. The results suggest that neural differences are too subtle for traditional features, indicating the need for more advanced approaches and larger datasets.

**Keywords–** EEG, linguistic rhythms, machine learning, signal processing, Lempel-Ziv complexity, channel clustering, TF analysis, classification models, Phase Locking Value, Louvain, Saltanaj

## 1 Introducció - Context del treball

La gran majoria de llengües del món es poden classificar segons el seu ritme lingüístic en tres grups: Mora -timed, Sylllable-timed i Stress-timed. Mora és el grup amb un ritme més ràpid, síl·labes curtes i regularitat en la llargada d'aquestes, mentre Stress té una gran variància de llargada de les síl·labes i Syllable es troba al mig dels dos grups. És demostrat, que l'activitat neuronal del cervell, que sincronitza amb el ritme de la parla, manté una major sintonia amb els idiomes del grup Mora-timed, independentment de la llengua materna de la persona (Silva Pereira et al., 2024) (Ekin Özer et al., 2023). És a dir, aquesta part del cervell, té una capacitat major de sincronitzar el nivell d'activitat de les neurones amb el ritme de parla i els seus patrons temporals. En els estudis mencionats, aquesta sincronització és calculada a partir del PLV. En aquest treball es buscarà trobar diferències significatives en l'activitat neuronal entre els grups segons Mora, Syllable i Stress, així com aplicar Machine Learning per a la seva classificació automàtica.

## 2 Objectius

L'objectiu general d'aquest treball és la classificació, mitjançant un model de Machine Learning, de les dades d'activitat neuronal del cervell segons el ritme lingüístic de l'idioma al qual hom està escoltant. Aquestes dades són

• E-mail de contacte: davidmarti03@gmail.com
• Treball tutoritzat per: Silvana Silva Pereira (Departament de Ciències de la Computació)
• Curs 2024/25

mesures de la potència al llarg del temps, per a diferents freqüències i canals, mentre se sent una petita frase. Són dades EEG cedides exclusivament per al treball per part del Grup de Recerca en Adquisició i Percepció de la Parla (SAP) de la Universitat Pompeu Fabra.

Per al compliment de l'objectiu general, es plantegen diversos objectius específics:

- Coneixement del context científic i naturalesa de les dades.
- Agrupació de canals per a la reducció de dimensionalitats.
- Preprocessament i anàlisi de les dades per a la creació de features
- Confecció d'un model adequat i afinament del mateix

## 3 Estat de l'art

Els estudis recents confirmen que el ritme de la parla s'afegeix als mecanismes d'entrainment neuronal: l'activitat cerebral es sincronitza amb l'envoltant acústica del discurs, especialment al voltant de la freqüència d'una síl·laba ( 5 Hz) (Silva Pereira et al., 2024). Aquesta sincronització es mesura sovint mitjançant el PLV o el seguiment cortical del senyal de parla.

El PLV és una mesura que s'utilitza per quantificar la sincronització de fase entre dos senyals (Namburi, 2011) (Schwartz, 2000). En l'àmbit de la neurociència i el processament de la parla, el PLV s'empra per mesurar específicament la sincronització entre:

- Les oscil·lacions de l'activitat neuronal (registrades, per exemple, mitjançant EEG).
- La periodicitat de l'estímul acústic de la parla, en particular l'embolcall d'amplitud del senyal de parla.

Aquesta sincronització s'analitza sobretot en la banda de freqüències theta (aproximadament entre 3 i 8 Hz), que coincideix amb el ritme sil·làbic de la parla contínua. El procés pel qual l'activitat cerebral segueix aquest ritme s'anomena seguiment neuronal (neural tracking).

Per exemple, (Varghese et al., 2022) van comparar parlants nadius d'anglès (ritme basat en l'estrès), francès (basat en la síl·laba) i japonès (basat en la mora) sentint senyals sintetitzats a la cadència dominant de cada ritme (Stress, síl·laba o mora). Van trobar que el seguiment cortical era dominant a 5 Hz en tots els grups, però amb modulacions específiques: els parlants de francès mostraven un seguiment significativament més intens a 5 Hz que els d'anglès o japonès, i també una millor sincronització del discurs a 10 Hz en comparació amb un estímul no-oral (Varghese et al., 2022). D'altra banda, He et al. (2024) van estudiar parlants nadius d'anglès i de xinès, focalitzant-se en ritmes Stress. Es va observar que només els parlants d'anglès (idioma Stress-timed) mostraven una fase-lleugerament augmentada a 2 Hz (ritme Stress anglès) i a 4 Hz (ritme sil·làbic), fenòmens absents en parlants de xinès (He et al., 2024). Aquestes troballes suggereixen que el seguiment neuronal dels ritmes Stress depèn de l'experiència lingüística.

D'altres aproximacions han explorat mesures de complexitat temporal. Pereira et al. (2024) van analitzar senyals


Fig. 1: Esquema del dataset.

EEG reconstruïts en zones temporals per investigar l'efecte de la llengua nadiua sobre el neural tracking. Amb eines de modularitat i complexitat (Lempel–Ziv), van observar que la complexitat EEG augmenta quan decreix la regularitat rítmica de l'estímul; és a dir, llengües amb ritmes menys regulars generen senyals més complexos (Silva Pereira et al., 2024)(Varghese et al., 2022). En conseqüència, conclouen que "les llengües amb ritmes menys complexes són més fàcilment seguibles per l'escorça neuronal".

En resum, la investigació recent (des de 2020) utilitza mesures com el PLV, la complexitat Lempel–Ziv i d'altres índexs per caracteritzar com el cervell respon als ritmes lingüístics. Aquests estudis demostren modulacions clares de l'activitat EEG segons la classe rítmica de la llengua escoltada.

## 4 Metodologia

### 4.1 Planificació

Per al seguiment de la feina feta durant el transcurs del treball s'ha fet un esquema de fases Appendix A.3 , on hi trobem les diferents etapes de treball. També s'ha fet el seguiment en un diari de treball, que serveix de suport per a la confecció dels diferents informes de progrés i finals.

#### 4.1.1 Canvis

El procés d'agrupació de canals per a la reducció de dimensionalitat ha presentat diversos desafiaments metodològics. Durant la fase inicial, dedicada a aquesta reducció, la inconsistència en els resultats obtinguts en generalitzar els grups de canals a escala de trial va requerir un replantejament de l'aproximació. Com a solució, es van definir les comunitats de canals mitjançant anàlisi visual de manera temporal, més endavant es va poder comprovar que l'agrupació detectada i l'anàlisi visual coincidien i s'ha modificat aquesta part del treball. A conseqüència d'aquests contratemps, l'etapa de classificació es va iniciar més tard del previst.

Es va optar per treballar inicialment amb un conjunt de dades més simple, que prové del mateix experiment. Aquesta decisió va permetre familiaritzar-se amb l'estructura del conjunt de dades, extreure conclusions preliminars sobre patrons globals (encara que amb resolució reduïda), i validar el flux de treball abans d'aplicar-lo a les dades completes.

## 4.2 Context de les dades

Aquest és un dataset creat arran de l'estudi prèviament esmentat, que demostra una major sensibilitat en la sincronització de l'activitat neuronal provinent de la cortesa neuronal i el ritme sil·làbic de l'input, quan aquest és més regular, propi d'idiomes com el japonès. (Silva Pereira et al., 2024)

Les dades EEG analitzades en aquest estudi mostren l'evolució de la potència del senyal al llarg del temps en diferents bandes de freqüència, obtingudes mitjançant una anàlisi temps-freqüència (TF) (Wikipedia contributors, 2023a). Aquesta tècnica descompon el senyal EEG en les seves components freqüencials i mostra com varien aquestes components al llarg del temps. En el nostre cas, l'anàlisi TF permet identificar en quins moments i freqüències hi ha una major activitat cerebral relacionada amb el processament dels diferents ritmes lingüístics.

