

***“MODELO DE RED NEURONAL BASADO EN
CAPSNET PARA EL ANÁLISIS DE HABLA
IMAGINADA”***

TESIS

QUE PARA OBTENER EL GRADO DE

***DOCTOR EN CIENCIAS EN INGENIERÍA
ELECTRÓNICA***

PRESENTA:

M.C. JOSÉ MANUEL MACÍAS MACÍAS

DIRECTOR DE LA TESIS:

DR. JUAN ALBERTO RAMÍREZ QUINTANA

CODIRECTOR DE LA TESIS:

DR. ALEJANDRO ANTONIO TORRES GARCÍA

CHIHUAHUA, CHIH., NOVIEMBRE DE 2023



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO®





Chihuahua, Chih., 31 de octubre de 2023

**JOSÉ MANUEL MACÍAS MACÍAS
PRESENTE**

En cumplimiento con los requerimientos para la obtención del grado de Doctor en Ciencias en Ingeniería Electrónica y a propuesta de su Comité Tutorial, la División de Estudios de Posgrado e Investigación le concede la autorización para imprimir la tesis titulada **"Modelo de red neuronal basado en CapsNet para el análisis de habla imaginada"**, dirigida por el Dr. Juan Alberto Ramírez Quintana con el siguiente contenido de capítulos:

- I Antecedentes
- II Estudio de vocales imaginadas
- III Propuesta de modelo basado en CapsNet
- IV Resultados de reconocimiento con CapsK-SI
- V Comparación de resultados utilizando CapsK-SI
- VI Conclusiones

ATENTAMENTE

*Excelencia en Educación Tecnológica®
"La técnica por el Engrandecimiento de México"*

**LEONCIO OCHOA CERVANTES
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN**





Chihuahua, Chih., 30 de octubre de 2023

LEONCIO OCHOA CERVANTES
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
PRESENTE

En cumplimiento con los requerimientos para la obtención del grado de Doctor en Ciencias en Ingeniería Electrónica, le notificamos que el documento de tesis del alumno **JOSÉ MANUEL MACÍAS MACÍAS** titulado **"Modelo de red neuronal basado en CapsNet para el análisis de habla imaginada"**, dirigido por el Dr. Juan Alberto Ramírez Quintana, ha sido aprobado y aceptado para su impresión.

Por lo anterior, proponemos le sea concedida la autorización de impresión correspondiente.

Agradeciendo la atención a la presente, quedamos de usted:

ATENTAMENTE

Excelencia en Educación Tecnológica®
"La técnica por el Engrandecimiento de México"

Juan Ramírez

DR. JUAN ALBERTO RAMÍREZ QUINTANA
DIRECTOR DE TESIS

DR. ALEJANDRO ANTONIO TORRES GARCÍA
MIEMBRO DEL COMITÉ TUTORIAL

DR. MARIO IGNACIO CHACÓN MURGUÍA
MIEMBRO DEL COMITÉ TUTORIAL

DR. LUIS FRANCISCO CORRAL MARTÍNEZ
MIEMBRO DEL COMITÉ TUTORIAL

DR. JESUS ALFONSO MEDRANO HERMOSILLO
MIEMBRO DEL COMITÉ TUTORIAL





CARTA CESIÓN DE DERECHOS

En la ciudad de Chihuahua el día 31 de octubre de 2023, la que suscribe, JOSÉ MANUEL MACÍAS MACÍAS con número de control D19060653, del Doctorado en Ciencias en Ingeniería Electrónica adscrito a la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Chihuahua, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección de el Dr. Juan Alberto Ramírez Quintana y cede los derechos del trabajo titulado **"Modelo de red neuronal basado en CapsNet para el análisis de habla imaginada"**, al Tecnológico Nacional de México y/o Instituto Tecnológico de Chihuahua para su difusión, divulgación, transmisión, reproducción, así como su digitalización con fines académicos y de investigación.

