Clasificación de Respuestas EEG en Infantes mediante TDA en Grafos de Conectividad Funcional

Nombre: Maria Ignacia Gothe
Correo classroom: mariaignacia.gothe@gmail.com
Correo UC: mgothe@uc.cl
Fecha: 28 de noviembre 2025

Abstract

Este trabajo investiga si las características topológicas de los grafos de conectividad funcional del electroencefalograma (EEG) permiten distinguir entre infantes escuchando audio con velocidad normal versus audio acelerado. Se analizaron 1,416 grabaciones EEG de 45 bebés de 3-5 meses utilizando Análisis de Datos Topológicos (TDA), específicamente homología persistente sobre grafos de conectividad construidos mediante correlación entre electrodos. Se extrajeron 220 características de los diagramas de persistencia para las dimensiones H0 (componentes conexas) y H1 (ciclos) en cinco bandas de frecuencia (delta, theta, alpha, beta, gamma). Un clasificador Random Forest con validación cruzada a nivel de sujeto (GroupKFold, k=5) alcanzó una precisión del 89.3% ± 1.7% (F1=0.893, ROC-AUC=0.951), significativamente superior al azar ($p<0.001$, d de Cohen=23.3). Las bandas beta (26.9%) y theta (23.4%) mostraron mayor poder discriminativo. Los resultados demuestran que el TDA captura diferencias robustas en la topología de conectividad cerebral inducidas por la velocidad del audio, sugiriendo patrones de procesamiento neural distintos en respuesta a estímulos auditivos con diferente tasa temporal.

Introducción

El estudio de cómo los infantes procesan el lenguaje hablado es fundamental para comprender el desarrollo cognitivo temprano. Investigaciones previas han demostrado que los bebés muestran preferencia por el habla dirigida a infantes (infant-directed speech), caracterizada por un tempo más lento, mayor entonación y pausas más prolongadas. Sin embargo, los mecanismos neurales subyacentes a estas preferencias permanecen parcialmente inexplorados. El electroencefalograma (EEG) proporciona una ventana no invasiva a la actividad cerebral con alta resolución temporal, siendo especialmente adecuado para estudios con poblaciones pediátricas. Tradicionalmente, el análisis de EEG se ha basado en métricas de amplitud, potencia espectral o conectividad funcional expresadas como valores escalares o matrices. Sin embargo, estos enfoques pueden perder información estructural importante sobre la organización topológica de las redes cerebrales.

El Análisis de Datos Topológicos (TDA) ofrece herramientas matemáticas para caracterizar la forma y estructura de datos complejos. En particular, la homología persistente permite identificar características topológicas (componentes conexas, agujeros, cavidades) que persisten a través de múltiples escalas, proporcionando descriptores robustos ante ruido y transformaciones continuas.

En este trabajo, aplicamos TDA a grafos de conectividad funcional derivados de EEG de infantes para responder la siguiente pregunta: ¿Pueden las características topológicas de la conectividad EEG distinguir entre infantes escuchando audio lento versus audio rápido? Nuestra hipótesis es que el

audio con velocidad normal (lento) induce patrones de conectividad cerebral topológicamente distintos comparados con el audio acelerado, y que estas diferencias son capturables mediante homología persistente y clasificables mediante aprendizaje automático.

Descripción del conjunto de datos

Se utilizaron señales EEG de varios canales recolectados de infantes mientras escuchaban un audio de una hablante femenina. El conjunto de datos comprende dos condiciones experimentales:

- **Audio lento (slow):** Velocidad normal de habla (~710 grabaciones)
- **Audio rápido (fast):** Mismo audio acelerado digitalmente (~706 grabaciones)

Los archivos siguen la nomenclatura bbXX_utYY.mat, donde XX identifica al sujeto (bebé) y YY el número de utterance (fragmento de audio). Se registraron 65 electrodos, de los cuales 47 fueron considerados válidos tras excluir aquellos ubicados en posiciones inferiores de la gorra que no asentaban correctamente en bebés de esta edad.

