Informe de progrés 2: Classificació de senyals EEG del cervell humà mitjançant tècniques d'aprenentatge automàtic

David Martí

---

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 de la cortesa auditiva, 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 Pereira et al. [2024] 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. Aquesta sincronització és calculada a partir del Phase Locking Value (PLV) A.1

2 Objectius

L'objectiu general d'aquest treball és la classificació, mitjançant un model de Machine Learning, de les dades segons el seu ritme lingüístic. És possible diferenciar els senyals EEG obtingudes durant l'escolta de fragments de frases a partir del ritme sil·làbic de l'idioma.

Per a assolir-lo 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

Diverses investigacions han mostrat que les llengües es poden agrupar segons el seu ritme lingüístic —Mora-timed, Syllable-timed i Stress-timed—, el qual es relaciona amb la regularitat temporal de les síl·labes Pereira et al. [2024], Özer et al. [2023]. S'ha comprovat que la sincronització cerebral en la banda theta és sensible a aquesta regularitat, independentment de la llengua materna dels participants. En concret, els idiomes amb ritmes més regulars, com els del grup Mora-timed, mostren uns valors més alts de Phase-Locking Value (PLV), cosa que suggereix que són més fàcils de seguir per part del còrtex auditiu Pereira et al. [2024], Özer et al. [2023].

Estudis recents han integrat el càlcul de la complexitat temporal mitjançant l'índex de Lempel-Ziv (LZ). Aquesta mètrica mesura la diversitat en els patrons d'un senyal Wikipedia contributors [2025]. El Superior Temporal Gyrus (STG) és una àrea del cervell crítica per al processament de la parla. S'ha observat que a la part esquerra del STG, la LZ-Complexity augmenta quan el ritme de la parla és menys regular (amb valors majors per a Stress-timed i menors per a Mora-timed). Aquesta tendència és coherent amb el comportament del PLV, ja que una menor variabilitat sil·làbica (com la que caracteritza Mora) afavoreix una sincronització més estable Pereira et al. [2024].

4 Metodologia i planificació

Per al seguiment de la feina realitzada durant el transcurs del treball s'ha fet un esquema de fases, 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 realització dels diferents informes de progrés i finals.

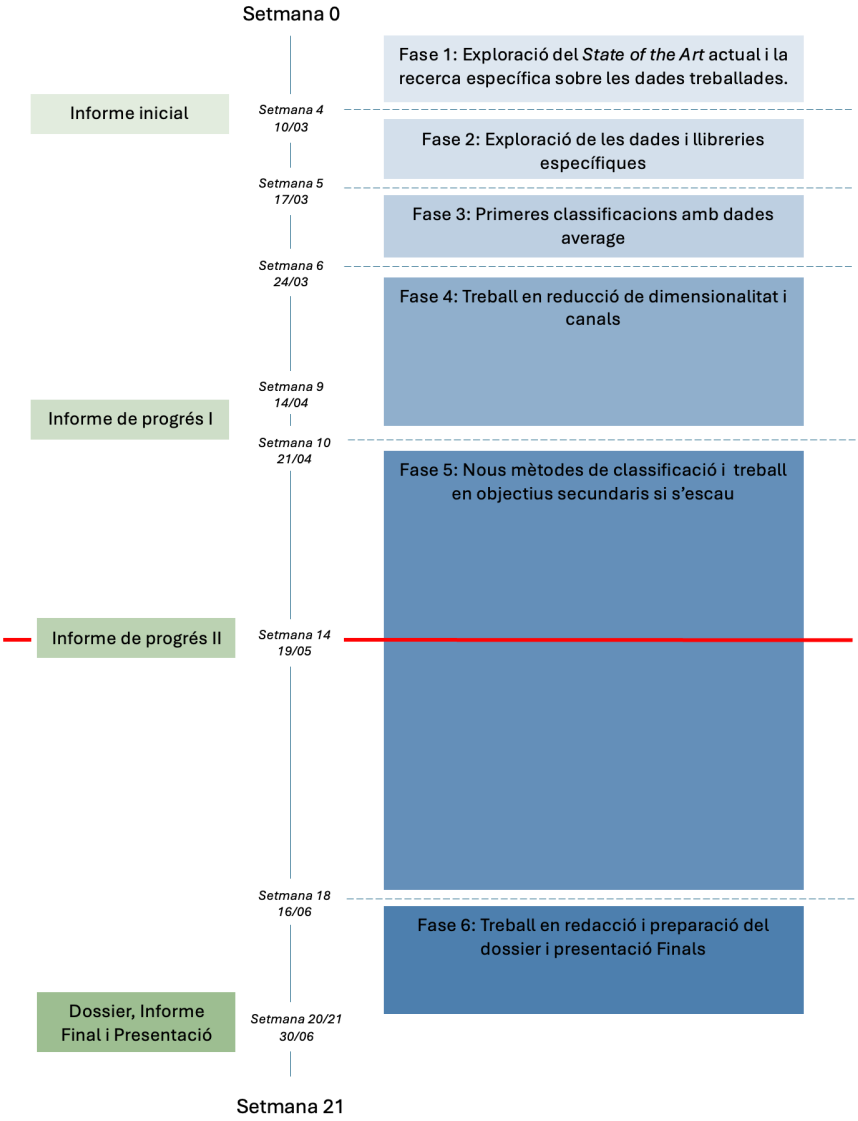

Figura 1: Esquema de fases del treball.

Per assegurar una bona pràctica a l'hora de programar, es crea un arxiu on trobem diferents funcions personalitzades per a mostrar les dades en les quals treballem sense necessitat d'executar el codi cada vegada. Aquestes funcions permeten agrupar per canals, subjects, grups de ritme i grups segons la llengua nativa. Així com un topoplot amb grups de canals diferenciats.

