

**Representación cerebral de atributos del habla durante el diálogo natural**

Juan Octavio Castro (0009-0005-6398-9236, joctavio287@gmail.com)<sup>1,2</sup>,  
Joaquin E Gonzalez (0009-0009-0473-6545,  
joaquin.gonzalez6693@gmail.com)<sup>1,2</sup>, Jazmin Vidal Dominguez  
(0009-0008-1306-7761, jazmin.vidald@gmail.com)<sup>1</sup>, Pablo E Riera  
(0000-0002-0635-9917, pablo.riera@gmail.com)<sup>1,3</sup>, Agustin Gravano  
(0000-0003-2812-6361, agravano@utdt.edu)<sup>4,5</sup>, y Juan E Kamienkowski  
(0000-0002-5725-6539, juank@dc.uba.ar)<sup>1,3,6</sup>

<sup>1</sup> *Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires - CONICET, Argentina;*

<sup>2</sup> *Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina;*

<sup>3</sup> *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina;*

<sup>4</sup> *Laboratorio de Inteligencia Artificial; Escuela de Negocios, Universidad Torcuato Di Tella, Argentina;*

<sup>5</sup> *CONICET, Argentina;*

<sup>6</sup> *Maestría de Explotación de Datos y Descubrimiento del Conocimiento, FCEyN-FI, UBA, Argentina*

**Resumen:** El estudio del habla en entornos naturales presenta desafíos para los enfoques tradicionales de análisis con electroencefalograma (EEG). En los últimos años, los modelos de codificación (*encoding*) y aprendizaje automático han avanzado considerablemente, facilitando una transición a diseños experimentales que contemplan estímulos dinámicos naturales, como el habla. En este trabajo se busca comprender cómo se codifican en el cerebro los distintos atributos del habla en el marco de un diálogo natural no guionado. Para ello se parte de características de bajo nivel (envolvente, frecuencia fundamental, espectrograma) y, luego, se utilizan atributos de alto nivel como los fonemas y las características fonológicas. Los resultados muestran que la inclusión de las características de alto nivel mejoran la predicción de la señal cerebral a partir del habla para todas las bandas de frecuencia consideradas. Además, las predicciones hechas sobre fonemas y características fonológicas indican que la sensibilidad neuronal es compatible con la hipótesis de un sistema de procesamiento jerárquico del lenguaje.

**Palabras clave:** procesamiento del habla, neurociencia cognitiva, modelos de codificación, EEG.

### **Neural representation of speech features during natural dialogue**

**Abstract:** Studying speech in natural environments presents significant challenges for traditional electroencephalography (EEG) analysis approaches. In recent years, encoding models and machine learning techniques have made substantial progress, enabling a shift toward experimental designs that incorporate naturalistic, dynamic stimuli such as speech. This study aims to understand how different speech attributes are encoded in the brain during unscripted natural dialogue. We begin by analyzing low-level features (envelope, fundamental frequency, spectrogram) and then incorporate higher-level features, such as phonemes and phonological attributes. The results show that including high-level features improves the prediction of neural responses from speech across all frequency bands. Moreover, predictions based on phonemes and phonological features suggest that neural sensitivity is consistent with a hierarchical language processing system.

**Keywords:** speech processing, cognitive neuroscience, encoding models, EEG.

### **1. Introducción**

En los últimos años, en el campo del procesamiento cognitivo y cerebral del habla hubo una transición desde diseños más experimentales (frases cortas, palabras e incluso fonemas aislados) a estudios cada vez más naturales (audiolibros, trailers de películas, etc.) (Hamilton y Huth, 2020). Una forma frecuente de estudiar la representación cerebral de estos estímulos continuos y complejos en la actividad cerebral es a través de modelos de *encoding*. En este enfoque, se utilizan atributos específicos de la señal del habla (envolvente, espectrograma, ocurrencia de fonemas, entre otros) como entrada para predecir la señal de EEG, bajo el supuesto de que las posibles no linealidades del sistema auditivo son capturadas por dichos atributos.

Generalmente, se utilizan modelos lineales —como la regresión ridge—, cuya ventaja principal frente a otros que podrían mostrar un desempeño mejor, es su interpretabilidad. En este sentido, los pesos ajustados por el modelo actúan como un filtro lineal característico del sistema neuronal (función de respuesta temporal o TRF, por sus siglas en inglés), revelando tanto la amplitud y la latencia en que cada atributo modula la actividad cerebral (Crosse et al., 2016).