El conjunto de datos final comprende EEG distribuidas equitativamente entre dos condiciones experimentales: audio lento (710 muestras, 50.1%) y audio rápido (706 muestras, 49.9%). Los 45 sujetos contribuyeron con un promedio de 31 grabaciones cada uno (rango: 21-40).

Los datos presentan varios desafíos que se deben tener en consideración al momento de analizarlos. En primer lugar, la dependencia intra-sujeto inherente al diseño experimental implica que las grabaciones del mismo individuo están correlacionadas debido a características anatómicas y fisiológicas idiosincrásicas. La validación cruzada estándar que divide aleatoriamente las muestras podría asignar grabaciones del mismo sujeto tanto al conjunto de entrenamiento como al de prueba, resultando en fuga de datos y sobreestimación sistemática de la capacidad de generalización. Para abordar este problema, implementamos una estrategia de validación cruzada a nivel de sujeto (GroupKFold con k=5), donde cada fold asigna aproximadamente 36 sujetos al entrenamiento y 9 sujetos completamente independientes a la prueba, asegurando cero solapamiento entre conjuntos.

Por otro lado la variabilidad inter-sujeto tambien presenta un desafio, dado que diferencias individuales en anatomía cerebral, maduración neural y estados cognitivos introducen heterogeneidad sustancial en los patrones de conectividad. Si bien esta variabilidad podría dificultar la detección de los patrones relacionados a distinguir entre cerebro en distintas condiciones expermientales, también fortalece la capacidad de clasificar datos nuevos ya que se tienen suficientes para entrenar un modelo en de individuos con características diversas por lo que logra generalizar bien. El diseño de agregación temporal (media y desviación estándar a través de ventanas) ayuda a estabilizar las estimaciones de características al promediar sobre fluctuaciones de corto plazo dentro de cada grabación.

Por ultimo, se desconocen los detalles de la creacion del set de datos por lo que potencialmente pueden haber variables de confusión que no fueron controladas explícitamente en este análisis. La edad específica de cada infante, el momento del día de las grabaciones, el estado de alerta o fatiga durante la sesión, y el orden de presentación de los audios (ej; primero lento o primer el rápido) podrían influir en los patrones de conectividad observados. Sin embargo, bajo el supuesto de asignación aleatoria de estas condiciones cualquier efecto sistemático de estas variables debería

distribuirse entre los conjuntos de entrenamiento y prueba por lo que su impacto en las estimaciones de rendimiento no debiese ser muy significativo

Preprocesamiento

Primero, cada registro EEG se descompuso en cinco bandas de frecuencia canónicas:

- Delta (0.5–4 Hz): asociada a procesos de sueño profundo y regulación atencional.
- Theta (4–8 Hz): vinculada a memoria de trabajo y segmentación temporal.
- Alpha (8–13 Hz): relacionada con estados de relajación e inhibición cortical.
- Beta (13–30 Hz): asociada a alerta, mantenimiento de estado y predicción temporal.
- Gamma (30–100 Hz): ligada a integración perceptual y procesos cognitivos de mayor frecuencia.

Esta descomposición por bandas permite estudiar cómo la velocidad del habla modula redes oscilatorias en distintos rangos de frecuencia, en lugar de trabajar con un único espectro global.

Sobre cada señal filtrada se aplicó una segmentación temporal en ventanas deslizantes (sliding windows) de duración fija y solapamiento constante. De este modo, cada ensayo se representa como una secuencia de ventanas de un segundo de duracion (250 muestras), 75 % de solapamiento que capturan la evolución temporal de la conectividad durante la presentación del estímulo.