En aquest experiment es va comptar amb un total de 48 participants, dividits entre parlants nadius d'anglès i d'espanyol (24 de cada llengua).

Es van utilitzar àudios d'oracions resintetitzades, repartides en parts iguals per ritmes lingüístics de la llengua en la qual es graven (Mora, Syllable i Stress), al voltant de 35 per a cadascun. Per al grup Mora es van triar frases en japonès, per al grup Syllable es va triar el català i castellà i per al grup Stress l'anglès i l'holandès. Es va aplicar el mètode de síntesi Saltanaj Appendix A.2 sobre les frases escollides per tal de mantenir la complexitat sil·làbica de les oracions originals, tot eliminant-ne el significat i així evitar possibles efectes de comprensió. (Silva Pereira et al., 2024).

La unitat bàsica d'anàlisi després del preprocessament és un "trial", que correspon a la representació d'una sola oració i un participant. Per a cada trial, el senyal EEG conté dades en una finestra temporal que abasta des d'un segon abans fins 3,5 segons després de l'inici de l'oració. A les dades originals es van enregistrar 64 canals EEG, 4 del quals serveixen de referència.

Per tal d'assolir una millor comprensió de les dades amb què es treballen, les seves dimensions i significat s'ha realitzat un esquema de dades Fig. 1. L'esquema explica visualment les dades. Podem veure a l'esquema cubs de 3 dimensions agrupats per classe (Mora, Syllable i Stress) i per l'idioma natiu dels participants. Per a cadascun d'aquests grups tenim 24 cubs, corresponents als 24 subjects (participants) de l'idioma nadiu corresponent que han participat en l'experiment, aquests són representats amb smileys grocs i blaus distingint la llengua. A cadascun d'aquests cubs, hi trobem dades agrupades en tres dimensions. Concretament, 35 matrius, una per cada trial, que guarden 90 valors de potència per a cada canal, formant l'evolució de la potència per canal i temps en cada trial i subject concrets.

# 5 DESENVOLUPAMENT

## 5.1 Preprocessament de les dades

Primerament, s'han eliminat els valors per a freqüències fora del rang entre 5 i 8 Hz. Això s'ha fet perquè la banda de freqüència theta, generalment definida entre 5 i 8 Hz, és fonamental per a la percepció de la parla. Els models neuronals de la percepció de la parla suggereixen que el cervell segmenta la parla en fragments similars a síl·labes mitjançant el neural entrainment en aquest rang de freqüències theta (Ekin Özer et al., 2023). Una síl·laba té una durada aproximada de 200 ms, la qual cosa correspon a una freqüència al voltant dels 5 Hz. Després s'ha col·lapsat aquesta dimensió fent la mitjana dels valors en aquestes freqüències. També s'ha realitzat un slicing en el temps per tal d'eliminar el primer segon i mig. Durant el primer segon, els participants encara no han començat a sentir res, i el mig segon restant s'elimina per evitar l'efecte del evoked potential (un augment especial de potència que es dona en el moment en què el cervell comença a reaccionar). Dels 60 canals, s'han descartat els 23 canals més exteriors, ja que és demostrat que canals exteriors presenten una quantitat de soroll major a la resta, a causa del comportament del senyal, que pot rebotar en els canals exteriors o ser afectat per cabells i petits moviments durant l'experiment (Frontiers in Psychiatry, 2025).

El resultat és un dataset amb les dimensions: 48 subjects x 3 classes x 35 trials (aprox) x 37 canals x 60 punts de temps. Cal puntualitzar que el nombre de trials no sempre és 35 i pot variar molt suaument, ja que alguns han estat descartats prèviament. D'aquests 48 subjects, 24 són natius en llengua espanyola i 24 en llengua anglesa.

Cal esmentar que durant les primeres visualitzacions, que es faran amb un conjunt de dades diferent, amb dades col·lapsades i amb menys dimensions, trobarem que hi ha 3 canals inferiors ("PO3", "POZ", "PO4") que afegeixen soroll al conjunt de les dades. Aquests canals seran descartats i l'estudi seguirà amb només 34 canals. També trobarem 90 valors en la dimensió temporal, ja que en el moment d'aquestes proves no s'havia realitzat l'slicing temporal, que s'ha aplicat directament a les dades més granulars.

## 5.2 Reducció de dimensionalitat

La reducció de la dimensionalitat del conjunt de dades EEG és un pas fonamental motivat per la necessitat de transformar senyals complexos i sorollosos en features significatives per a la seva anàlisi i interpretació. Les mesures d'EEG són el resultat de la superposició de múltiples senyals a l'espai dels sensors, cosa que fa que una localització precisa de les fonts cerebrals sigui impossible sense aplicar restriccions addicionals i que l'anàlisi de les dades en la seva forma raw sigui molt difícil. Això implica un conjunt de dades amb molt de soroll, del qual necessitem agrupar-ne els canals en grups similars, perdent-ne la mínima informació. Per a això utilitzarem tècniques com l'algoritme de Louvain.

### 5.2.1 Assajos amb dades average

Durant les primeres setmanes, s'ha treballat amb un dataset on la dimensió "trial"és col·lapsada aplicant la mitjana. Així s'ha assolit amb èxit una primera fase de coneixement de les dades i familiarització amb aquestes. Les dimensions d'aquest dataset són: 48 subjects (24 nadius en llengua anglesa i 24 en llengua castellana) x 3 classes (Mora, Syllable i Stress) x 60 canals x 90 punts de temps. Cada subject és un participant en l'experiment i cada classe (Mora, Syllable i Stress) determina el ritme lingüístic de l'idioma de les frases que han estat gravades. D'aquests 60 canals, se'n descarten 23, ja que els canals més exteriors contenen molt

de soroll i un senyal alterat (Frontiers in Psychiatry, 2025).

El primer pas en la detecció de comunitats ha estat l'ús de la tècnica PCA. Aquesta tècnica ens permet saber quins canals són els que expliquen en major part el comportament global. Hem aplicat aquesta tècnica a escala de subject i després hem buscat quins són els canals que més apareixen com a components principals. Tot i ser una bona pràctica exploratòria, i ajudar-nos a entendre de manera més profunda el dataset, aquesta tècnica no ens ha aportat informació valuosa per a la reducció de dimensionalitat del dataset. Si bé en podem concloure que els canals C1 i CP1 expliquen en gran part la variància de les dades, aquesta informació no és útil per aconseguir una agrupació de canals, que ens permeti reduir aquesta dimensió.

En el camí cap a la reducció de la dimensionalitat, volem reduir el nombre de canals intentant perdre el mínim d'informació possible, agrupant aquests en comunitats i col·lapsant aquestes en un sol senyal corresponent a la mitjana de tots els canals de la mateixa comunitat. S'ha calculat la matriu de correlació entre canals per a cada subject i classe (Mora, Syllable i Stress). D'aquesta manera, identifiquem mitjançant un valor concret per a cada parella de canals, la semblança en el seu comportament. S'ha creat una matriu amb la mitjana dels valors en totes les matrius calculades, una per cada trial, i aplicat l'algorisme de detecció de comunitats Louvain al graf creat usant els valors de la matriu com a pesos per les relacions entre cada parella de canals.