JOSÉ MANUEL MACÍAS MACÍAS





DECLARACIÓN DE ORIGINALIDAD

En la ciudad de Chihuahua el día 31 de octubre de 2023, la que suscribe, JOSÉ MANUEL MACÍAS MACÍAS con número de control D19060653, del Doctorado en Ciencias en Ingeniería Electrónica adscrito a la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Chihuahua, manifiesta ser autor intelectual del presente trabajo de Tesis titulado **"Modelo de red neuronal basado en CapsNet para el análisis de habla imaginada"**, bajo la dirección de el Dr. Juan Alberto Ramírez Quintana; y manifiesta que el contenido es original y que las fuentes de información consultadas para su fundamentación están debidamente citadas y referenciadas.

JOSÉ MANUEL MACÍAS MACÍAS



Chihuahua, Chihuahua, noviembre de 2023

Dra. María Elena Álvarez-Buylla Roces
Directora de Conacyt

Sirva la presente para saludarla e informarle que a la fecha he obtenido el grado de **Doctor en Ciencias en Ingeniería Electrónica** en la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Chihuahua, motivo por el cual agradezco todo el apoyo brindado por este Consejo que usted representa, con el otorgamiento de una beca que permitió dedicarme de manera exclusiva a la realización de mis estudios de posgrado y de esta manera lograr el objetivo principal del convenio establecido.

Sin otro particular quedo a sus órdenes no sin antes reiterar mi agradecimiento.
Atentamente

A handwritten signature in black ink, consisting of stylized, overlapping loops and lines, positioned above the printed name.

José Manuel Macías Macías

Si la vida fuera perfecta no tendría chiste.

Alejandro Macías Macías

En tres palabras puedo resumir todo lo que he aprendido de la vida: La vida continúa.

Robert Frost

Cuanto más alto coloque el hombre su meta, tanto más crecerá.

Friedrich von Schiller

A mis padres y hermanos, en especial a Alejandro

Porque gracias a su apoyo incondicional y amor, he logrado llegar a ser lo que soy.

Agradecimientos

A Dios

Por colmarme de bendiciones y darme la capacidad necesaria de enfrentar los retos de la vida.

A Alejandro

Lamentablemente no pudiste acompañarme en el término de este proceso, te guardo en mi corazón, sabes que te quiero hermano. Gracias por enseñarme siempre, y más hoy después de tu partida. Pondré tu firma en esta tesis como lo hiciste en la de maestría. Te la dedico. Al rato nos vemos.

A mis padres y hermanos

Porque gracias a su apoyo desde mi nacimiento he logrado alcanzar metas. Gracias por todas las correcciones y consejos que me brindan todos los días. Gracias por apoyarme y siempre estar para mí, y por soportar mi forma de ser.

A Melissa

Porque gracias a su apoyo, cariño y amor he logrado superar situaciones en mi vida, y he podido seguir adelante en cada meta que me propongo. Gracias por ser ese respaldo incondicional todos los días, te amo.

A mis amigos y compañeros

Agradezco a Melissa por ayudarme a alcanzar las metas que me propongo cada día y darme consejos en cada momento. Agradezco a mis compañeros de laboratorio, Eduardo, Miguel, Armando, Edgar, Flavio, Sahyra, Javier por haberme ayudado en el examen predoctoral, artículos, trabajos, programas y demás. También por sus consejos y el ambiente de cordialidad.