4.1 Canvis

Fins al moment, les etapes del treball no han estat alterades. S'han experimentat contratemps durant l'agrupació de canals per a la reducció de la dimensionalitat de les dades. La falta de consistència en els resultats després de generalitzar les comunitats a partir dels grups obtinguts individualment, han fet triar les comunitats en base a proves visuals amb aquest propòsit. Així doncs, s'ha començat amb cert retard l'etapa de classificació.

5 Context de les dades

Per a la creació del dataset es va comptar amb un total de 48 participants, dividits entre parlants nadius d'anglès i d'espanyol (24 de cada llengua) Pereira et al. [2024]Özer et al. [2023].

Es van utilitzar 60 d'oracions resintetitzades, repartides en parts iguals per ritmes lingüístics de la llengua en la qual es graven (Mora, Syllable i Stress). Es va aplicar el mètode de resíntesi Saltanaj A.3 per tal de mantenir la complexitat sil·làbica de les oracions originals, tot eliminant-ne el significat i així evitar possibles efectes de comprensió Pereira et al. [2024]Özer et al. [2023].

La unitat bàsica d'anàlisi després del preprocessament és un "trial", que correspon a la presentació d'una sola oració a un participant. Per a cada trial, el senyal EEG es va segmentar en una finestra temporal que abastava des d'un segon abans fins 3,5 segons després de l'inici de l'oració. Les dades originals es van enregistrar amb 60 canals EEG. 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.

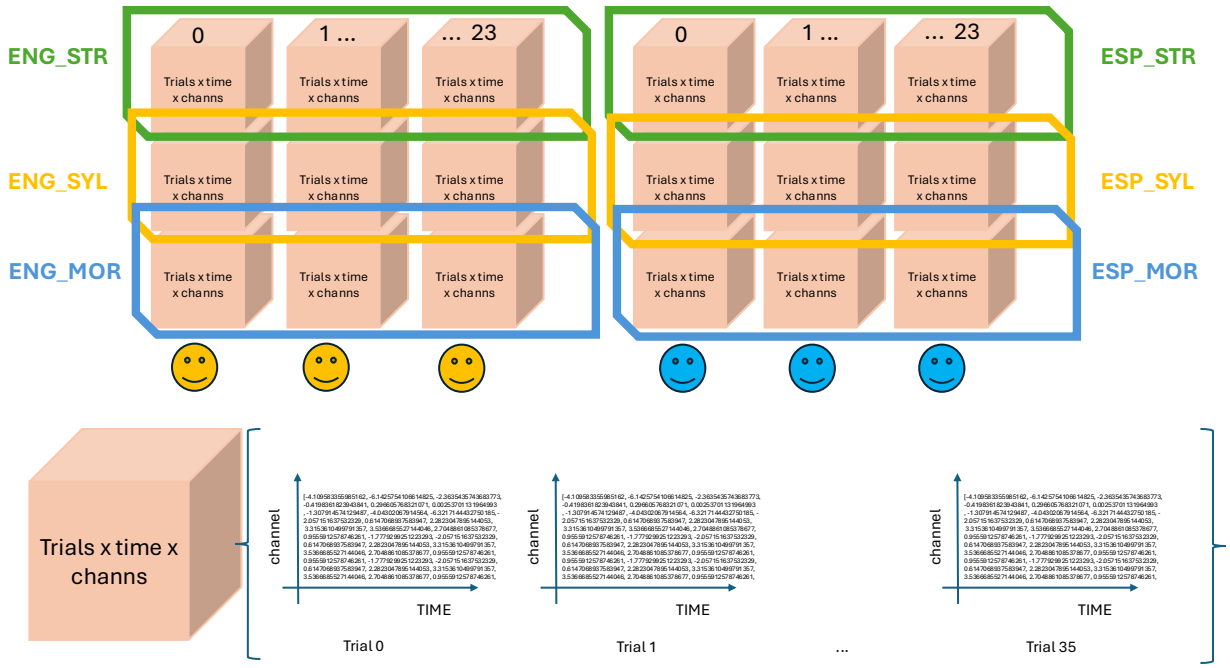

Figura 2: Esquema del dataset.

6 Desenvolupament

6.1 Assajos amb dades average

Durant les primeres setmanes, s'ha treballat amb un dataset on la dimensió "trial"és col·lapsada aplicant la mitjana. Les dimensions són: 48 subjects x 3 classes x 60 canals x 90 punts de temps. D'aquests 60 canals, se'n descarten 23, com s'ha fet en investigacions anteriors, ja que els canals més exteriors contenen molt de soroll i un senyal alterat.

En el primer pas cap a la reducció dels canals, s'ha fet una matriu de correlació entre canals per a cada subject.

D'aquesta manera, identifiquem mitjancçant un valor concret per a cada parella de canals, la semblança en el seu comportament i a la vegada, serveix com a pes de les arestes en el graf sobre el qual podem aplicar l'algorisme de Louvain. S'ha creat una matriu amb la mitjana de tots aquests valors i aplicat l'algorisme de detecció de comunitats Louvain a aquesta matriu. Els resultats mostren dues grans comunitats, una situada als canals superiors i una altra per als inferiors.

El següent 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. En el següent histograma 3 trobem les vegades que cada canal ha estat seleccionat com un dels 4 més importants, després d'executar PCA per a cadascun dels subjects (144). Destaquem en aquest cas que els canals amb més importància són canals a la part esquerra del mapa.