Para cada ventana y cada banda de frecuencia se calculó una matriz de correlación de Pearson entre todos los pares de electrodos (47×47). Estas matrices de correlación cuantifican la conectividad funcional lineal entre regiones, midiendo la similitud temporal de las señales EEG. Sin embargo, los algoritmos de homología persistente operan de forma más natural sobre métricas de distancia que sobre medidas de similitud, por lo que las matrices de correlación se transformaron en matrices de distancia mediante la regla

$$\mathrm{d_{ij}} = 1 - |\mathrm{r_{ij}}|,$$

donde r_{ij} es el coeficiente de correlación entre los electrodos i y j. Con esta transformación, valores cercanos a 0 representan conexiones fuertemente correlacionadas (alta similitud) y valores cercanos a 1 representan conexiones débiles. Además, se impuso explícitamente que la diagonal fuera cero y que las matrices fueran simétricas, lo que garantiza las propiedades básicas requeridas para construir un complejo de Vietoris–Rips a partir de la matriz de distancias.

El resultado de este preprocesamiento es, para cada ensayo y cada banda de frecuencia, una colección de matrices de distancia bien definidas que sirven como entrada directa al pipeline de TDA descrito en la siguiente sección.

Proceso de Análisis Topológico de Datos

A partir de cada matriz de distancias se extrajeron características topológicas mediante Análisis Topológico de Datos (TDA). Para ello, se construyó un complejo de Vietoris–Rips sometido a una filtración progresiva: inicialmente cada electrodo se considera un punto aislado y, a medida que el umbral de distancia aumenta, se van añadiendo aristas entre pares de electrodos cercanos y, posteriormente, símplices de orden superior (triángulos, tetraedros, etc.). Esta filtración genera una familia de grafos anidados que describe cómo se organiza la red de conectividad a distintas escalas.

Sobre esta filtración se calculó la homología persistente en dos dimensiones: H0 (componentes conexas) e H1 (ciclos). Las clases de H0 corresponden a grupos de electrodos que forman un conjunto conectado entre sí pero separado de otros grupos, mientras que las clases de H1 representan bucles cerrados de conectividad funcional. Cada característica topológica se describe por dos valores: su nacimiento (umbral en el que aparece por primera vez) y su muerte (umbral en el que desaparece o se fusiona con otra). La diferencia entre ambos define la persistencia; valores altos de persistencia indican estructuras topológicas robustas frente a cambios en el umbral, mientras que persistencias muy cortas suelen asociarse a ruido o a conexiones poco estables.

De cada diagrama de persistencia se extrajeron 11 características escalares para H0 y otras 11 para H1. Estas incluyen: número total de características finitas, número de características esenciales, media y desviación estándar de los tiempos de nacimiento, media y desviación estándar de los tiempos de muerte, media, desviación estándar, máximo y suma total de las persistencias, y la entropía de persistencia normalizada. Esta última cuantifica cuán concentrada o dispersa está la "masa" de persistencias: valores altos indican muchas características con persistencias similares, mientras que valores bajos reflejan que unas pocas estructuras dominan el diagrama. Estas 22 características por banda constituyen la representación topológica de cada matriz de conectividad que se utiliza posteriormente para el aprendizaje estadístico.

Aprendizaje Estadístico y Clasificación

Como cada grabación contiene múltiples ventanas temporales, las características se agregaron calculando la media y desviación estándar a través de todas las ventanas. Esto produce 44 características por banda de frecuencia (22 características × 2 estadísticos). Con cinco bandas, cada grabación quedó representada por un vector de 220 características. Este vector captura tanto las tendencias promedio de la topología de conectividad (mediante las medias) como su variabilidad temporal (mediante las desviaciones estándar), proporcionando una descripción completa de los patrones de conectividad durante el estímulo auditivo.