Esta metodología se ha aplicado al estudio de la percepción del habla, enfocándose en distintos atributos como el espectrograma, los fonemas y las características fonológicas. Las representaciones obtenidas son robustas para la predicción del EEG en frecuencias de hasta 8 Hz, en línea con el fenómeno de sincronización cortical, donde la actividad cerebral se acopla con la prosodia (Di Liberto et al., 2015). Estos atributos se estudiaron también en modelos multivariados para analizar la contribución individual de diferentes atributos (Desai et al., 2021). Recientemente, se han extendido los resultados para ciertos atributos (envolvente y espectrograma) a situaciones de diálogos naturales no guionados, donde la

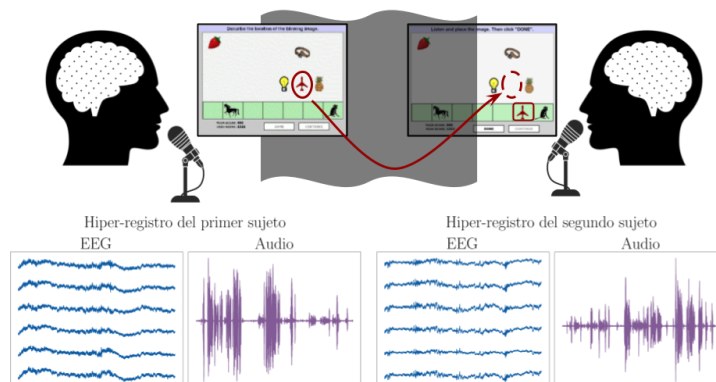
escucha se vuelve activa (Gonzalez et al., 2024), mostrando inclusive una mejora en el desempeño.

En el presente trabajo nos proponemos expandir la exploración en la situación de diálogos no guionados a nuevos atributos del habla, comparando las representaciones de atributos de bajo nivel como la envolvente (Env.) y la frecuencia fundamental ( $F_0$ ) del hablante, con atributos de mayor nivel como los fonemas y las características fonológicas.

## 2. Métodos experimentales

### 2.1. Datos

Se utilizaron datos de registro simultáneo de EEG y habla obtenidos durante un diálogo no guionado, mientras ambos participantes realizaban una tarea diseñada para generar intercambios (Fig. 1). Las sesiones incluyeron entre 15 y 30 ensayos de 1 a 5 minutos cada uno (total=5,3 horas). La adquisición y preprocesamiento de los datos de habla y la sincronización del EEG se describen, respectivamente, en Gravano et al. (2023) y Gonzalez et al. (2024).



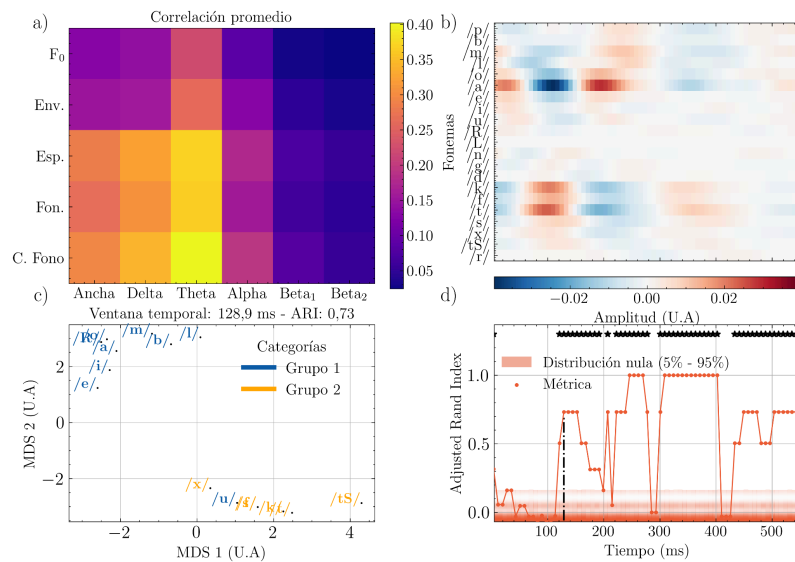
**Fig. 1.** Esquema de la tarea y del registro. Los sujetos ( $N=18$ , 9 mujeres y 9 hombres, edad= $26\pm 6$  años) se encontraban frente a una pantalla, separados por una cortina opaca que no permitía la comunicación visual. La tarea consistía en que el participante de la derecha lograra ubicar el objeto faltante en la misma posición que lo veía el participante 1, en base a sus indicaciones.

Para el análisis, la señal de EEG fue filtrada en diferentes bandas de frecuencia. Entre los atributos novedosos más relevantes se destacan los fonemas (Fon.) y las características fonológicas (C. Fono.) que fueron identificados automáticamente con Phonet (Vásquez-Correa et al., 2019), una implementación de Python que utiliza un banco de redes neuronales diseñadas para calcular en forma simultánea e independiente las probabilidades condicionales de que una dada ventana temporal pertenezca

a un determinado fonema y clase fonológica. Para ajustar las TRFs de cada atributo se construyó una regresión lineal múltiple con regularización *Ridge* por participante. Los parámetros fueron estimados utilizando validación cruzada en 5 particiones, mientras que el hiperparámetro de regularización fue optimizado con el fin de maximizar la correlación en una grilla de 32 valores equiespaciados dentro de una escala logarítmica. El detalle del esquema de análisis y preprocesado de los atributos puede encontrarse en Gonzalez et al. (2024). En el presente trabajo se realizaron optimizaciones sobre dicho esquema utilizando las librerías PyTorch (Paszke et al., 2019) y MNE (Gramfort et al., 2013).