L'algorisme de Louvain és un mètode de detecció de comunitats àmpliament utilitzat en l'anàlisi de xarxes complexes, destacant per la seva eficiència computacional. Aquest algorisme optimitza iterativament la modularitat, una mesura que quantifica la qualitat de l'agrupament en comunitats. La seva capacitat per identificar estructures jeràrquiques en grafs el converteix en una gran eina per ajudar-nos a definir les comunitats de canals (Wikipedia contributors, 2023b).

Després d'aplicar aquest algorisme amb diferents valors en la resolució, fet que influeix en el nombre de comunitats en el resultat, trobem que en la gran majoria de casos, els canals frontals, a la part superior del topoplot, formen part de la mateixa comunitat. Així doncs, entenem que els canals frontals tenen moltes possibilitats de formar una comunitat. És per això que ordenem els canals col·locant els canals segons la seva alçada al topoplot (de superior a inferior) i observem que els canals superiors es comporten de manera diferent. Vegem la diferència en el comportament de la potència al llarg del temps per a un trial concret Fig. 2. Aquest gràfic ha estat realitzat posteriorment, quan ja es treballa amb dades sense col·lapsar sobre els trials.

A partir d'ara, buscarem trobar diferències entre tres grups de canals: canals frontals, canals a l'esquerra i canals a la dreta des de la visió. En l'estudi que hem fet fins ara, s'aprecia un comportament diferent entre canals segons la seva posició. Quan estan a prop, aquests es comporten de manera semblant, com hem vist amb el grup superior. Per tant, en els següents intents per a detectar comunitats buscarem una detecció similar a aquesta, que divideixi els canals inferiors i de mitjana alçada segons la seva posició lateral, segons si estan a la dreta o a l'esquerra del topoplot.Fig. 3.

Per a trobar aquestes diferències, visualitzem com oscil·la la potència al llarg del temps per a les diferents classes i grups de canals.

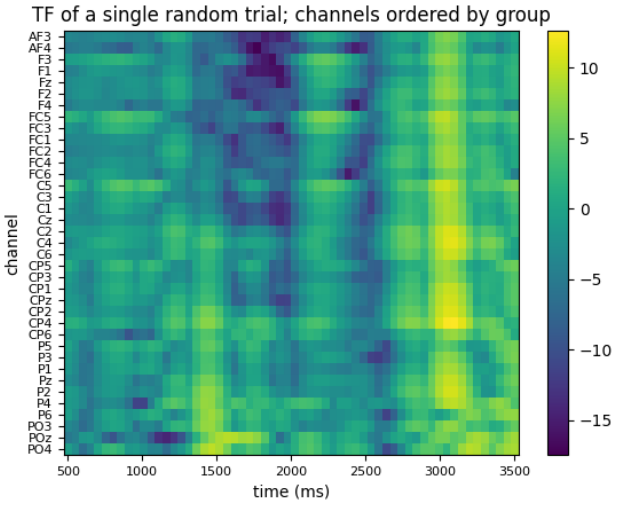


Fig. 2: Potència al llarg del temps ordenant canals de superior a inferior

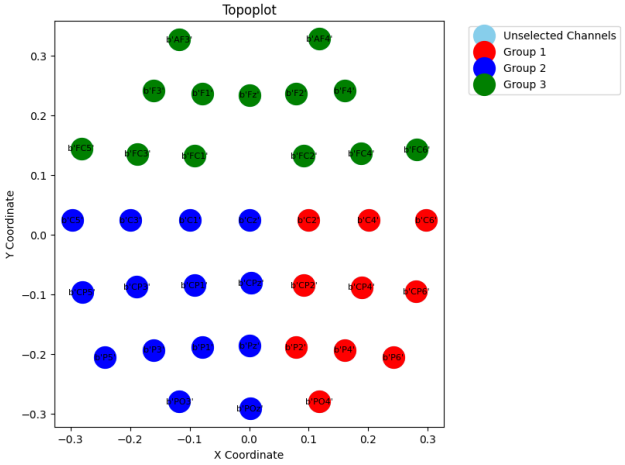


Fig. 3: Hipòtesi de possibles comunitats

Podem veure que existeix diferència en els valors de potència, tant entre classes (Mora, Syllable, Stress), com entre grups de canals segons la posició (esquerra, centrals, dreta). Vegem la diferència entre grups Mora i Stress a les Fig. 4 i Fig. 5 entre canals frontals, que com es pot apreciar, s'agrupen molt bé entre ells. També observem que els canals PO, que es veuen en blau i taronja a la Fig. 6, s'allunyen de la resta de canals, tan agrupats per grups com per classe. Això passa en els canals PO3, POz i PO4. Aquests canals embrutarien el senyal a l'hora de col·lapsar-los per grups segons la posició. A més, són els canals més exteriors dels que encara no han estat descartats, i això implica una gran probabilitat que el seu senyal es trobi contaminat. És per això que es decideix eliminar aquests canals del grup de canals usats. A partir d'ara, treballarem amb els 34 canals restants.

### 5.2.2 Reducció de dimensionalitat amb les dades granulars

Per a la detecció de comunitats, s'ha aplicat l'algorisme de Louvain sobre la matriu de correlació de canals en cada trial i subject diferenciat. Això dóna com a resultat una agrupació diferent per a cada trial i subject. Volem saber quin és

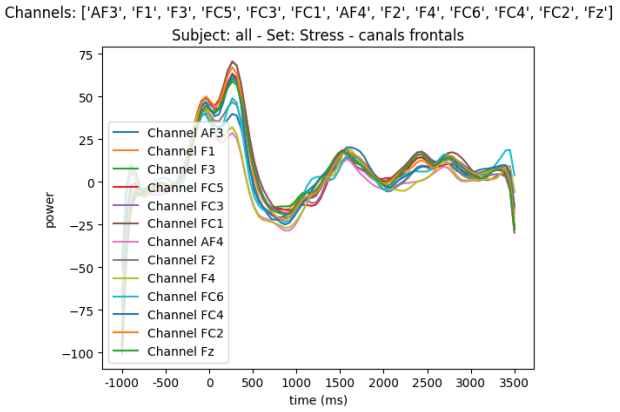


Fig. 4: Power x Time separant les dades en grups i classes: Stress, canals frontals

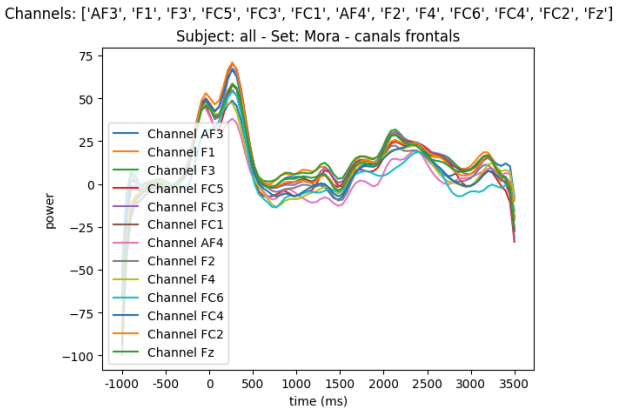


Fig. 5: Power x Time separant les dades en grups i classes: Mora, canals frontals