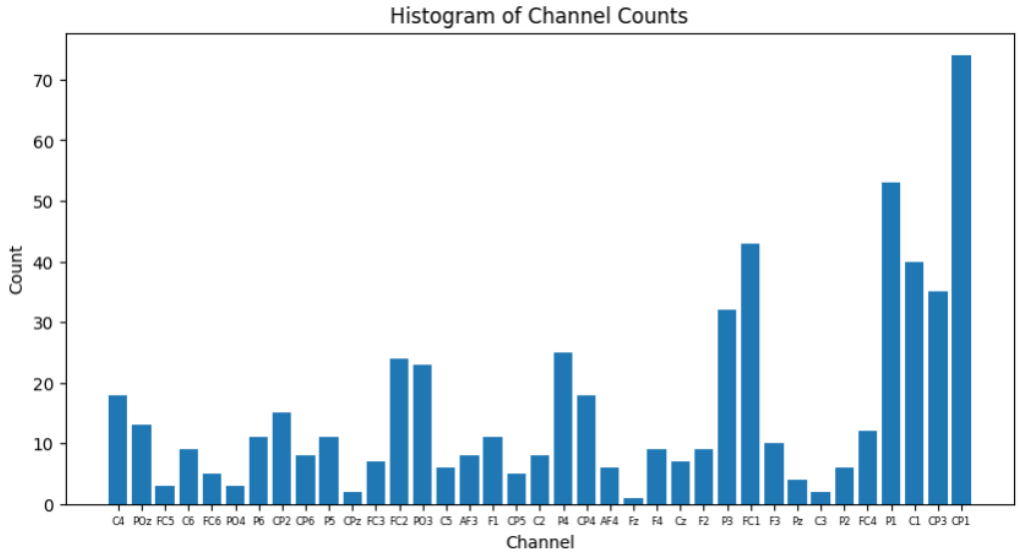

Figura 3: Canals més repetits en els resultats de PCA.

Entre tots els canals en destaquen els canals 'C1' 'FC1' i 'P1' com els canals més repetits. Aquests pertanyen tots a la part central esquerra del cervell, aquesta informació és coherent amb investigacions prèvies, que relacionen l'STG esquerre amb el processament de la parla. 4.

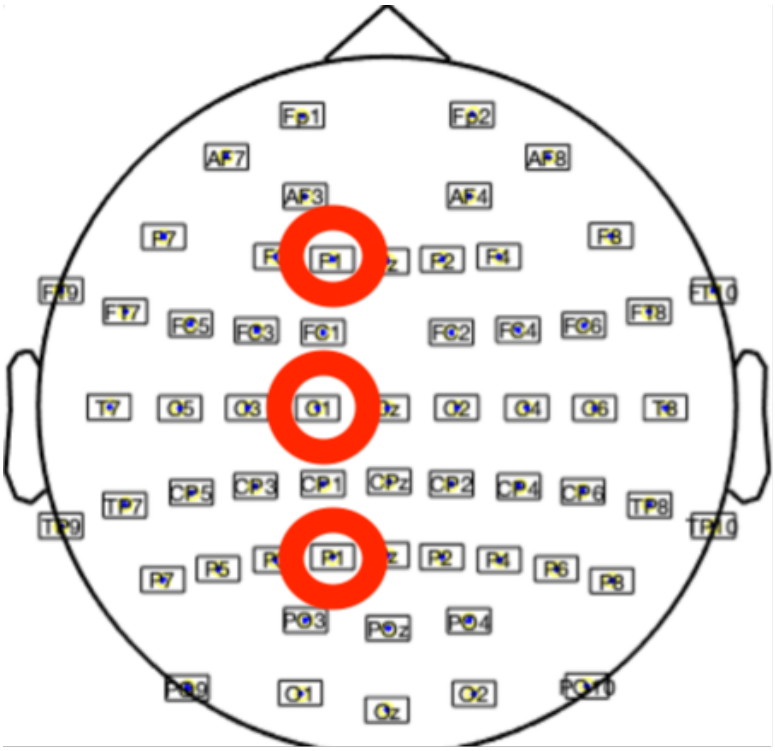

Figura 4: Canals destacats.

Aquesta informació ens indica quins canals són importants, però no els separa en comunitats, seguint els passos de la investigació feta sobre aquestes dades Pereira et al. [2024], buscarem trobar diferències entre tres grups de canals: canals frontals, canals a l'esquerra i canals a la dreta des de la visió. Per tant, en els següents intents per a detectar comunitats s'intentarà trobar una detecció similar a aquesta.