A mi comité de tesis

Agradezco al Dr. Juan Ramírez por haberme guiado en este proyecto, haberme ayudado a esclarecer cómo es mejor comunicar las ideas que tengo en mi cabeza, por todas las preguntas

que se me hacían para sustentar en mayor medida el proyecto de tesis. Agradezco al Dr. Alejandro Torres García por sus contribuciones y preguntas acerca de las interfaces cerebro-computadora, modelos matemáticos y demás, por haber ayudado a generar un impacto y guiarme en la investigación de habla imaginada. Agradezco al Dr. Mario Chacón por haberme permitido ocupar un espacio en el laboratorio y por cuestionar mi proyecto de tesis para generar un mayor impacto en la sociedad, por haberme brindado apoyo mediante consejos de cómo se debe de ser doctor. Agradezco al Dr. Francisco Corral por sus consejos y preguntas filosóficas que siempre me ayudaron a ver de más ángulos el trabajo de mi proyecto de tesis. También agradezco al Dr. Alfonso Medrano por haber formado parte del comité de tesis y el contribuir a generar mayor impacto en la comunidad científica.

Al Tecnológico Nacional de México y CONACYT

Agradezco al Tecnológico Nacional de México por el apoyo a este trabajo bajo el proyecto “Diseño de plataforma tecnológica de internet de las cosas para generar conectividad informática en sistemas de procesamiento inteligente de señales enfocados a cuidados médicos y discapacidad”, Clave 14044.22-P.

RESUMEN

MODELO DE RED NEURONAL BASADO EN CAPSNET PARA EL ANÁLISIS DE HABLA IMAGINADA

M.C. José Manuel Macías Macías
Doctorado en Ciencias en Ingeniería Electrónica
División de Estudios de Posgrado e Investigación del
Instituto Tecnológico de Chihuahua
Chihuahua, Chih. 2023
Director de Tesis: Dr. Juan Alberto Ramírez Quintana
Codirector de Tesis: Dr. Alejandro Antonio Torres García

El habla imaginada es la acción de imaginar la pronunciación de palabras, fonemas o vocales sin generar movimientos en los músculos que se involucran para la articulación de lenguaje. En los últimos años se han generado diversos trabajos para el desarrollo de interfaces cerebro-computadora que consisten en el reconocimiento de habla imaginada en señales electroencefalográficas. Sin embargo, la mayoría de estos trabajos no presentan un análisis profundo acerca de los patrones y características que aportan información en el reconocimiento de habla imaginada. Por ello, esta tesis propone un modelo novedoso de análisis de señales EEG para clasificación de las vocales /a/ y /u/, fonemas /iy/, /piy/, /tiy/, /diy/, /uw/, /m/ y /n/, y palabras /pat/, /pot/, /knew/ y /gnaw/, que permite identificar propiedades relevantes relacionadas con el habla imaginada en la clasificación. El modelo se basa en procesamiento de propiedades estadísticas de señales EEG y redes neuronales tipo cápsula. Para el análisis de vocales se utilizó el método de patrones espaciales comunes y una red tipo capsula denominada Cápsulas para Vocales (CapsVI). Para el análisis de fonemas y palabras se utilizó estadística multivariada y ANOVA con el propio modelo CapsVI. A la versión del modelo de análisis de fonemas se le denomina cápsulas para Habla imaginada (CapsK-SI). Los resultados demostraron que el modelo desarrollado logra reconocer con éxito vocales, grupos de fonemas y palabras. Además, este modelo brinda información que puede ser interpretable y relacionada con las teorías de neurociencia acerca de cuáles son las áreas cerebrales y características involucradas en la producción de habla imaginada. Consecuentemente, se puede concluir que las redes neuronales tipo cápsula logran encontrar patrones descriptivos en las señales EEG de vocales. En el caso de reconocimiento de CapsK-SI se concluye que las áreas cerebrales en la producción de habla

imaginada se centralizan en el hemisferio izquierdo del cerebro, específicamente las áreas de la corteza premotora, el área de Broca, la corteza auditiva secundaria, giro temporal inferior y central, y el área de Wernicke. También se brinda evidencia experimental de que las características estadísticas extraídas de las señales EEG pueden ser reconocidas por las cápsulas para encontrar patrones relacionados con diferencias fonológicas en habla imaginada.