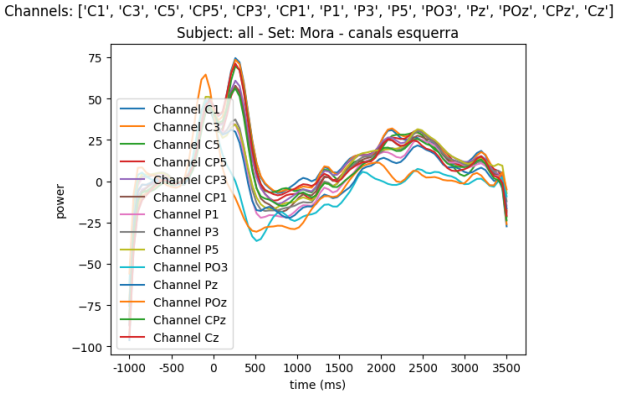


Fig. 6: Power x Time separant les dades en grups i classes: Mora, canals esquerra

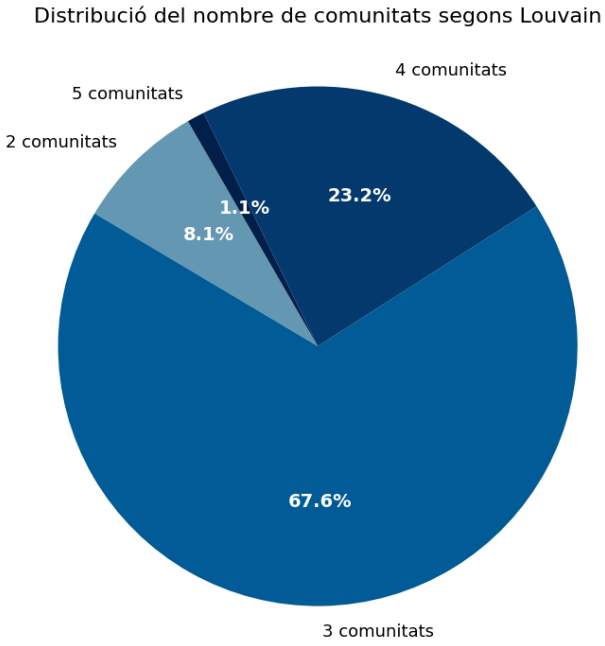


Fig. 7: Distribució del nombre de comunitats trobades cada vegada que s'executa Louvain (per trial i subject)

el nombre de comunitats òptim, que respecti al màxim les comunitats locals creades en cada trial. És per això que fem un recompte de quantes comunitats han sortit en cadascuna de les vegades que hem aplicat Louvain Fig. 7.

Aquest gràfic mostra clarament que el nombre de comunitats que hem de buscar és de 3. Ja que en la gran majoria de vegades l'algorisme ha trobat 3 comunitats. Així doncs, reforcem la hipòtesi plantejada en els assaigs previs amb dades col·lapsades i la idea que podem separar de manera òptima els canals en tres comunitats segons la seva localització (frontal, esquerra i dreta).

Per a buscar aquestes 3 comunitats, el següent pas ha estat crear una matriu quadrada on per cada parella de canals guardem el nombre de vegades que aquests han coincidit en una mateixa comunitat per l'algorisme de Louvain, calculat com en el cas anterior sobre la matriu de correlació entre els canals. En total, aquest algorisme s'executa més de 5040 vegades (35 trials x 3 classes x 48 subjects). Són més de 5040, ja que alguns dels grups contenen més de 35 trials.

Per a crear aquesta matriu, hem creat una matriu de zeros quadrada de mida 34 x 34 per apuntar les vegades que cada parella de canals coincidia en un grup. Després, hem iterat sobre les més de 5000 agrupacions, i en cadascuna d'elles hem sumat 1 al valor de la parella a la matriu, en cas que fossin part de la mateixa agrupació. El resultat és una matriu quadrada amb un nombre més alt per a les parelles que més freqüentment formen part del mateix grup. Hem visualitzat aquesta matriu quadrada en forma de graf, on cadascun dels nodes pertany a un canal i cadascuna de les relacions al valor de cada parella a la matriu Fig. 8. Per a una millor visualització, es mostra el 60% de les relacions més fortes, i les arestes creixen en amplada i intensitat del color quan el seu valor és major.

En aquest cas, a la Fig. 8, trobem una distinció clara entre els dos grups principals, un agrupa els nodes superiors i l'altre els inferiors. Dit això, podríem separar, basant-nos

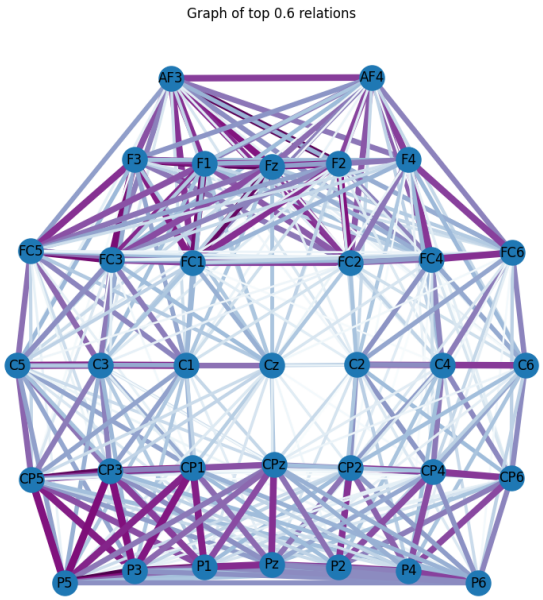


Fig. 8: Graf que relaciona els canals amb pesos segons les vegades que han format part de la mateixa comunitat

en l'anàlisi visual, el grup inferior en dos grups més, on les quatre primeres columnes formen el grup esquerre, i les 3 últimes el grup a la dreta, seguint així la hipòtesi plantejada anteriorment.

A la Fig. 8, trobem que la quarta fila de canals podria pertànyer visualment tant al grup de canals superiors com als grups inferiors.

En l'últim pas per a l'agrupació de canals, utilitzem les agrupacions individuals per trial i subject, que hem calculat per a decidir el nombre de comunitats òptim a la Fig. 7. Ordenant per freqüència aquestes agrupacions, vegem que la més repetida (18 vegades) és exactament l'agrupació proposada anteriorment 3, però sense els canals inferiors que hem eliminat (PO3, PO4 i POz). Tot i repetir-se només 18 vegades, ja apareix el doble que la segona i la tercera agrupació més repetida, i aquestes mostren només petites variacions, sobretot en els canals de la quarta fila. Tenint en compte que a l'agrupació més repetida tots els canals de la quarta fila formen part dels grups de l'esquerra i la dreta, i que en les altres agrupacions més repetides només alguns d'aquests formen part del grup de canals frontals, però mai tots a la vegada, queda definida l'agrupació de canals com a la 3, però eliminant els canals PO.

## 5.3 Model de classificació

### 5.3.1 Context

És important comprovar si les diferències entre grups per ritme sil·làbic es reflecteixen de manera clara i mesurable en el senyal EEG per tal que un model sigui capaç de classificar-les. Per aquest motiu, hem generat un gràfic on es mostra l'evolució de la potència del senyal al llarg del temps per a cada una de les tres condicions (Mora, Syllable i Stress). A la Fig. 9 es mostra la potència mitjana calculada amb tots els trials de cada grup. A més, s'hi ha afegit l'error estàndard (SEM) per visualitzar la variabilitat entre

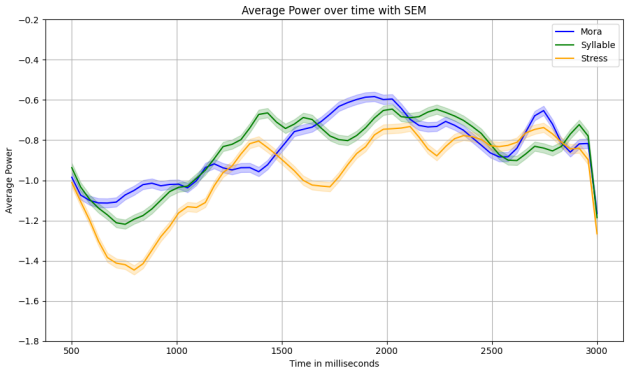


Fig. 9: Comparació Power mitjana calculada amb tots els trials, segons grups (Mora, Syllable, Stress), usant només dades extretes d'experiments on la llengua nativa del participant era l'anglès.

trials dins de cada classe.