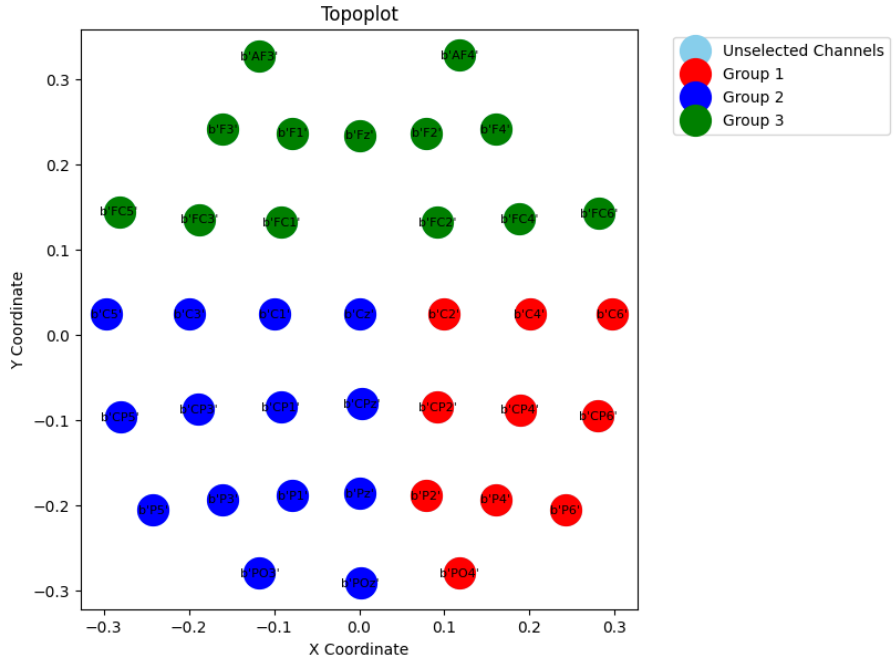

Figura 5: Hipòtesi de possibles comunitats

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

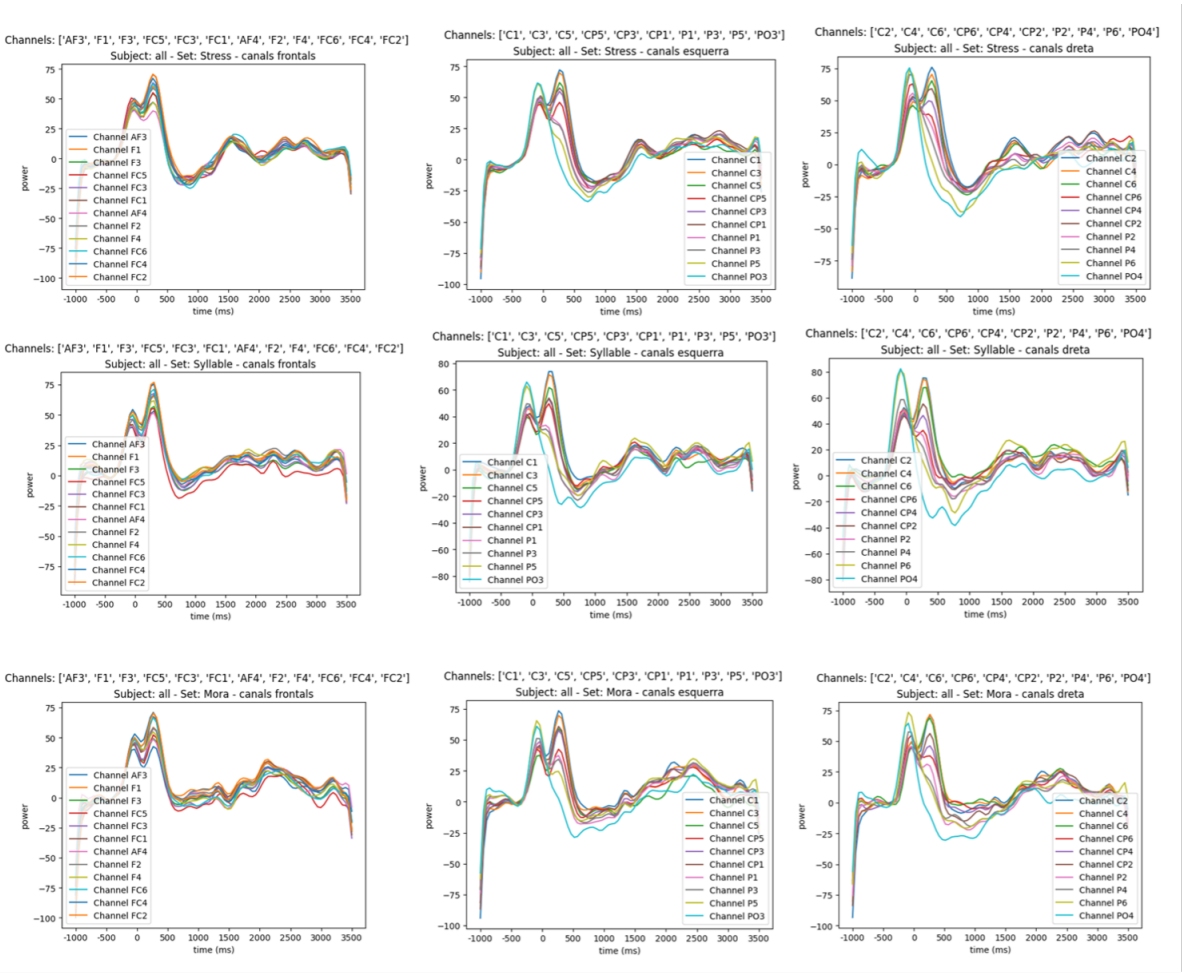

Figura 6: Power x Time en dades Mora

Podem veure a la figura 5, que existeixen diferències significatives en els valors de potència, tant entre classes (Mora, Syllable, Stress), com entre grups de canals (esquerra, centrals, dreta). També observem, que els canals PO, que es veuen en blau a les figures de canals esquerre i canals drets, s'allunyen de la resta. És per això que es decideix eliminar aquests canals del grup de canals usats. A partir d'ara, treballarem amb els 34 canals restants.

6.2 Preprocessament de les dades

A partir d'aquest punt, s'ha usat un dataset amb les dades sense agrupar per trial ni freqüència. Durant el següent pas es fa un preprocessament d'aquest conjunt de dades. Per a realitzar aquest pas s'han canviat detalls a les funcions i alguna nomenclatura en el codi. Per a continuar treballant sense perdre coherència hem seguit el mateix tractament en la freqüència que s'havia realitzat en les dades avg en anteriors recerques Pérez et al. [2015]. Primerament, s'han eliminat els valors per a freqüències fora del rang entre 5 i 8 Hz. Després s'ha col·lapsat del tot 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 on els senyals encara no han estat afectades pel so emès en l'experiment, que correspon als primers 30 punts de temps. El resultat és un dataset amb les

dimensions: 48 subjects x 3 classes x 35 trials (aprox(.) x 37 canals x 90 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 previament.