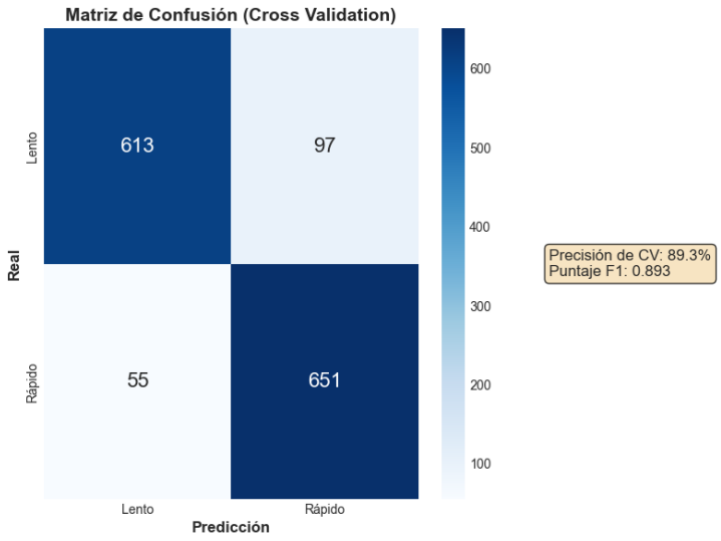
Como clasificador se utilizó Random Forest. Este modelo es adecuado por tres razones principales: es robusto frente a características ruidosas o poco informativas, maneja bien problemas con un número de variables relativamente alto respecto al tamaño muestral y entrega medidas de importancia de características que facilitan la interpretación científica. En lugar de asumir relaciones lineales simples, como haría una regresión logística, Random Forest puede capturar interacciones complejas entre las características topológicas sin requerir conjuntos de datos masivos ni arquitecturas profundas como en redes neuronales.

El modelo se configuró con 100 árboles de decisión y una profundidad máxima de 10 niveles. Estos parámetros son conservadores y habituales en la literatura: un número suficiente de árboles para

estabilizar las predicciones y una profundidad limitada para evitar que los árboles memoricen el conjunto de entrenamiento. No se realizó una búsqueda exhaustiva de hiperparámetros, porque el objetivo principal del trabajo no es optimizar al máximo la precisión, sino demostrar que las características topológicas contienen información discriminativa real. Mantener una configuración estándar y moderadamente regularizada también reduce el riesgo de sobreajuste asociado a la selección agresiva de hiperparámetros.

Resultados

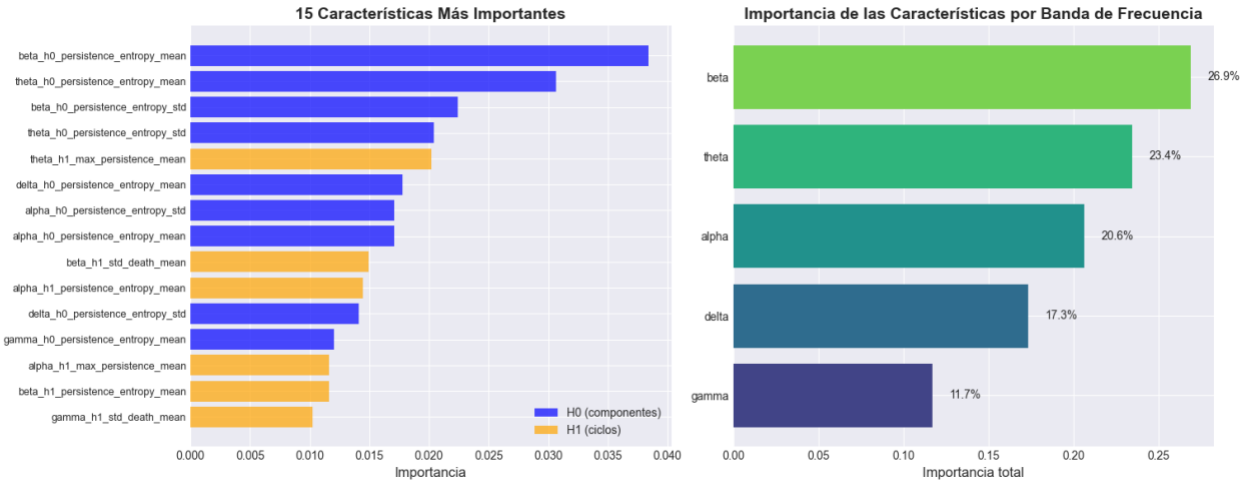
El modelo Random Forest entrenado sobre los 220 descriptores topológicos por ensayo alcanzó una precisión media de 89.3 % ± 1.7 % en validación cruzada GroupKFold a nivel de sujeto, claramente por encima del 50 % esperable por azar. Las precisiones por fold se mantuvieron en un rango estrecho (86.2 %–90.9 %), lo que indica estabilidad del modelo entre distintas particiones de sujetos. El puntaje F1 ponderado fue 0.893 y el área bajo la curva ROC 0.951, reflejando una capacidad discriminativa excelente entre las dos condiciones de audio.



