ISSN 1870-4069

Estudios sobre el reconocimiento del habla imaginada mediante señales EEG

Arturo Iván Sandoval Rodríguez, Francisco Hiram Calvo Castro

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Laboratorio de Ciencias Cognitivas Computacionales,
México

asandovalr2021@cic.ipn.mx, fcalvo@ipn.mx

Resumen. El habla imaginada es un área con una gran cantidad de posibilidades para mejorar la calidad de vida de las personas que han sufrido de accidentes que les impidan comunicarse mediante el habla o que no puedan realizar actividad motriz. Es por eso que este tema de estudio resulta interesante, mediante el estudio del cerebro, más específicamente, señales electroencefalográficas (EGG) posibilitar una nueva forma de comunicación mediante una interfaz cerebro-computadora. En el presente trabajo se realiza un estudio del estado del arte, así como una comparativa de trabajos recientes. Finalmente se concluye con una propuesta de que es lo que se puede innovar teniendo en cuenta el estudio del estado del arte realizado.

Palabras clave: Habla imaginada, señales EEG, interfaz cerebro-computadora, comunicación asistida, estado del arte.

Studies on Imagined Speech Recognition Using EEG Signals

Abstract. Imagined speech is a field with great potential to improve the quality of life for individuals who have suffered accidents that prevent them from communicating through speech or performing motor activities. For this reason, the topic is of particular interest, as it explores the brain — more specifically, electroencephalographic (EEG) signals — to enable a new form of communication through a brain–computer interface (BCI). This work presents a study of the state of the art in imagined speech recognition, along with a comparison of recent research. Finally, it concludes with a proposal on potential innovations based on the insights gained from the state-of-the-art analysis.

Keywords: Imagined speech, EEG signals, brain–computer interface, assistive communication, state of the art.

1. Introducción

Para comenzar es necesario dar una definición de lo que es el habla imaginada, en palabras de D'zmura et al., (2009) corresponde a imaginar la pronunciación del habla en ausencia de salida articulatoria y acústica. Ahora bien, dentro del mundo de las ciencias cognitivas, el cerebro es el principal objeto de estudio, aunque debido a la complejidad de su entendimiento, a lo largo de los años se han ido desarrollado herramientas para tener un mejor entendimiento de este. Una de las más utilizadas son las Interfaces Cerebro-Computadora o BCI (por sus siglas en inglés, Brain Computer Interface), que en palabras de Valerdi et al., (2019) son sistemas que miden la actividad del Sistema Nervioso Central y la convierten en salidas que reemplazan, restauran, aumentan, suplementan o mejoran las salidas naturales de dicho sistema y, por lo tanto, cambian las interacciones en curso entre el ser humano y su ambiente interno o externo.

En otros estudios se define como un sistema informático que adquiere señales cerebrales, las analiza y las traduce en comandos que se transmiten a un dispositivo de salida para llevar a cabo una acción deseada [2].

Estas interfaces o bien sistemas están basadas en señales electroencefalográficas (EEG), las cuales resultan ser una excelente opción de estudio pues no son invasivas, son relativamente simples, económicas e insensibles a ambientes con grandes cantidades de ruido audible [3].

El EEG estándar es una exploración indolora, no invasiva, de bajo coste, que puede ser de gran utilidad en la práctica clínica. Se realiza colocando electrodos de superficie adheridos al cuero cabelludo por un gel conductor. Se posicionan de acuerdo con el sistema internacional 10-20 [4, 5].

Cada canal de registro mide la diferencia de voltaje entre dos electrodos (uno es el activo y otro el de referencia). Lo habitual es que se usen de 16 a 24 derivaciones en cada montaje. Los distintos pares de electrodos se combinan constituyendo los montajes. Hay dos tipos básicos de montajes: bipolar (transversal y longitudinal) y monopolar (o referencial). [7] El bipolar registra la diferencia de voltaje entre dos electrodos colocados en áreas de actividad cerebral, mientras que el monopolar registra la diferencia de potencial entre un electrodo ubicado en una zona cerebral activa y otro colocado sobre un área sin actividad o neutra [7].