6.3 Reducció de dimensionalitat

Explicació de tot el procés realitzat per a la divisió de comunitats i gràfics d'on finalment acabem traient les conclusions. Cal explicar el doble Louvain,

Per a la detecció de comunitats, s'ha aplicat l'algorisme de Louvain sobre la matriu de correlació de canals de cada trial. El següent histograma mostra el recompte de trials que han obtingut des d'una a sis comunitats.

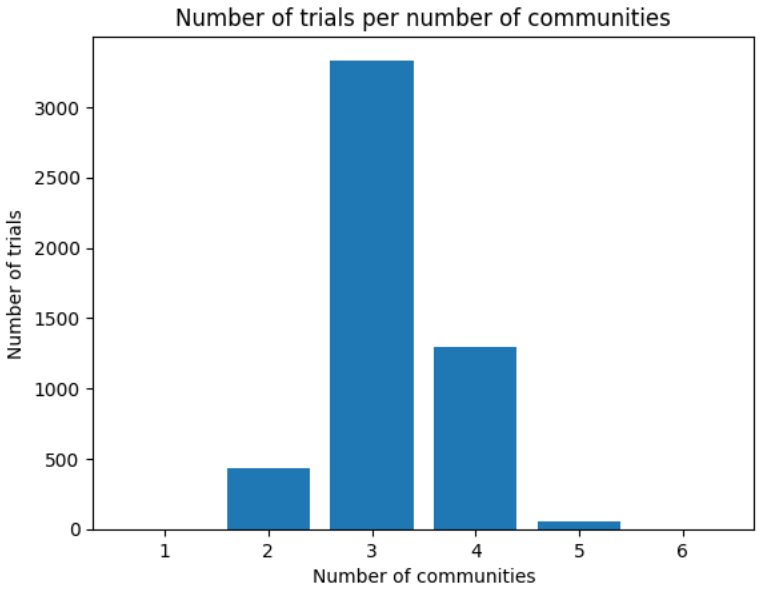

Figura 7: Recompte de trials per nombre de comunitats trobades

Aquest histograma 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.

Per a buscar aquestes 3 comunitats, el següent pas ha estat crear una matriu quadrada on per cada parella, s'han sumat les vegades en les que han coincidit en la mateixa comunitat. D'aquesta matriu en surt un graf, on els nodes són els diferents canals i les arestes el nombre de vegades en què dos canals han coincidit en la mateixa comunitat. Sobre aquest graf s'ha aplicat un algorisme que va eliminant arestes de menys a més pes fins que queden només 3 comunitats connectades al graf. Usant aquesta metodologia, ens assegurem de guardar les relacions més comuns entre canals. El resultat ha estat la formació de tres comunitats de la següent forma.

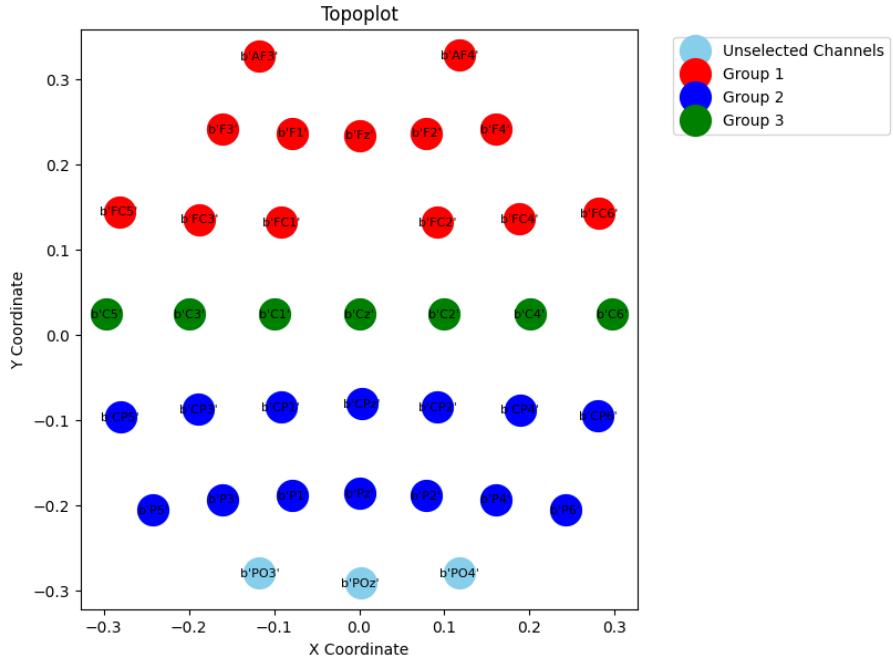

Figura 8: Comunitats generalitzades

Aquest topoplot, mostra tres comunitats alineades de forma horitzontal. Si bé aquest pas verifica que hi ha una diferència important entre els canals superiors i inferiors, no mostra la importància que té la posició lateral dels nodes (part esquerra i dreta). Es continua buscant, doncs, una manera de trobar aquesta diferència en les comunitats.

També s'ha aplicat l'algorisme de Louvain sobre la mateixa matriu en què s'han recomptat les vegades en què cada parella de nodes apareix junta en la mateixa comunitat.