El fet que aquestes tendències siguin diferenciables fa pensar que un model com la regressió logística multiclasse podria captar aquestes relacions a partir de features derivades directament del senyal EEG. S'han realitzat tres versions d'aquest gràfic diferenciant per l'idioma natiu de la persona que participa. Tant per a natius en llengua anglesa, castellana i un últim on hem usat els dos conjunts. Hem observat que les diferències són més grans en els gràfics que comparen les dades de la mateixa llengua. Vegem a la Fig. 9, les dades per a participants nadius en llengua anglesa.

### 5.3.2 Construcció del model

S'ha construït un model de classificació binària per predir a quin ritme lingüístic correspon cada trial a partir de característiques temporals del senyal EEG. S'usaran les dades Mora i Stress, ja que les llengües del grup Mora generen una activitat EEG més regular i sincronitzada, amb valors més alts de PLV i menys complexitat temporal, al contrari que les llengües del grup Stress, i a més són els dos grups més allunyats entre ells. No utilitzem el grup Syllable per així dur a terme una classificació binària simplificant l'estudi.

La unitat experimental o sample a classificar prové de l'activitat neuronal d'una persona(subject) per a un trial concret. Les dimensions d'aquest sample són 3x60 (60 valors de potència al llarg del temps per a les 3 agrupacions de canals). Aquestes dades passen per un procés de binarització i s'utilitzen per a calcular les features del model.

Per la seva interpretabilitat, s'ha optat per una classificació usant l'algoritme SVM i un kernel no lineal i pel model de regressió logística, ambdós models s'aplicaran i es compararan els seus resultats per tal de veure quin és més eficient. El percentatge de divisió de dades train/test serà del 70% a train i 30% a test. També s'ha decidit que s'usaran 7 features per cada grup de canals. En total 21 features per trial. Com a mesura de precaució, farem servir regularització L1, que tria automàticament les més útils per al model.

### 5.3.3 Features del model

Les features escollides per al model han de representar correctament la informació que creiem que és essencial per a

distingir entre grups (Mora i Stress). És per això que hem triat un conjunt de features variat, procurant que mesurin tant el nivell de la potència com la seva variabilitat i evolució en el temps.

Per a algunes d'aquestes features, hem calculat prèviament la binarització del senyal a partir de la potència mitjana més la desviació estàndard calculada per trial i subject. Per a binaritzar el senyal, s'han convertit a 0 tots els valors que no superaven la mitjana més una desviació típica i a 1 tots aquells més grans o iguals a aquest valor. Aquesta és una tècnica que ens permet reduir la dimensionalitat de les dades, aplicant ara càlculs sobre les dades binaritzades que definiran les features. Vegem a Fig. 10 i Fig. 11 com el senyal es binaritza.

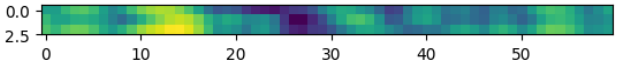


Fig. 10: Potència agrupada per grups de canals al llarg del temps sense binaritzar, un sol trial d'exemple

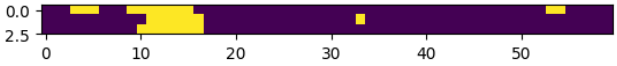


Fig. 11: Potència agrupada per grups de canals al llarg del temps binaritzada, un sol trial d'exemple

Calcularem les següents features:

- **Potència mitjana:** Valor mitjà de la potència. Indica el nivell mitjà d'activitat. És calculada a partir dels valors de potència directament
- **Desviació típica:** La desviació típica de la potència indica la variabilitat d'aquesta.
- **Màxims i mínims locals (indicador pos/neg):** Codifica els canvis de tendència. És calculada a partir del senyal derivat.

La resta són features calculades a partir de les dades binaritzades.

- **Nombre de transicions 0→1:** Comptatge de canvis al senyal binari, mesura la variabilitat del senyal.
- **Longitud màxima de ratxa de 1s i de 0s:** Indica la persistència d'activitat o inactivitat. D'aquestes dues features és previst escollir-ne la més informativa.
- **Moment temporal mitjà dels 1s:** Reflecteix quan tendeixen a aparèixer els pics d'activitat.
- **Complexitat de Lempel-Ziv (LZ) binaritzada:** Mesura l'heterogeneïtat/irregularitat global del patró.

### 5.3.4 Anàlisi de features

L'anàlisi de les features és un pas clau per a conèixer quina informació pot aportar cadascuna al model, si cal modificar el càlcul d'alguna d'elles o fins i tot eliminar-ne i afegir-ne de noves. Per això hem volgut visualitzar el comportament de les features calculades a partir de dades Mora respecte a les calculades amb dades Stress. Els trials Mora han estat reduïts aleatòriament per a usar el mateix nombre de trials Mora que Stress.

Degut a les diferències que trobem entre les dades que provenen dels experiments realitzats en persones natives en llengua castellana i anglesa, hem decidit visualitzar la diferència entre Mora i Stress sobre dades d'una sola llengua nativa, aquesta és l'anglès a menys que es puntualitzi el contrari. Aquesta diferència entre en les dades segons la llengua nativa, ja és present al paper (Silva Pereira et al., 2024), on s'analitzen diferències entre els valors de la LZ-complexity entre aquests dos grups. És per això que s'entrenarà el model també en versions on només s'utilitzin les dades d'una mateixa llengua.

En el cas de la **potència mitjana**, s'observa que els valors per a Mora són lleugerament superiors als de Stress, però aquesta diferència és mínima i les distribucions es troben massa superposades Fig. 12. Aquesta superposició massiva indica que la potència mitjana, per si sola, seria una feature poc útil per discriminar entre grups, ja que no captura una separació clara a nivell poblacional.

Pel que fa a la **desviació típica**, les diferències només són apreciables en el subgrup de participants nadius de llengua espanyola Fig. 13. En aquest cas concret, Stress mostra una variabilitat lleugerament major, però aquest patró no es reprodueix en parlants nadius d'anglès ni en l'anàlisi global. Aquesta, però, és una diferència clara, que podria ser útil en un model específic per a dades on la llengua materna és el castellà.

Quant a la complexitat **Lempel-Ziv**, l'histograma inicial mostra una distribució molt similar a la de les altres features: les figures per a Mora i Stress gairebé se superposen, amb només una lleugera diferència en les mitjanes. Aquesta semblança inicial podria fer pensar que, com passa amb la potència mitjana o les transicions, aquesta feature tampoc aporta informació útil per al model. No obstant això, per explorar millor el seu potencial, vam anar més enllà de l'anàlisi estàtica i vam calcular l'evolució temporal de la LZ complexity. En lloc d'usar un únic valor global per trial, vam calcular la complexitat en cadascun dels 60 punts temporals i vam traçar les corbes mitjanes per als grups Mora i Stress Fig. 14. En aquest gràfic visualitzem Mora i Stress per a les dues llengües possibles separades durant els últims 10 valors (on la diferència és més notable). Vegem que la diferència més gran entre elles és entre Mora i Stress quan la llengua nativa és el castellà. Tot i que les línies no arriben a creuar-se, la distància entre elles és mínima al llarg de tot l'interval temporal. És a dir, aquesta diferència és tan petita que resulta poc rellevant a la pràctica. Així doncs, no considerem útil al model aquest càlcul temporal tampoc. Per al model utilitzarem el valor de la complexitat LZ de tota la seqüència, que correspon a l'últim valor del gràfic.