2. Estado del arte

Los trabajos que se encontraron donde se hace hincapié en el habla imaginada se detallan a continuación:

2.1. Decoding imagined, heard, and spoken speech: classification and regression of EEG using a 14 channel dry-contact mobile headset

(Clayton et al., 2020) [8]: En este trabajo los autores hacen énfasis en que los dispositivos de tipo científico de recolección de datos EEG resultan ineficientes

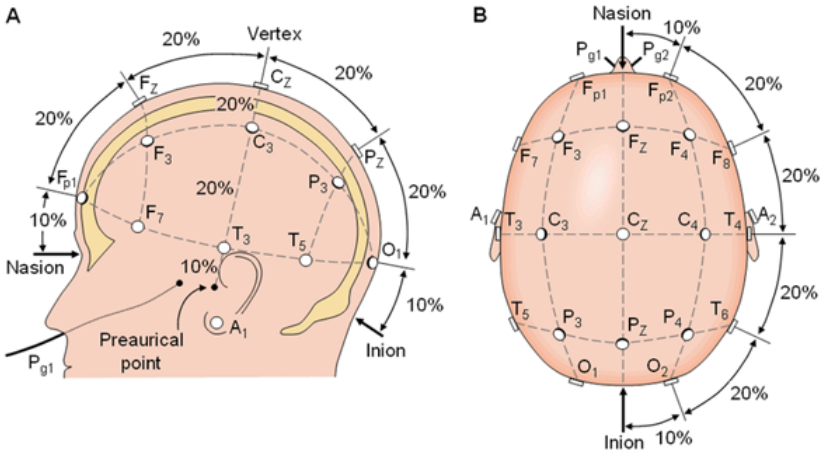

Fig. 1. Sistema Internacional 10-20 para la colocación de los electrodos extracraneales. Las letras señalan el área (Fp, prefrontal; F, frontal; C, central; P, parietal; T, temporal y O, occipital), mientras que los números designan el hemisferio (pares del derecho, nones del izquierdo) y los electrodos de la línea media se señalan con una "z "; por lo que Fz se encuentra frontalmente en la línea media [6].

en los escenarios cotidianos, por esa razón optaron por usar un dispositivo de tipo comercial que, si bien ofrece una menor fidelidad con menos electrodos, resulta ser más práctico para el día a día.

Sus objetivos primarios son:

- Evaluar la efectividad del dispositivo "ligero"para la decodificación del habla, al comparar el rendimiento con el de dispositivos de grado investigativo.
- Investigar diferentes modelos de machine learning para la clasificación del habla y tareas de regresión con EEG.
- Presentar un nuevo corpus o bien base de datos, con licencia de grabaciones de EGG evocadas por voz además de resultados y código.

La base de datos presentada fue llamada FEIS (por sus siglas en ingles Fourteen-channel EEG for Imagined Speech) y contiene grabaciones de 21 participantes de habla inglesa, estas grabaciones se realizaron con un dispositivo portátil, de 14 canales con electrodos secos (el Emotiv EPOC+) [9].

Las grabaciones EGG constan en 160 ensayos, los cuales comprenden 6 fonemas en 10 repeticiones, aleatorias para mantener la atención del participante. Cada prueba tiene cuatro "épocas"de cinco segundos. Este procedimiento se ilustra en la figura 2.

Primero se realiza una época de "descanso", en la cual a los participantes se les muestra en el monitor la palabra "REST" y tratan de despejar su mente, esto también reduce la carga cognitiva. Después viene una época de "estímulo" en la cual a los participantes se les reproduce su propia grabación del fonema

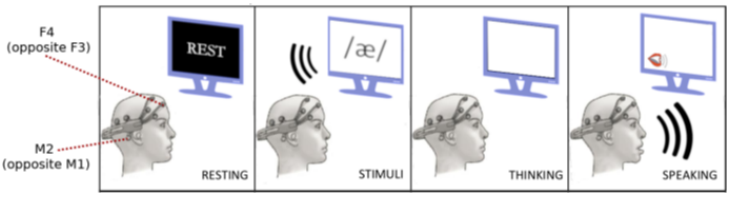

Fig. 2. Los participantes escuchan cinco veces repetidas un fonema (el cual es su propia voz grabada previamente), luego se imaginan hablando este fonema cinco veces (con el mismo ritmo), y después hablan abiertamente el fonema cinco veces [8].

Tabla 1. Resultados de los modelos empleados.

Porcentaje de precisión (Desviación estandar)			
Modelo	Escuchar	Pensar	Hablar
SVM	64.0(16.5)	69.0(13.2)	63.7(21.2)
CNN	51.2(5.7)	49.0(6.1)	49.4(6.3)

en bucle cinco veces, y se les muestra en el monitor una representación IPA del fonema. Lo siguiente es una época de "pensamiento", la cual consiste en mostrar a los participantes el monitor en pantalla negra y que se imaginen repitiendo el fonema, pero sin realizar ningún movimiento. Finalmente, la época del "habla", donde a los participantes se les muestra una imagen de una boca para que posteriormente vocalicen el fonema.