CONTENIDO

LISTA DE FIGURAS	xiii
LISTA DE TABLAS	xiv
INTRODUCCIÓN.....	1
CAPÍTULO 1. ANTECEDENTES	4
1.1 Producción y adquisición de lenguaje	4
1.1.1 Adquisición de lenguaje basado en sonido o articulado	5
1.2 Habla imaginada	6
1.2.1 EEG DaSalla	8
1.2.2 KaraONE.....	12
1.2.3 EEG Arizona	13
1.2.4 EEG Argentina	14
1.2.5 INAOE-IMAGINEDSPEECH DATASET.....	15
1.3 Estado del arte de habla imaginada.....	15
CAPÍTULO 2. ESTUDIO DE VOCALES IMAGINADAS.....	18
2.1 Fisiología relacionada con la fonología de vocales y consonantes.....	18
2.2 Modelos de aprendizaje profundo para el análisis de vocales imaginadas.....	20
2.2.1 Redes neuronales convolucionales.....	20
2.2.2 Redes neuronales tipo cápsula	22
2.2.3 Optimizadores y funciones de aprendizaje	25
2.2.4 Algoritmos de entrenamiento en las redes cápsula	28
2.3 Redes profundas para vocales imaginadas.....	29
2.3.1 Modelo de red neuronal sCNN	30
2.4 Modelo de red neuronal sCapsNet.....	31
2.5 Experimentos y resultados	33
2.5.1 Resultados de reconocimiento de sCNN y CapsNet	34
2.5.2 Comparación con otros trabajos.....	35
CAPÍTULO 3. PROPUESTA DE MODELO BASADO EN CAPSNET.....	37

3.1 Arquitectura del modelo propuesto.....	37
3.2 Capa de entrada y capa convolucional.....	39
3.2.1 Capa de cápsulas primarias PC	39
3.2.2 Capa clasificadora CC.....	40
3.2.3 Capa de reconstrucción	42
3.3 Resultados de clasificación.....	44
3.3.1 Comparación con otros trabajos.....	46
3.3.2 Análisis de los <i>kernels</i> en la capa convolucional	47
CAPITULO 4. RESULTADOS DE RECONOCIMIENTO CON CAPSK-SI.....	51
4.1 Base de datos	51
4.2 Selección de canales	54
4.3 Selección de características	55
4.4 Resultados de clasificación de CapsK-SI.....	61
4.5 Generación de mapas de actividad cerebral.....	67
4.6 Discusión	72
CAPITULO 5. COMPARACIÓN DE RESULTADOS UTILIZANDO CAPSK-SI	73
6. CONCLUSIONES.....	76
BIBLIOGRAFÍA	80
ANEXO A. CREACIÓN DE BASE DE DATOS.....	94
A.1 Justificación	94
A.2 Protocolo de adquisición.....	96
A.3 BCI para la adquisición de señales EEG de habla imaginada.	97
A.4 Composición de la base de datos.....	100

LISTA DE FIGURAS

Figura 1.1. a) Proceso de habla articulada entre dos sujetos b) Proceso de habla imaginada. W: área de Wernicke, B: área de Broca, A: área auditiva, M: corteza motora.	6
Figura 1.2 Modelo de doble flujo de procesamiento de habla Hickok-Poeppel [38].	7
Figura 1.3. Protocolo de grabación para SI [42].	10
Figura 2.1 Posición de labios y lengua en la generación de vocales.	19
Figura 2.2 Diagrama de articulación de vocales.	19
Figura 2.3 Modelo de una red neuronal CNN.	21
Figura 2.4 Arquitectura de CapsNet simple para el reconocimiento de dígitos [94].	23
Figura 2.5 Modelo de red neuronal sCNN.	30
Figura 2.6 Modelo de red neuronal CapsVI.	32
Figura 3.1 Arquitectura del modelo propuesto CapsVI.	39
Figura 3.2 Grafico del modelo de entrenamiento para CapsVI.	43
Figura 3.3 Matrices de confusión para la clasificación de vocales imaginadas.	45
Figura 3.4 Señales de entrada al modelo CapsNet-K. a) Habla imaginada de /a/. b) Estado de control.	48
Figura 3.5 Resultado de la convolución para el habla imaginada de la vocal /a/.	48
Figura 4.1 Áreas cerebrales asociadas al habla.	55
Figura 4.2 Arquitectura del modelo CapsK-SI.	61
Figura 4.3 Activaciones cerebrales en la imaginación de palabras y fonemas bilabiales.	69
Figura 4.4 Activaciones cerebrales en la imaginación de palabras y fonemas nasales.	70
Figura 4.5 Activaciones cerebrales en la imaginación de consonantes y vocales.	71
Figura A.1 Diagrama de articulación de sonidos en el tracto oral.	96
Figura A.2 Protocolo de adquisición de señales EEG.	97
Figura A.3 Cython biosensing board [139].	98
Figura A.4 Distribución de canales para la BCI de habla imaginada.	100