En relació a la resta de features derivades de la binarització (**transicions 0→1, longitud de ratxes d'uns/zeros, màxims locals i moment temporal dels pics**), els histogrames revelen distribucions pràcticament idèntiques per als dos grups Fig. 15. La similitud en els patrons de variabilitat temporal, la persistència de les ratxes d'activitat i la distribució temporal dels pics d'energia és notable, fins al punt que les corbes de freqüència se superposen gairebé completament. Aquesta manca de divergència indica que cap d'aquestes mesures, en la seva forma actual, aporta informació discriminativa rellevant per al model.

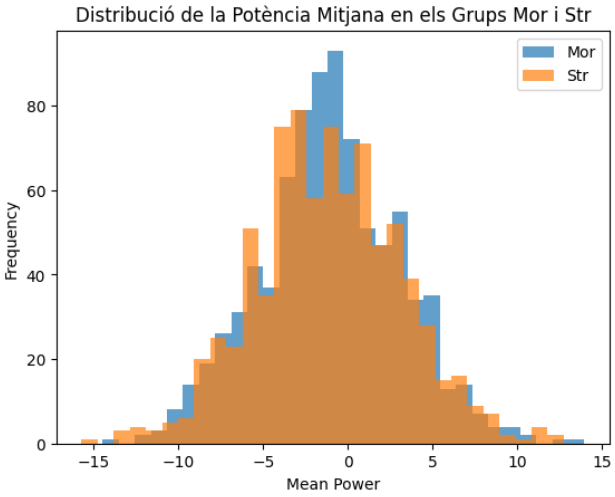


Fig. 12: Distribució de la mitjana de la potència per a Mor i Str

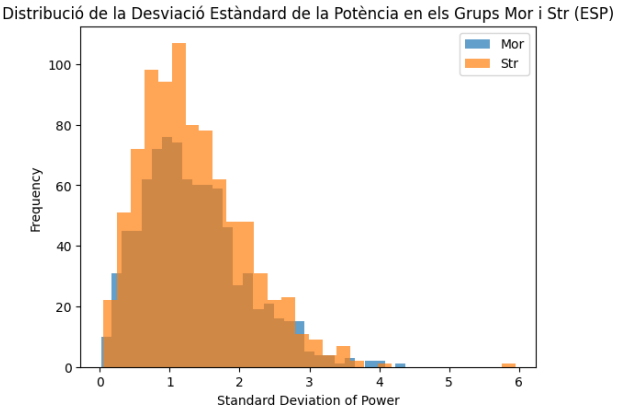


Fig. 13: Distribució de la desviació típica de la potència per a Mor i Str

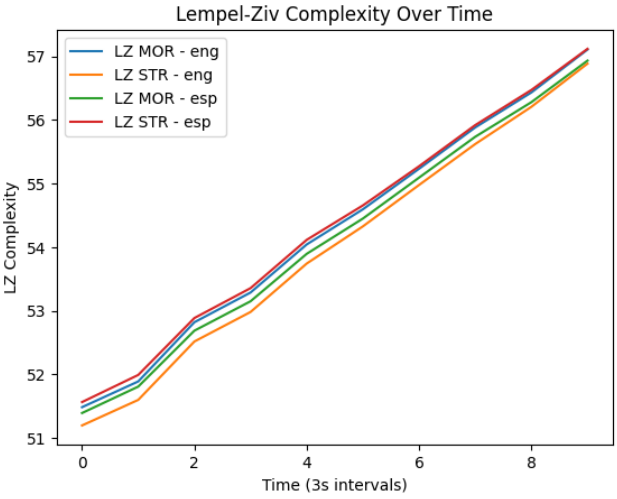


Fig. 14: Evolució dels últims 10 valors de la complexitat LZ, Mora vs Stress

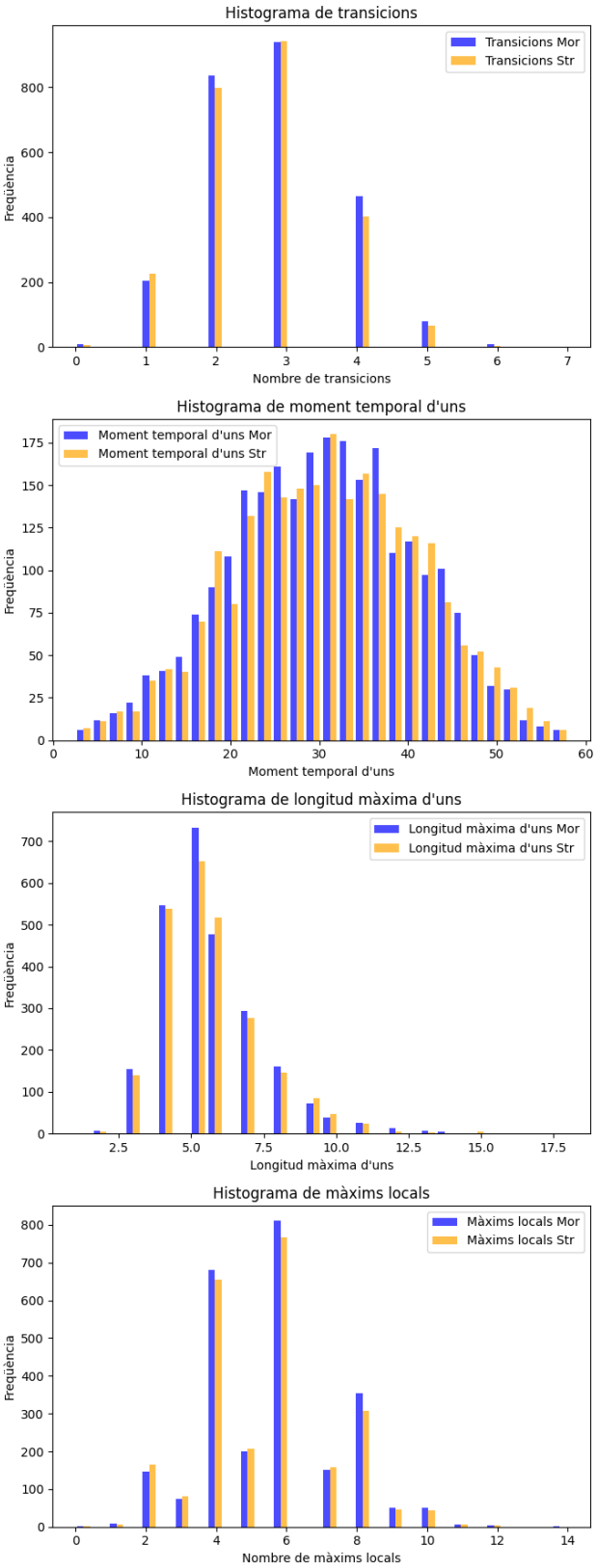


Fig. 15: Histogrames que mostren el nombre de trials de cada grup segons els valors de les features corresponents.