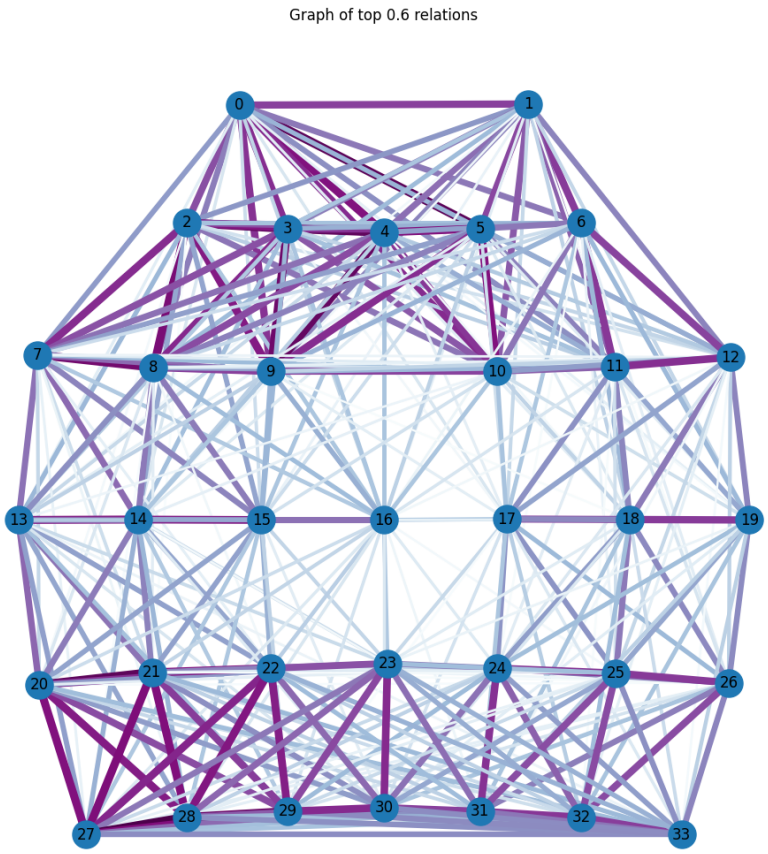

Figura 9: Relacions entre canals

En aquest cas, a la figura 8, trobem una distinció de dos grups principals, un agrupa els nodes superiors i l'altre els inferiors. Dit això, podem separar els grups inferiors en dos grups, on les quatre primeres columnes formen el grup esquerre, i les 3 últimes el grup a la dreta. Per a comprovar-ho, es visualitza la potència al llarg del temps, i es comparen els comportaments dels diferents grups de canals. Només es mostren dades de la classe Mora, ja que el comportament dels canals és diferent quan canvia la classe, i sumar-les totes embrutaria el gràfic.

A la figura 8, trobem que la quarta fila de canals podria pertànyer tant al grup de canals superiors com als grups inferiors. Per tant, s'ha decidit realitzar dues vegades el gràfic amb les diferents possibilitats d'agrupació. A la figura 9, els canals de la quarta fila han estat assignats al grup de canals frontals. En canvi, a la figura 10, han estat assignats als grups de canals inferiors, a l'esquerra i a la dreta depenent de la columna.