LISTA DE TABLAS

Tabla 1.1 Resumen de otros trabajos relacionados con habla imaginada.	16
Tabla 2.1 Resultados de clasificación de sCNN.....	34
Tabla 2.2 Resultados de clasificación de sCapsNet.	35
Tabla 2.3 Comparación entre modelos con sCNN y sCapsNet.	36
Tabla 3.1 Resultados de desempeño en la validación cruzada del sujeto 1.....	44
Tabla 3.2 Resultados de desempeño en la validación cruzada del sujeto 2.....	44
Tabla 3.3 Resultados de desempeño en la validación cruzada del sujeto 3.....	45
Tabla 3.4 comparación de resultados con otros trabajos de la literatura.	47
Tabla 3.5 Desempeño promedio de validación cruzada con 12 <i>kernels</i> para el sujeto 1.	49
Tabla 3.6 Desempeño promedio de validación cruzada con 12 <i>kernels</i> para el sujeto 2.	49
Tabla 3.7 Desempeño promedio de validación cruzada con 12 <i>kernels</i> para el sujeto 3.	49
Tabla 3.8 Desempeño promedio de validación cruzada con 12 <i>kernels</i> para todos los sujetos.	50
Tabla 4.1 Clasificación de consonantes de acuerdo con el lugar y la manera de articulación. .	53
Tabla 4.2 Propuesta de categorización de palabras y fonemas para KaraOne.	53
Tabla 4.3 Selección de canales y su localización en áreas cerebrales.	55
Tabla 4.4 Características seleccionadas en la base de datos de KaraONE.	59
Tabla 4.5 Ocurrencia de cada característica a través de los sujetos utilizando ANOVA.	60
Tabla 4.6 Resultados de la clasificación bilabial de KaraONE.	63
Tabla 4.7 Resultados de la clasificación NS de KaraONE.	64
Tabla 4.8 Resultados de la clasificación C/V de KaraONE.	65
Tabla 4.9 Resultados de la clasificación /iy/ de KaraONE.....	66
Tabla 4.10 Resultados de la clasificación /uw/ de KaraONE.....	67
Tabla 5.1 Comparación de resultados del modelo CapsK-SI con modelos de la literatura.	74
Tabla A.1 Localización de electrodos en BCI para habla imaginada.....	98

INTRODUCCIÓN

La electroencefalografía (EEG) es una de las técnicas no invasivas más utilizadas en el área clínica para el diagnóstico y seguimiento de actividad cerebral con el fin de conocer la actividad cerebral de un paciente o sujeto. El origen de la señal eléctrica subyace en las células piramidales de la corteza cerebral, donde la sumatoria del flujo del potencial de acción generado por la corriente de millones de neuronas en los potenciales postsinápticos, excitatorios e inhibitorios, es amplificada y combinada por los tejidos para luego ser registrada en el cuero cabelludo, donde se colocan electrodos que captan la diferencia de voltaje entre ellos, dado que conforman dipolos eléctricos [1]. La información captada por los electrodos se le denomina señal EEG [2].