### 5.3.5 Evolució del model

Tot i que l'anàlisi preliminar de les features ja indicava que no eren especialment discriminatives, vam decidir provar diferents configuracions de models per descartar qualsevol possibilitat d'èxit. Els resultats, però, van confirmar les sospites inicials: cap de les variants provades va assolir una eficiència destacable, amb precisions properes a l'atzar en tots els casos.

Vam experimentar amb diferents kernels no lineals amb SVM i lineals amb la regressió logística, però ni la versió lineal ni la no lineal van mostrar millores significatives. El kernel no lineal, tot i la seva capacitat teòrica per capturar relacions complexes, no va poder extreure patrons útils de les features disponibles, cosa que suggereix que el problema no rau en la linealitat de les dades sinó en la manca d'informació discriminativa en les features seleccionades.

Per descartar que el problema fos degut a la variabilitat entre llengües maternes, vam entrenar models separats: un amb totes les dades, un altre només amb participants nadius d'anglès i un tercer només amb nadius de castellà. Cap d'aquestes aproximacions va millorar els resultats.

També vam provar a reduir les features a les teòricament més prometedores (mitjana, desviació típica i LZ complexity), el model va simplificar-se, però la precisió va romandre igualment baixa. La poca separabilitat intrínseca d'aquestes variables va impedir qualsevol progrés significatiu.

En resum, els models van fallar sistemàticament, independentment de la tècnica emprada. Aquesta consistència en els mals resultats apunta a un problema fonamental: les features actuals no capturen les diferències reals entre els ritmes Mora i Stress en aquest context experimental. La manca de divergència clara en les distribucions, juntament amb la inconsistència entre llengües maternes, explica per què cap dels diferents models no va poder generalitzar.

## 6 Resultats

Malauradament, cap dels models avaluats va assolir resultats significatius, amb el millor model (regressió logística entrenada només amb dades de participants amb un anglès nadiu i les variables de potència mitjana, desviació estàndard i LZ complexity) mostrant una accuracy de 51±0.1% i AUC de 0.55 - lleugerament superior a l'atzar però insuficient per a una classificació útil. La Fig. 16 revela que aquest rendiment limitat podria atribuir-se a l'escassedat de dades, ja que la corba d'aprenentatge mostra que la predicció en dades test millora amb més mostres d'entrenament, tot i així els resultats són massa propers a l'atzar per assegurar que la classificació milloraria amb un volum de dades major. Això, unit a la superposada distribució de característiques entre classes, confirma que les variables seleccionades no capturen diferències discriminatives suficients en les dades EEG actuals.

La matriu de confusió mostra que, tot i que l'accuracy global del model és propera a l'atzar, no es tracta d'un simple biaix cap a una única classe. Les prediccions es distribueixen de manera relativament uniforme entre totes les categories (Mora i Stress), amb valors similars tant a la diagonal (encerts) com fora d'aquesta (errors). Això indica que el model no ha après a generalitzar patrons discriminatius Table 1.

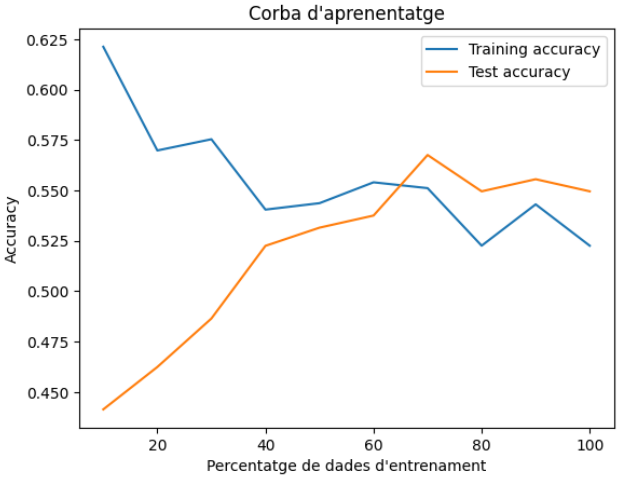


Fig. 16: Corba d'aprenentatge: mostra la accuracy en les dades train i test a mesura que el model s'entrena amb més dades.

TAULA 1: Matriu de confusió del model de classificació

| Real \ Predicció | Mora | Stress |
|---|---|---|
| **Mora** | 153 | 103 |
| **Stress** | 139 | 104 |

## 7 Conclusions

Les primeres exploracions realitzades amb el conjunt de dades més simple, van facilitar la familiarització amb l'estructura del dataset, la identificació preliminar de tendències globals i la validació del flux de treball, tot i que amb una resolució reduïda. Les observacions fetes durant aquesta etapa, com l'agrupament natural dels canals frontals o la presència de soroll en certs canals, van ser confirmades més endavant amb les dades granulars, demostrant la utilitat d'aquesta aproximació simplificada. Així, les proves amb dades avg no només van servir com a punt de partida per a l'anàlisi posterior, sinó que també van proporcionar una base sòlida per a prendre decisions informades en etapes més avançades del projecte.

Respecte a la reducció de dimensionalitat de les dades granulars, durant el treball hem pogut analitzar en profunditat el comportament dels diferents canals EEG. Mitjançant tècniques com l'algoritme de Louvain, hem observat que canals situats a posicions properes presenten patrons d'activitat similars. Aquest fet ens ha permès identificar una estructura clara en tres grups diferenciats segons la seva localització: frontal, esquerra i dreta. Aquesta agrupació, que coincideix amb la hipòtesi plantejada durant el període de primeres exploracions, s'ha demostrat consistent al llarg de múltiples trials i subjectes. La validació d'aquesta agrupació mitjançant la freqüència en anàlisis independents i l'anàlisi visual, confirma que s'ha dut a terme amb èxit l'agrupació de canals per a la reducció de dimensionalitat de les dades.

Els resultats de la classificació mostren que les features utilitzades no han capturat adequadament les diferències en l'activitat neuronal entre els ritmes lingüístics. Això es reflecteix en el baix rendiment dels models de classificació, que no van superar l'atzar. La causa principal podria ser que

les diferències reals en la sincronització neuronal són més subtils del que s'havia previst, i no queden ben reflectides en les mesures temporals simples que vam emprar. També és un hàndicap el limitat volum de dades per entrenar el model, sobretot quan només entrenem amb dades de participants d'una sola llengua nativa.

Com a possibles millores, es podria aplicar finestres temporals per analitzar fases concretes del processament del llenguatge, en lloc de considerar tot el senyal de manera global. Això permetria detectar diferències subtils en els moments clau de sincronització neuronal. Una altra via seria augmentar la mida del dataset, ja que amb més dades es podrien obtenir resultats més robustos. Així, es podrien provar mètodes més avançats, com models de deep learning (per exemple, xarxes neuronals convolucionals o transformers), que podrien capturar patrons complexos que els enfocaments tradicionals no han pogut detectar.

## AGRAÏMENTS

Vull expressar el meu agraïment a Silvana Silva Pereira per la seva tutorització i guia durant tot el treball. Així mateix, vull reconèixer a la UAB per les eines i formació rebudes durant el grau en Enginyeria de Dades i agrair al Grup de Recerca en Adquisició i Percepció de la Parla (SAP) de la UPF per facilitar l'accés al conjunt de dades.

# APÈNDIX

## A.1 LZ-Complexity

La Complexitat de Lempel-Ziv (LZ) és una mesura de la diversitat de patrons en una seqüència (contributors, 2025). Es defineix com el nombre de subcadenes diferents identificades en analitzar una seqüència de manera incremental. Formalment, per a una seqüència $S$ de longitud $n$, la complexitat $C(S)$ reflecteix el nombre mínim de passos necessaris per reconstruir $S$ mitjançant còpies de subcadenes prèviament observades (Ruffini, 2017).