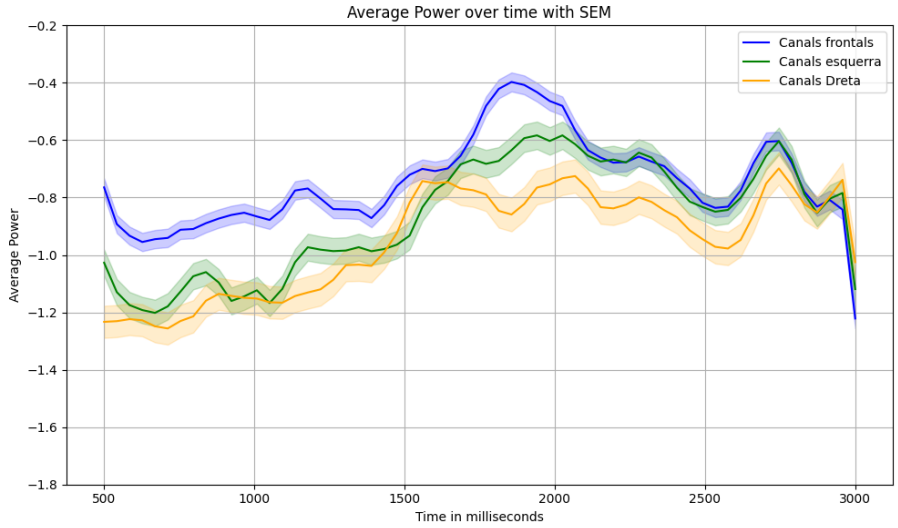

Figura 10: Power x time agrupant la 4a fila al grup superior

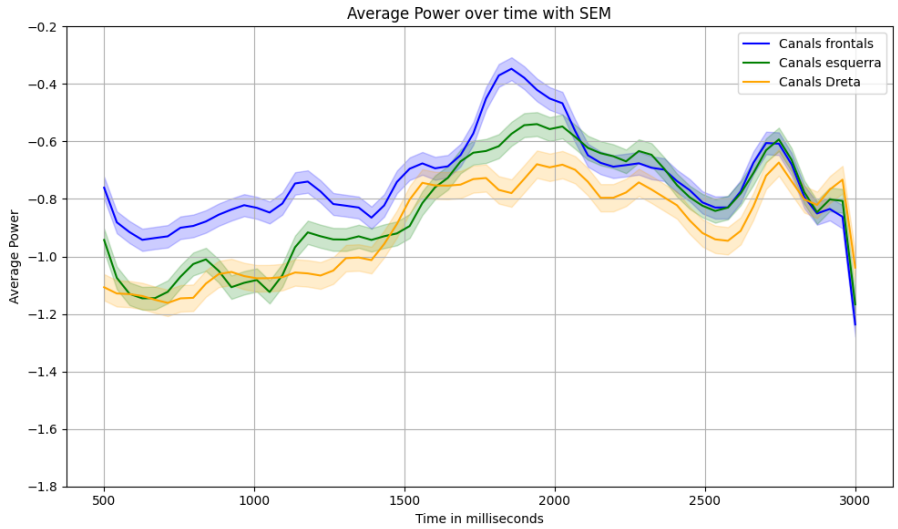

Figura 11: Power x time agrupant la 4a fila als grups inferiors

Basant-nos en les visualitzacions, hem decidit que la figura 10 mostra lleugerament millor les diferències entre canals, i per tant hem escollit l'agrupació que es mostra a la figura 11.

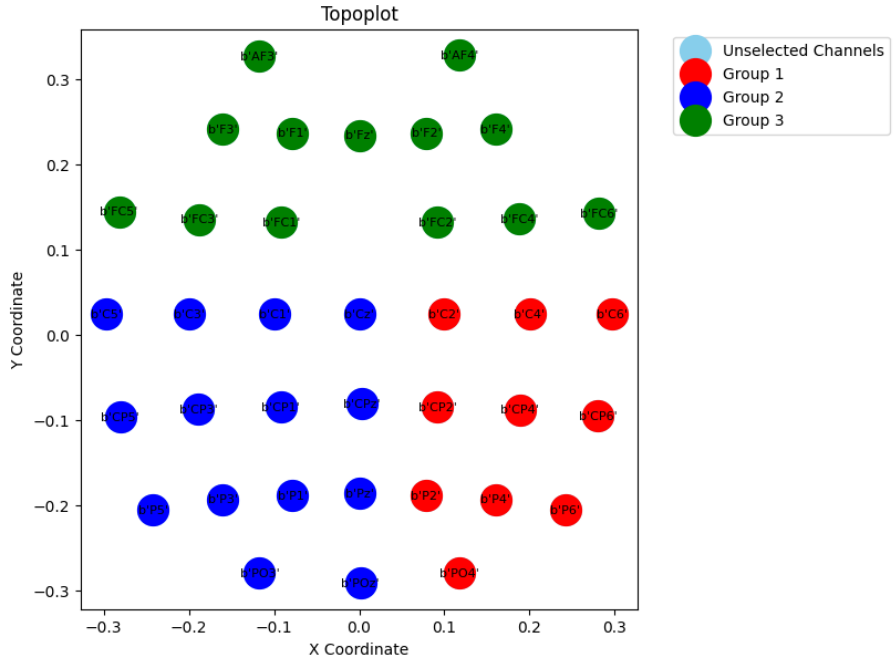

Figura 12: Agrupació de canals triada

6.4 Model de classificació

6.4.1 Context

És important comprovar si les diferències entre classes lingüístiques es reflecteixen de manera clara i mesurable en el senyal EEG. Per aquest motiu, hem generat un gràfic on es mostra l'evolució de la potència mitjana del senyal al llarg del temps per a cada una de les tres condicions (Mora, Syllable i Stress). A més, s'hi ha afegit l'error estàndard (SEM) per visualitzar la variabilitat entre trials dins de cada classe.

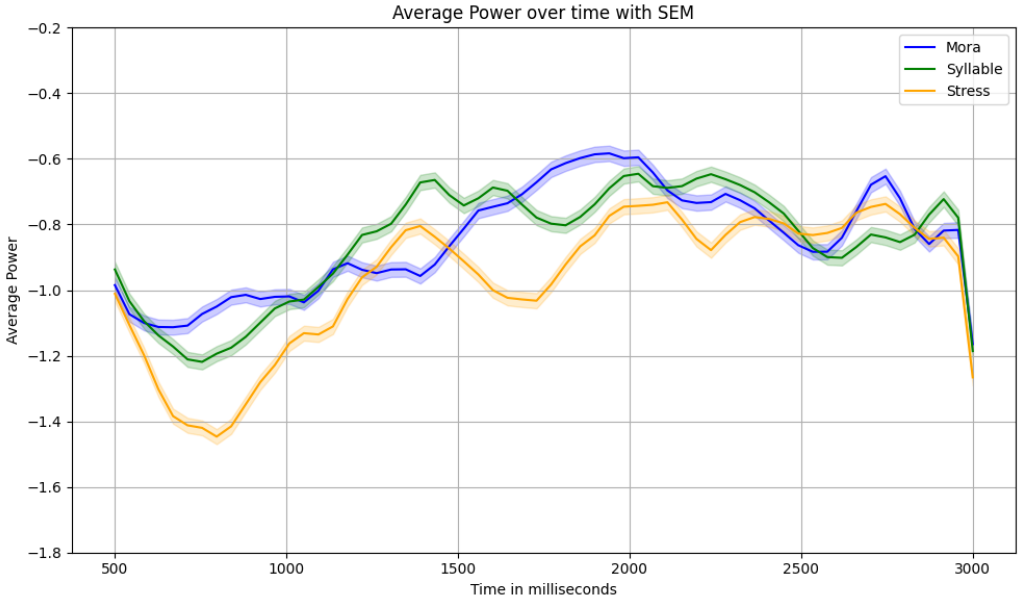

Figura 13: Comparació Power segons classes ambdós idiomes natius

El fet que aquestes tendències siguin clares i diferenciades fa pensar que un model com la regressió logística multiclasse podria captar aquestes relacions a partir de features derivades directament del senyal EEG. També s'han realitzat dues versions d'aquest gràfic diferenciant per l'idioma natiu de la persona que escolta. Les diferències són més grans, però creiem que és millor no crear més d'un model, per a tenir un volum adequat de dades en el training.