La matriz de confusión obtenida a partir de las predicciones de validación cruzada (Figura X) muestra un rendimiento equilibrado entre clases: se clasificó correctamente el 86 % de las muestras de audio lento y el 92 % de las muestras de audio rápido. La precisión por clase fue aproximadamente 92 % para lento y 87 % para rápido, mientras que el recall fue 86 % y 92 %, respectivamente. Este patrón indica que el clasificador no presenta sesgos sistemáticos hacia una condición particular.

La significancia estadística del rendimiento se evaluó mediante un test de permutación con 1 000 iteraciones. Al permutar aleatoriamente las etiquetas lento/rápido, la distribución nula de precisión se concentró en torno al 50 %, y solo 0.1 % de las permutaciones alcanzó una precisión igual o superior a la observada ($p = 0.001$). El tamaño del efecto, medido con la d de Cohen = 23.30, muestra que la precisión real se sitúa muchas desviaciones estándar por encima del nivel esperado por azar. De forma complementaria, el análisis bootstrap con 1 000 remuestreos a nivel de sujeto produjo un intervalo de confianza del 95 % de [92.1 %, 96.4 %], bien por encima del 50 %.

El análisis de importancia de características indica que las bandas de frecuencia no contribuyen por igual al rendimiento del modelo. Al agrupar la importancia por banda, beta aporta alrededor de 26.9 % de la importancia total, seguida de theta (23.4 %) y alpha (20.6 %); delta y gamma contribuyen 17.3 % y 11.7 %, respectivamente (Figura Y). En conjunto, beta, theta y alpha concentran cerca del 71 % de la información discriminativa. A nivel de invariantes topológicos, las características asociadas a ciclos (H1) muestran una contribución algo mayor que las de componentes conexas (H0), aunque ambas dimensiones participan de manera sustantiva.



En resumen, los resultados cuantitativos y las visualizaciones asociadas (matriz de confusión, distribuciones por sujeto y gráficos de importancia de características) muestran de forma consistente que las características topológicas de los grafos de conectividad EEG permiten distinguir de manera fiable entre audio lento y rápido en infantes, generalizando a sujetos no vistos y con un rendimiento muy superior al esperado por azar.

Discusión

El estudio demuestra que la topología de los grafos de conectividad EEG captura diferencias sistemáticas entre condiciones de audio de diferentes velocidades. La dominancia de la entropía de persistencia como característica discriminativa sugiere que no son las magnitudes absolutas de conectividad las que difieren, sino la distribución y diversidad de escalas de conectividad. Audio lento versus rápido podría inducir patrones de sincronización más o menos jerárquicos, reflejados en distribuciones de persistencia distintas.

La relevancia de los ciclos (H1) es particularmente interesante. En neurociencia de redes, los ciclos pueden representar bucles de retroalimentación donde la información circula entre regiones antes de integrarse. Que estos sean más sensibles que las componentes conectadas (H0) sugiere que las diferencias no están tanto en qué regiones están conectadas, sino en cómo están conectadas y qué rutas de procesamiento se utilizan.