Para el modelado se realizó una división de los datos del 80/10/10, para el entrenamiento, prueba y validación respectivamente. Para la máquina de vectores de Soporte (por sus siglas en inglés SVM, Support Vector Machine) se usó un 80/20, entrenamiento/prueba con validación cruzada de cinco iteraciones. Para cada tipo de habla, escuchada, imaginada y hablada, los datos fueron divididos en épocas de cinco segundos. Las características son extraídas de estas divisiones y son usadas para el entrenamiento del clasificador usando las etiquetas correspondientes al tipo de fonema, y un modelo de regresión usando las características extraídas del estímulo de audio que fue escuchado, imaginado y hablado. Para la clasificación se probaron dos tipos de modelo, máquina de vectores de soporte(SVM) y las redes neuronales convolucionales (por sus siglas en inglés, Convolutional Neural Networks).

Los resultados más destacables de este trabajo se muestran en la tabla 1. En esta se puede observar que el modelo SVM tuvo mejor rendimiento, sin embargo, los mismos autores mencionan que es posible obtener mejores resultados con las CNN aunque esto requeriria de mayor investigación.

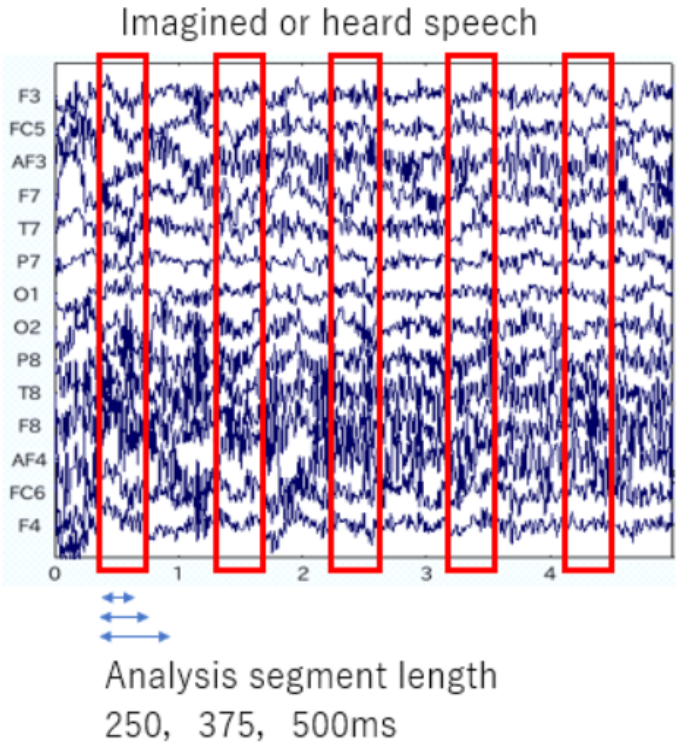

Fig. 3. Extracción de caracteristicas de señal EGG.

2.2. Classification of Imagined and Heard Speech Using Amplitude Spectrum and Relative Phase of EEG

(Sakai et al., 2021) [10]: En esta investigación precisamente se retoma el trabajo de Clayton et al., pues principalmente para su estudio se hace uso de la base de datos FEIS, sin embargo, este trabajo se enfoca en la amplitud del espectro y la fase relativa de las señales EGG. Para aumentar el número de muestras, dividen la información en cinco partes (de un segundo cada uno) y es tratada cada parte como información separada. Con el fin de investigar en qué periodo de tiempo las características aparecen, los experimentos fueron llevados a cabo con tres longitudes de segmento de análisis de 250, 375 y 500 ms desde el inicio de cada estímulo o pensamiento repetido, esto es ilustrado en la figura 3.

Como clasificador, se usó una máquina de vectores de soporte (SVM) y redes neuronales (NN). Por una parte, la SVM realiza una validación cruzada de cinco iteraciones en la información preprocesada de cada participante con un kernel gaussiano. Por otro lado, la red neuronal usada tiene una capa oculta con el mismo número de nodos que de dimensiones de entrada. Se aplicó ReLU como función de activación, la función sigmoide se aplicó en la capa final, entropía cruzada como la función de perdida, y Adam como función de optimización. Se