6.4.2 Construcció del model

S'ha construït un model de classificació multiclasse per predir a quin ritme lingüístic correspon cada trial a partir de característiques temporals del senyal EEG. Per la seva simplicitat i interpretabilitat, s'ha optat per una regressió logística multiclasse.

Per complementar el model multiclasse, s'ha entrenat també un model de classificació binària per distingir específicament la classe Mora de les altres dues (Syllable i Stress). Aquesta decisió es basa en l'evidència que les llengües del grup Mora generen una activitat EEG més regular i sincronitzada, amb valors més alts de PLV i menys complexitat temporal. Aquesta regularitat fa que creguem que Mora sigui més fàcil de distingir, fet que pot augmentar la precisió del model i facilitar la interpretació dels resultats.

6.4.3 Features del model

Per tal de transformar les sèries temporals en representacions binàries, i calcular les features del model, primer es realitza una binarització basada en la mitjana del senyal. Concretament, cada mostra s'etiqueta com a 1 si el seu valor és superior a la mitjana del trial corresponent, i com a 0 en cas contrari.

Per a aquests dos models, s'ha decidit que usarem 6 features per cada grup de canals. En total 18 features per trial. Com a mesura de precaució, farem servir regularització L1, que tria automàticament les més útils per al model.

Calcularem les següents features:

- **Proporció d'1s:** percentatge de mostres igual a 1 per trial, indica el nivell mitjà d'activitat.
- **Nombre de transicions 0→1:** comptatge de canvis binaris al senyal, mesura la variabilitat.
- **Longitud màxima de ratxa de 1s i de 0s:** indica la persistència d'activitat o inactivitat.
- **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.
- **Derivada binaritzada (indicador pos/neg):** codifica tendències globals d'increment o decrement.

6.4.4 Comportament del model

Aquí explicarem amb gràfics el comportament dels models i diferents enfocaments rellevants.

6.5 Resultats dels classificadors

Aquí mostrem els resultats del model gràficament i els expliquem.

7 Conclusions parcials

Tot i tenir definit el model i les features que s'usaran. Encara falta desenvolupament per a poder analitzar els resultats de la classificació.

Respecte a la reducció de dimensionalitat. Durant el treball hem vist com es comporten els canals. Sabem que canals propers es comporten de manera semblant. I no només això, sinó que podem agrupar-los en 3 grups segons la seva localització. Així doncs, s'ha dut a terme amb èxit l'agrupació de canals per a la reducció de dimensionalitat de les dades.

A Conceptes

A.1 Phase Locking Value (PLV)

El Phase-Locking Value (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 què és rellevant el PLV en aquest estudi?

- El PLV permet mesurar fins a quin punt les neurones del còrtex auditiu (com l'àrea STG) es "sincronitzen" o segueixen el ritme de la parla escoltada.
- Un PLV més alt indica una major sincronització, és a dir, una millor "sintonia" entre l'activitat cerebral i el ritme del senyal de parla.
- Estudis previs han mostrat que les llengües amb ritmes més regulars, com les de tipus mora-timed, generen valors de PLV més elevats que les syllable-timed o stress-timed. Això suggereix que els ritmes més regulars són més fàcils de seguir per part del còrtex auditiu. Pereira et al. [2024]

A.2 LZ-Complexity

La Complexitat de Lempel-Ziv (LZ) és una mesura de la diversitat de patrons en una seqüència Wikipedia 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 al model a diferenciar entre les 3 classes de trials.

A.3 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. 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 [Accessed: 2025]

Referències

Praneeth Namburi. Programming language vulnerabilities. *Praneeth Namburi's Blog*, August 2011. URL https://praneethnamburi.com/2011/08/10/plv/.

Ege Ekin Özer, Silvana Silva Pereira, Judit Ciarrusta, and Nuria Sebastian-Galles. Neural entrainment in theta range is affected by speech properties but not by the native language of the listeners. *bioRxiv*, 2023. doi: 10.1101/2023.07.11.548540. bioRxiv preprint.

Silvana Silva Pereira, Ege Ekin Özer, and Nuria Sebastian-Galles. Complexity of STG signals and linguistic rhythm: a methodological study for EEG data. *Cerebral Cortex*, 34:1–13, 2024. doi: 10.1093/cercor/bhad549. URL https://doi.org/10.1093/cercor/bhad549.

A. Pérez, M. Carreiras, M. Gillon Dowens, and J. A. Duñabeitia. Differential oscillatory encoding of foreign speech. *Brain Language*, 147:51–57, 2015.

Franck Ramus. Ecoute de mots resynthétisés. Online, Accessed: 2025.

Giulio Ruffini. Lempel-ziv complexity: A framework for biomedical signal analysis. *arXiv*, 2017.

Robert S. Schwartz. Biological aging and the genesis of atherosclerosis. *Journal of the American College of Cardiology*, 35(6):1651–1652, May 2000. doi: 10.1016/S0735-1097(00)00613-0. URL https://pubmed.ncbi.nlm.nih.gov/10619414/.

Wikipedia contributors. Lempel–ziv complexity, 2025. URL https://en.wikipedia.org/wiki/Lempel%E2%80%93Ziv_complexity. Accés el 21 de maig de 2025.