Sin embargo el proyecto presenta varias limitaciones. Primero, no se realizaron comparaciones directas con métodos tradicionales como características espectrales o métricas estándar de teoría de grafos. Esta es una limitación importante y n podemos afirmar que TDA es superior a otros enfoques, solo que es efectivo. Sería valioso en trabajo futuro comparar TDA contra otras metodologias, pero aun asi TDA ofrece ventajas conceptuales. Es invariante a transformaciones continuas (robusto a ruido) y multi-escala . Estas propiedades lo hacen atractivo para análisis de redes cerebrales complejas donde la organización jerárquica y multi-escala es importante.

Segundo, no controlamos múltiples variables que podrían influir en conectividad EEG: edad específica, hora del día, estado de alerta, orden de presentación de estímulos. Aunque la validación cruzada a nivel de sujeto debería distribuir estos efectos equitativamente, un diseño experimental más controlado permitiría conclusiones más definitivas.

Ademas los parámetros TDA (dimensión máxima H1, longitud de arista 2.0) se seleccionaron según práctica estándar, pero no se exploró sistemáticamente el espacio de parámetros. Es posible que otras configuraciones capturen información adicional o diferente. Sin embargo, búsqueda exhaustiva de parámetros conlleva riesgo de sobreajuste.

Hallazgos Inesperados

La baja contribución de la banda gamma (11.7%) fue sorprendente. Gamma (30-100 Hz) se asocia típicamente con procesamiento cognitivo de alto nivel y binding de información. Esperaríamos que diferencias en velocidad de audio, que afectan carga de procesamiento temporal, se reflejen en gamma. Su baja importancia sugiere que las diferencias críticas ocurren en niveles más básicos de procesamiento auditivo (theta, beta) más que en integración cognitiva de alto nivel.

Los resultados establecen claramente que la velocidad del audio escuchado en infantes efectivamete induce patrones de conectividad cerebral topológicamente distintos en infantes. Esto puede tener implicaciones para la compresion sobre como funciona procesamiento auditivo.

Conclusión

Este estudio demostró que las características topológicas de grafos de conectividad EEG distinguen exitosamente entre condiciones de audio lento y rápido en infantes (89.3% de precisión, $p < 0.001$). La pregunta científica planteada se responde afirmativamente: diferentes velocidades de audio inducen patrones de conectividad cerebral topológicamente distintos que son detectables, robustos y generalizables entre sujetos.

La principal contribución del Análisis Topológico de Datos fue proporcionar características cuantitativas multi-escala de la organización funcional de redes cerebrales. La entropía de persistencia y las propiedades de ciclos (H1) capturaron diferencias que reflejan reorganización de circuitos de conectividad según la tasa temporal del estímulo. Las bandas beta, theta y alpha fueron las más discriminativas, consistente con su rol en procesamiento auditivo.

Las limitaciones principales incluyen ausencia de comparación con métodos tradicionales, tamaño de muestra moderado (45 sujetos), y falta grid search o comparacion entre modelos distintos al random forest. Extensiones futuras valiosas incluirían abordar tales limitaciones.

Declaración de reproducibilidad

El análisis se realizó con las siguientes versiones:

- Python: 3.14
- NumPy: (para operaciones numéricas y arrays)
- pandas: (para manipulación de datos tabulares)
- matplotlib: (para visualizaciones)
- seaborn: (para gráficos estadísticos)
- scikit-learn: (para machine learning: RandomForestClassifier, StandardScaler, validación cruzada, métricas)
- ripser: (para cálculo de homología persistente)
- tqdm: (para barras de progreso)

Todas las librerías están disponibles mediante pip o conda. El código fue desarrollado y ejecutado en macOS, pero no deberian haber problemas de compatibilidad. (Aún no lo subo a github, pero lo subire a la brevedad a https://github.com/Ignaciagothe)

Los parámetros clave establecidos para reproducibilidad:

- Semilla aleatoria: 42 (fijada en `np.random.seed(42)`)
- Validación cruzada: GroupKFold con k=5
- Random Forest: 100 árboles, profundidad máxima 10
- TDA: dimensión máxima H1, longitud de arista máxima 2.0
- Permutaciones: 1,000 iteraciones
- Bootstrap: 1,000 iteraciones