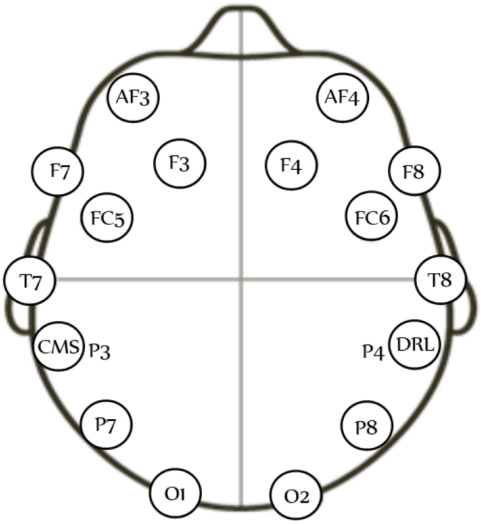

Fig. 4. Localización de los electrodos en el kit Emotiv [11].

dividió la información de la siguiente manera, 80 % para entrenamiento, 10 % para validación y 10 % para pruebas. Finalmente se obtuvo una mejor precisión usando tanto el espectro de amplitud como la fase relativa, lo que indica la efectividad de ambas en el reconocimiento del habla imaginada. Respecto a los modelos, las redes neuronales presentan una mejor clasificación, por otro lado, los experimentos con SVM no mejoraron los resultados de trabajos previos.

2.3. Análisis de señales electroencefalográficas para la clasificación de habla imaginada

(Torres et al., 2012)[3] Este trabajo se enfoca en la identificación de cinco palabras a través del habla imaginada. Para la metodología que se presenta está dividida en cinco etapas, las cuales son: Adquisición de la actividad cerebral, Preprocesamiento, Extracción de características, Selección de características, y Clasificación. Para la parte de adquisición de datos se usó el kit EMOTIV, el cual tiene catorce canales de alta resolución. Para la adquisición de datos se tiene un mayor interés a los canales F7, FC5, T7, P7(ver fig. 4) puesto a que están más cercanas a las áreas del modelo Geschwind-Wernicke, lugar donde las señales EGG están más relacionadas a la producción del habla.

Además, se usó un pequeño protocolo para delimitar el inicio y fin de una palabra imaginada. Un conjunto de muestras entre ambos etiquetados (inicio y fin) es llamado ventana o instancia. El preprocesamiento consiste en filtrar la información obtenida de los canales previamente mencionados mediante un filtro FIR pasa bandas en el rango de 4 a 25 Hz. Cabe mencionar que la duración de cada ventana del habla imaginada es variable, tanto para un sujeto como para diferentes sujetos, por lo que resulta innecesario establecer un tamaño

específico para todas las ventanas. En este punto, las ventanas con 256 muestras y una frecuencia entre 4 a 25 Hz se mantienen para el uso de creación de la información experimental. Las ventanas con un menor número de muestras a 256 son completadas con ceros y aquellas que sobrepasen las 256 muestras son descartadas.

Para la extracción de características se usa la Transformada Wavelet Discreta, su ecuación está definida por [12]:

$$W(j,k) = \sum_j \sum_k f(x) 2^{-j/2} \psi(2^{-j}x - k). \quad (1)$$

La DWT (por sus siglas en inglés, Discrete Wavelet Transform) provee una representación wavelet altamente eficiente mediante la restricción de la variación en la traslación y la escala, usualmente a potencias de dos. Por lo que se aplica la DWT con seis niveles de descomposición, usando como wavelet madre a una Daubechies de segundo orden(db2). Se obtiene un vector con 269 coeficientes wavelet por cada ventada en cada canal de interés. Por lo tanto, los coeficientes en el mismo intervalo de tiempo que pertenecen a los cuatro canales son concatenados en el siguiente orden F7-FC5-T7-P7. En este punto se obtiene un vector con 1076 características y su etiqueta de clase es obtenida.

Para la selección de características, los subconjuntos que sean mayores a 25 Hz son descartados, después el subconjunto seleccionado consiste en los coeficientes D2 al D6 y a la aproximación A6(estos coeficientes obtenidos previamente de la DWT) lo cual reduce el tamaño del vector de características y al mismo tiempo reduce el impacto del problema de la dimensionalidad. Con esto, cada ventana de cada canal es representado con 140 coeficientes wavelet. Después, los coeficientes DWT de ventanas en el mismo intervalo de tiempo son concatenadas como parte de la etapa de extracción de características.

Finalmente tres tipos de clasificadores fueron entrenados y probados: Maquinas de vectores de soporte (SVM), Random Forest (RF) y Naive Bayes(NB). La precisión de la clasificación es obtenida a través de una validación cruzada de 10 iteraciones, los resultados preliminares(realizados a los datos de tres participantes) se muestran a en la tabla 2.

Tabla 2. Porcentaje de precisión de cada uno de los tres clasificadores utilizando los vectores de características reducido.