La amplitud de las señales EEG dependerá del potencial neuronal, el área de la corteza cerebral involucrada, sincronización de potenciales, orientación de los dipolos y transformación causada por los elementos interpuestos, tales como músculos en la cabeza [3]. De esta manera, las señales EEG brindan un panorama de la situación neurológica del cerebro.

EEG tiene diversas aplicaciones, dado que es una técnica con las características de accesibilidad, bajo costo, sensibilidad y especificidad [4]. Entre sus aplicaciones destaca el diagnóstico de epilepsia [5]–[8] y desarrollo de sistemas interfaz cerebro computadora (BCI por sus siglas en inglés) [4], [9]. Existen diferentes tipos de estudios de EEG; el EEG estándar consiste en un paciente en reposo y en vigilia, donde se pueden realizar diferentes procedimientos como fotoestimulación, potenciales evocados, hiperventilación, apertura o cierre palpebral [10]. El otro tipo de EEG consta de la privación del sueño enfocada en realizar un estudio con el paciente sin permitirle haber concebido el sueño de forma natural. Lo anterior implica que las horas de privación de sueño deben estar estandarizadas y variar de acuerdo con la edad del paciente. Hasta este momento se han mencionado dos tipos de EEG con fines de diagnóstico aplicados en epilepsia. Sin embargo, también existe un tipo de EEG que consiste en grabar señales y almacenarlas para estudiar modelos matemáticos de reconocimiento, con la intención de desarrollar sistemas BCI o evaluar métodos de procesamiento computacionales [11], [12].

Los sistemas BCI son un conjunto de dispositivos que permiten la comunicación entre el cerebro y una computadora utilizando mediciones de actividad neuronal creando una retroalimentación al sujeto de prueba. Actualmente los sistemas BCI son altamente utilizados por personas con discapacidad [13], sin embargo, es necesario el estudio y análisis con aplicaciones especializadas [14].

La actividad cerebral que se adquiere en BCI puede ser extraída mediante técnicas invasivas o no invasivas. Las técnicas invasivas se definen como aquellas en las que el cuerpo es invadido por un agente externo, ya sea un dispositivo, aguja, instrumento quirúrgico o algún agente clínico. Por ejemplo, la Electrocorticografía (ECoG) que consiste en la introducción de electrodos en la capa cortical del cerebro y posee mayor exactitud en la medición de señales a comparación de EEG, pero es riesgosa para el paciente. Por otro lado, las técnicas no invasivas no involucran instrumentos que rompan la piel o se introduzcan físicamente en el cuerpo. Un ejemplo de dicha técnica es la resonancia magnética funcional (IRMf) es una técnica no invasiva que utiliza la resonancia magnética para observar activaciones en el cerebro generadas por diferentes tareas, no se necesita suministrar ninguna sustancia y generalmente no representa riesgos para el paciente. Sin embargo, la IRMf y otras técnicas no invasivas son muy costosas. De esta manera, las BCI basadas en EEG son menos costosas y se basan en técnicas no invasivas para la adquisición de la actividad cerebral [15].

Las aplicaciones de sistemas BCI comenzaron a desarrollarse para ayudar a personas con discapacidad a comunicarse, desarrollando sistemas de computación que permitan el control de sistemas electrónicos aplicados a sillas de ruedas o brazos robóticos [16]. Asimismo, una de las primeras aplicaciones que se introdujeron fue el desarrollo de sistemas relacionados con el habla. Sin embargo, ante la complejidad que representa el desarrollo de sistemas precisos en su funcionamiento, hoy en día la comunidad científica ha enfocado sus esfuerzos en el desarrollo de métodos y algoritmos que aporten fiabilidad y alto desempeño en las aplicaciones de los sistemas BCI. Se estima que aproximadamente un 60% de los trabajos relacionados con sistemas BCI enfocados al desarrollo y prueba de algoritmos, mientras que el restante es el desarrollo de aplicaciones [14].