L'algorisme clàssic de Lempel-Ziv es resumeix en els següents passos:

1. **Inicialització**: Dividir la seqüència en subcadenes començant pel primer símbol.
2. **Cerca de patrons**: Per a cada posició $i$, trobar la subcadena més llarga que ja hagi aparegut anteriorment.
3. **Divisió**: Si no es troba una coincidència, afegir un nou símbol al diccionari de patrons.
4. **Complexitat**: El resultat final $C(S)$ és el nombre total de subcadenes úniques.

En aquest treball, usem la LZ-complexity com una de les features en els models de classificació. Aquesta ens aporta informació sobre la regularitat del ritme, i ajuda el model a diferenciar entre les 3 classes de trials.

## A.2 Mètode de Saltanaj

El mètode de Resíntesi Saltanaj és una tècnica utilitzada per crear estímuls auditius en estudis sobre el processament de la parla, amb l'objectiu específic de manipular les propietats rítmiques del discurs tot eliminant-ne el significat i la familiaritat. (Silva Pereira et al., 2024)

El mètode consisteix a substituir els fonemes de les oracions originals pels seus equivalents lingüístics. Aquests equivalents són fonemes considerats comuns en la majoria de llengües.

- Totes les fricatives es substitueixen per /s/.
- Totes les vocals es substitueixen per /a/.
- Tots els líquids (laterals i ròtics) es substitueixen per /l/.
- Totes les consonants oclusives (stops) es substitueixen per /t/.
- Totes les nasals es substitueixen per /n/.
- Totes les semivocals es substitueixen per /j/.

**Exemple:** Si l'oració original en anglès *"The next local elections will take place during the winter"* es transforma mitjançant aquest mètode en *"sa natst latl alatsans jal taat tlaas tjalan sa janta"*. (Ramus, 2025)

## A.3 Esquema de fases

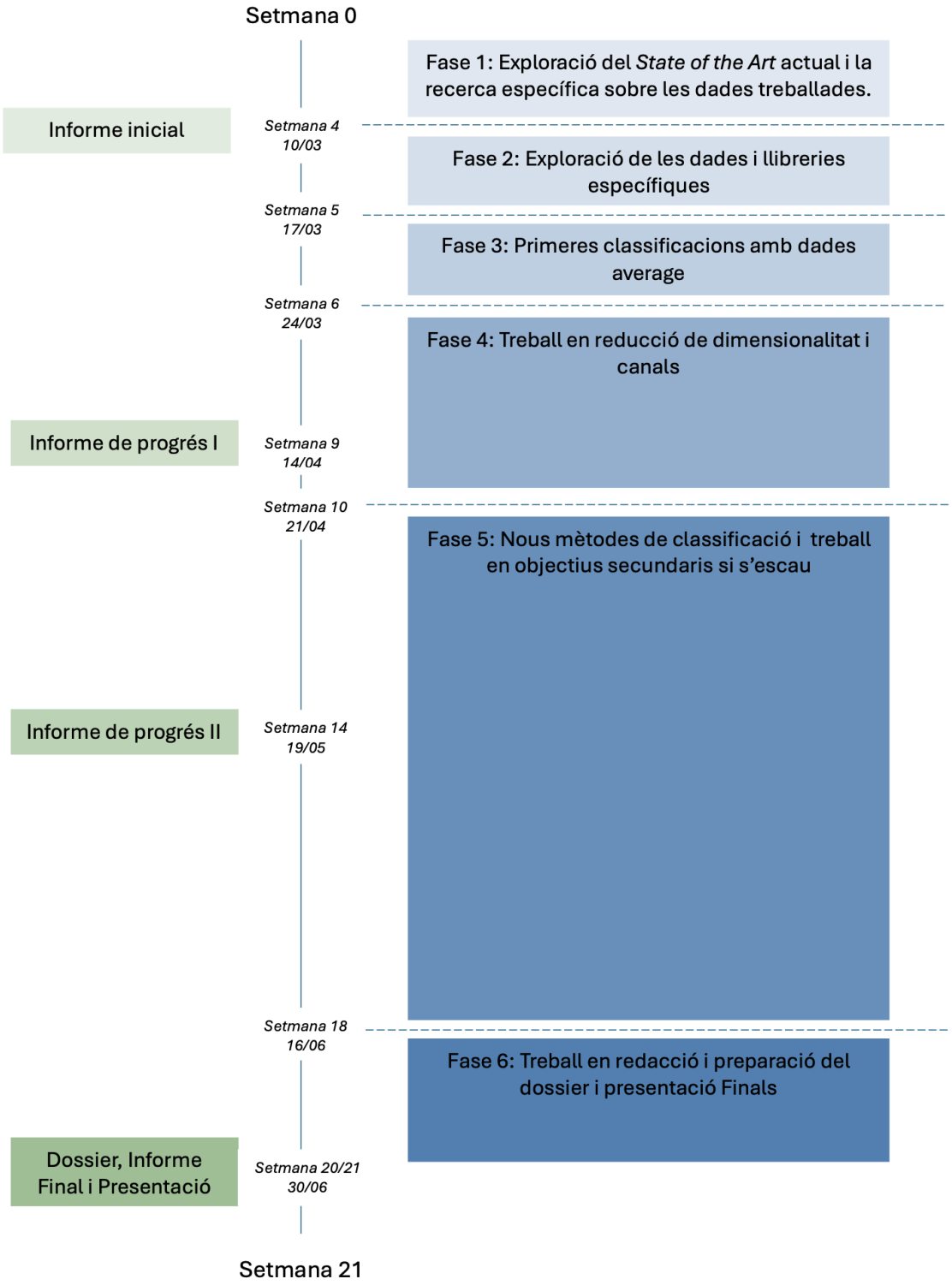


Fig. 17: Esquema de fases del treball.

## REFERÈNCIES

contributors W (2025) Lempel–ziv complexity Accés el 21 de maig de 2025.

Ekin Özer E, Silva Pereira S, Ciarrusta J, Sebastian-Galles N (2023) Neural entrainment in theta range is affected by speech properties but not by the native language of the listeners. *bioRxiv* bioRxiv preprint.

Frontiers in Psychiatry (2025) Identifying relevant eeg channels for subject-independent emotion recognition using attention network layers Pàgina web de Frontiers in Psychiatry Accedit el 14 de juny de 2025.

He Y, Shaw JA, Price CJ, Ding N (2024) The cortical encoding of hierarchical linguistic rhythms reflects language experience. *Cerebral Cortex* 34:619–627.

Namburi P (2011) Programming language vulnerabilities. *Praneeth Namburi's Blog* .

Ramus F (2025) Ecoute de mots resynthétisés Online Last accessed: [Insert Date].

Ruffini G (2017) Lempel-ziv complexity: A framework for biomedical signal analysis. *arXiv* .

Schwartz RS (2000) Biological aging and the genesis of atherosclerosis. *Journal of the American College of Cardiology* 35:1651–1652.

Silva Pereira S, Özer EE, Sebastian-Galles N (2024) Complexity of STG signals and linguistic rhythm: a methodological study for EEG data. *Cerebral Cortex* 34:bhad549.

Varghese L, Grube M, Richardson D, Goswami U (2022) Neural speech tracking is modulated by rhythmic class. *bioRxiv* .

Wikipedia contributors (2023a) Análisis tiempo-frecuencia [Online; accés 15-Juny-2024].

Wikipedia contributors (2023b) Louvain method — Wikipedia, the free encyclopedia [Online; accés 14-Juny-2024].