Participante	NB	RF	SVM
S1	23.35	24.08	23.35
S2	17.09	31.63	24.78
S3	35.75	41.21	18.18

Con esto se procedió a usar a los clasificadores para entrenamiento y prueba en los 21 participantes, y como en los resultados preliminares se obtuvo mayor

precisión con RF, se experimentó con está teniendo aún mejores resultados bajo el nombre de "Bagging-RF".

3. Conclusiones y trabajo futuro

A pesar de que hay trabajos previos sobre el habla imaginada, esta sigue siendo un mundo poco explorado, por lo que siguen existiendo muchísimas posibilidades por crear e innovar. Cabe mencionar que los trabajos previos tienen su alta relevancia y podrían funcionar como punto de inicio para futuras investigaciones. La selección de los trabajos para esta investigación se debe a que son estudios recientes y que tienen su importancia en este campo, la identificación del habla imaginada, además se cuenta con su base de datos por lo que se plantea en primera instancia replicar su trabajo para posteriormente proponer un mayor desarrollo. Con esta premisa se propone realizar un estudio donde con un conjunto de silabas, separadas por pocos segundos, se empiecen a identificar palabras sencillas, claro esta con su respectiva limitación a cierto número de sílabas y las posibles combinaciones que resulten para las palabras a identificar.

Agradecimientos. Este trabajo fue apoyado por el CONACyT, COFAA y el Instituto Politécnico Nacional, IPN-EDI, IPN-SIP gracias a los proyectos SIP 2083 y 20210189.

Referencias

1. D'Zmura, M., Deng, S., Lappas, T., Thorpe, S., Srinivasan, R.: Toward EEG sensing of imagined speech. In: Human-Computer Interaction. New Trends, ed J. A. Jacko (Berlin; Heidelberg: Springer Berlin Heidelberg), 40–48 (2009)
2. Alonso-Valerdi, L. M., Arreola-Villarruel, M. A., Argüello-Garcia, J.: Interfaces Cerebro-Computadora: Conceptualización, Retos de Rediseño e Impacto Social. Mexican Journal of Biomedical Engineering, 40(3), 1–18 (2019)
3. Jerry J. Shih, Dean J. Krusienski, Jonathan R. Wolpaw: Brain-Computer Interfaces in Medicine. Mayo Clinic Proceedings (2012)
4. Torres-García, A., Reyes-Garcia, C. A., Villaseñor-Pineda, L.: Toward a silent speech interface based on unspoken speech. In: BIOSIGNALS 2012 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing (2012)
5. Homan, R.W., Herman, J., Purdy, P.: Cerebral location of international 10-20 system electrode placement. Electroencephalogr Clin Neurophysiol, 66, 376–382 (1987)
6. Myslobodsky, M.S., Coppola, R, Bar-Ziv, J., Weinberger, D.R.: Adequacy of the International 10-20 electrode system for computed neurophysiologic topography. J Clin Neurophysiol 7, 507–518 (1990)
7. Novo-Olivas, C,, Guitiérrez, L,, Bribiesca, J,: Mapeo Electroencefalográfico y Neurofeedback. (2010)

8. F. Ramos-Argüelles, G. Morales, S. Egozcue, R.M. Pabón, M.T. Alonso: Técnicas básicas de electroencefalografía: principios y aplicaciones clínicas, An. Sist. Sanit. Navar. 32 (Supl. 3): 69–82 (2009)
9. Clayton, J., Wellington, S., Valentini-Botinhao, C., Watts, O.: Decoding imagined, heard, and spoken speech: Classification and regression of EEG using a 14-channel dry-contact mobile headset. In: Proceedings Interspeech 2020 (Vol. 2020-October, pp. 4886-4890). International Speech Communication Association (2020) https://doi.org/10.21437/Interspeech.2020-2745
10. Emotiv EPOC+. [Online]. Available: https://www.emotiv.com/epoc/
11. R. Sakai, A. Kai. S. Nakagawa: Classification of Imagined and Heard Speech Using Amplitude Spectrum and Relative Phase of EEG. In: IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), pp. 373-375 (2021) doi: 10.1109/LifeTech52111.2021.9391883.
12. Torres-García, A.A., Reyes-García, C.A., Villaseñor-Pineda. L., et al.: Análisis de Señales Electroencefalográficas para la Clasificación de Habla Imaginada. Rev Mex Ing Biomed. 34(1):23–39 (2013)
13. Priestley M.: Wavelets and timedependent spectral analysis. Journal of Time Series Analysis, vol. 17, no. 1, pp. 85-–103 (2008)