Las estrategias mentales que se han utilizado mayormente para el desarrollo de modelos en BCI son los potenciales evocados, imaginería motora y potenciales corticales lentos [17]–[23]. En los últimos años han surgido nuevas estrategias mentales como habla imaginada. Sin embargo, los avances en la investigación del habla imaginada no han sido significativos a comparación de las estrategias mentales como potenciales evocados e imaginería motora. La mayoría de los trabajos relacionados a habla imaginada se centran en el desarrollo de modelos y reporte de desempeños, más no en brindar una explicación detallada de lo que ocurre en el cerebro cuando esta estrategia mental se origina [24]. Por ello, el tema principal de esta tesis está enfocado en el análisis de habla imaginada en señales EEG por medio de un modelo de red neuronal tipo cápsula (CapsNet) para estudiar sus propiedades y lograr el reconocimiento de sílabas, palabras o fonemas. De acuerdo con la información existente en la literatura acerca de la producción de habla articulada, se generan patrones de señales entre áreas cerebrales asociadas a la articulación de lenguaje. En el caso de habla imaginada, se tiene como hipótesis que existen áreas que generan señales cuando se produce el habla articulada, y por medio de estas señales se puede reconocer el habla.

El presente trabajo se conforma de cinco capítulos. En el capítulo I se presentan los antecedentes necesarios para entender el presente proyecto de tesis. En el capítulo II se muestra estudio de vocales imaginadas por medio de modelos de redes neuronales. El capítulo III muestra la descripción del modelo propuesto para el análisis de palabras y fonemas imaginados. El capítulo IV detalla los resultados de experimentación obtenidos. Finalmente, en el capítulo V se encuentran las conclusiones y trabajo futuro que se propone.

CAPÍTULO 1. ANTECEDENTES

Este capítulo describe los antecedentes del presente proyecto de tesis necesarios para comprender cada una de las herramientas que aborda esta investigación. Consecuentemente, la sección 1.1 de este capítulo presenta la producción de lenguaje y fonética, la sección 1.2 introduce el concepto de habla imaginada. Luego, la sección 1.3 detalla la técnica de electroencefalografía para obtener señales cerebrales y la sección 1.4 describe las interfaces cerebro-computadora. Finalmente, en la sección 1.5 se incluye un estado del arte sobre el habla imaginada.

1.1 Producción y adquisición de lenguaje

El ser humano siempre tiene la necesidad de comunicarse con sus semejantes y con los elementos que componen su entorno, tales como animales, dispositivos electrónicos o máquinas. Para lograrlo, el ser humano ha creado diferentes métodos de comunicación que incluyen desde la producción de sonidos con sus cuerdas vocales, hasta la creación de objetos para reproducir sonidos de animales y codificar intenciones o alertas [25]. De acuerdo con la evolución del ser humano, los métodos de comunicación se tornan más complejos por las propias actividades de vida de un humano; se puede decir que los métodos comunicación desarrollados por el humano conforman un lenguaje, el cual se tiene que adquirir por medio de un aprendizaje. Asimismo, la comunicación implica la idea de mutualidad y reciprocidad teniendo el intercambio de información con éxito adaptativo. La comunicación tiene diferentes vías de expresión, entre las cuales resaltan la comunicación gestual y el lenguaje.

De acuerdo con la Real Academia Española (RAE), el lenguaje se define como la facultad del ser humano de expresarse y comunicarse con los demás a través del sonido articulado o de otros sistemas de signos [26]. En otras palabras, el lenguaje puede ser entendido como el modo de relacionarse que surge en el vivir en comunidad lingüística [27]. Dicho esto, el lenguaje se define como el intercambio de información, el cual constituye un fenómeno biológico racional y expresa en su base: función, contenido y forma. Tal y como se ha mencionado, el lenguaje puede ser basado en sonido o movimientos corporales. Para el presente proyecto de tesis nos enfocaremos en el lenguaje basado en sonido o articulado, específicamente a la respuesta

cerebral que se produce al imaginar la articulación de lenguaje, ya que involucra el tema de estudio de interés relacionado con el habla.