### 3. Resultados

Para evaluar el impacto y el desempeño del modelo en distintas bandas de frecuencia del EEG, para cada electrodo y sujeto, se calculó la correlación de Pearson entre la señal original y aquella predicha por cada atributo de la señal de habla del interlocutor. Al promediar los resultados en sujetos y electrodos (Fig. 2a), se observa un desempeño superior en la banda Theta, en línea con la hipótesis de sincronización cortical en baja frecuencia ( $\leq 8$  Hz). Dentro de esta banda, el desempeño mejora a medida que aumenta la complejidad de los atributos. Incorporar atributos de segundo nivel (fonemas y características fonéticas) aumenta el poder predictivo en todas las bandas respecto a los atributos previamente estudiados (Gonzalez et al., 2024), y además, alcanza correlaciones superiores a trabajos previos sin tareas activas (Di Liberto et al., 2015).



Alpha (8-13 Hz), Beta<sub>1</sub> (13-19 Hz) y Beta<sub>2</sub> (19-25 Hz). **b)** TRF promedio por canal y sujeto para cada fonema. **c)** Proyección de las respuestas en un espacio de dimensión reducida (MDS) para una ventana centrada en 128,9 ms **d)** Coincidencia de los agrupamientos de las TRFs (manual y automático) en función del tiempo.

Las TRFs, promediadas entre canales y sujetos, se separan cualitativamente acorde a la forma funcional en dos grupos (Fig. 2b): el primero —constituido por los fonemas /m/, /l/, /b/, y las vocales— presenta la respuesta característica de la envolvente para este tipo de tareas, mientras que el segundo —/f/, /k/, /t/, /s/, /x/ y /tʃ/— muestra una respuesta invertida en signo. Con el fin de comprender si estos resultados están en concordancia con la hipótesis de que la segmentación semántica de los fonemas ocurre en forma jerárquica, se subdividieron las TRFs en ventanas de tiempo móviles y superpuestas de 93,75 ms. Para cada ventana, la TRF de cada fonema fue proyectada en un espacio euclidiano cuya distancia se definió en base a la disimilaridad entre los posibles pares de fonemas (MDS) (Fig. 2c), donde aquellas respuestas con correlaciones fuertes se encuentran más cerca.

Se agruparon las respuestas en cada ventana utilizando el algoritmo k-medias (k=2; Pedregosa et al., 2011). Luego, se contrastaron los grupos con el etiquetado manual mencionado previamente, obteniendo valores de ARI (Adjusted Rand Index) significativos a partir de los ~130 ms. Se observó que la segmentación crece conforme aumenta la latencia de la respuesta (Fig. 2d), consistentemente con la hipótesis planteada.

#### 4. Conclusiones

Brevemente, se muestra que la incorporación de atributos complejos, como fonemas y características fonológicas, en diseños más naturales donde la escucha se torna activa, mejora considerablemente el desempeño de los modelos de *encoding* para la predicción de EEG. Más aún, la segmentación semántica de los fonemas es compatible con un sistema jerárquico inclusive en diálogos no guionados. A futuro, se pretende estudiar la respuesta de cada grupo por separado con el objetivo de verificar si éstos se separan en subgrupos acorde a otras características, a latencia mayores de los 130 ms. Asimismo, se espera replicar el análisis para las características fonológicas, con el objetivo de estudiar la jerarquía entre las clasificaciones semánticas en el cerebro.

#### Bibliografía

- Crosse, M., Di Liberto, G., Bednar, A., & Lalor, E. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10, 604.
- Desai, M., Holder, J., Villarreal, C., Clark, N., Hoang, B., & Hamilton, L. (2021). Generalizable EEG encoding models with naturalistic audiovisual stimuli. *Journal of Neuroscience*, 41(43), 8946-8962.

- Di Liberto, G., O'Sullivan, J., & Lalor, E. (2015). Low-frequency cortical entrainment to speech reflects phoneme level processing. *Current Biology*, 25(19), 2457-2465.
- Gravano, A., Kamienkowski, J., & Brusco, P. (2023). UBA Games Corpus. <http://hdl.handle.net/11336/191235>.
- Gonzalez, J., Nieto, N., Brusco, P., Gravano, A., & Kamienkowski, J. (2024). Speech-induced suppression during natural dialogues. *Communications Biology*, 7(1), 291.
- Gramfort, A. et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7, 267
- Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5), 573-582.
- Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12, 2825-2830.
- Vásquez-Correa, J., Klumpp, P., Orozco-Arroyave, J., & Nöth, E. (2019). Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. *INTERSPEECH*, 60, 61.