Los componentes del lenguaje basado en sonido son: fonemas, morfemas, sintaxis, léxico, semántica, prosodia y discurso. Los fonemas son la unidad mínima fonológica que resulta de la abstracción o descripción teórica de los sonidos de la lengua [28]. Los morfemas son un fragmento mínimo capaz de expresar un significado, por ejemplo, en las palabras “/niño/” y “/niña/”, /a/ es un morfema. La sintaxis es la combinación de las palabras en frases para formar oraciones que tengan sentido. El léxico es el conjunto de todas las palabras de un lenguaje. La semántica es el significado de todo el léxico y combinaciones con sentido. La prosodia es la entonación que puede modificar la percepción o sentido de una palabra o frase. El discurso es constituir una narración a partir de frases [29].

1.1.1 Adquisición de lenguaje basado en sonido o articulado

La adquisición de lenguaje articulado es un proceso integral basado en la neuropsicología. Dicha adquisición se desarrolla de la infancia hasta la adultez, conformando el periodo prelingüístico y la etapa lingüística. El periodo prelingüístico comprende desde el nacimiento hasta los 18-24 meses y se caracteriza porque en este periodo no existe un lenguaje como tal, es meramente comunicación a través de sonidos y gestos [30]. La conducta del infante está llena de movimientos incontrolados y automatismos, pero a medida que se produce el desarrollo de interconexiones neuronales en el cerebro, lo automático se convierte en voluntario y lo no controlado en controlado. Específicamente, el infante se comunica a través de llantos, sonrisas, sonidos, balbuceos, y poco a poco comienza a emitir sonidos vocálicos y señas [31].

Por otra parte, la etapa lingüística comprende de los 2 a los 5-7 años, donde el niño adquiere el lenguaje en su totalidad. Se calcula que al día adquiere un promedio de 10 palabras en su léxico, además, adquiere mayor número de habilidades motrices [32]. En cuanto a lo cognitivo, el niño se encuentra en un periodo de acceso progresivo a la inteligencia representativa; cada objeto se representa por una imagen mental. Alrededor de los 3 años, desaparecen las dificultades para pronunciar diptongos y se produce un avance significativo en la producción de

consonantes. En este punto, y conforme avance en edad, se conformará el lenguaje en su totalidad, adquiriendo la capacidad de articular sin problemas.

1.2 Habla imaginada

El habla imaginada (SI por sus siglas en inglés) es la acción de imaginar la pronunciación del habla de una palabra o vocal sin la generación de movimiento o sonido alguno [33]–[35]. Una de las formas de explicar SI, es necesario recurrir al proceso fisiológico que se lleva en la utilización del lenguaje articulado, detallado generalmente por el modelo de lenguaje Geshwind-Wernicke que se ilustra en la Figura 1.1a. Este modelo sugiere que las áreas de Broca, Wernicke, el fascículo arqueado, la circunvolución angular, la corteza motora y auditiva (41 y 42 de las áreas de Brodmann) están involucradas en la producción y articulación de lenguaje al mantener una conversación. El flujo natural de información va de la corteza auditiva al área de Wernicke donde se realiza una comprensión cognitiva dándole un significado a la información recibida en el fascículo arqueado [36], [37]. Luego de ello, en el área de Broca el significado cognitivo se transforma en una estructura gramatical de una frase, y se almacena en la memoria para la articulación de las palabras. En el caso de habla imaginada en la Figura 1.1b, no existe una inclusión de la corteza motora o área auditiva, y constantemente se encuentra ligado el proceso del área de Broca a Wernicke y viceversa, lo que podríamos llamar, hablar consigo mismo.

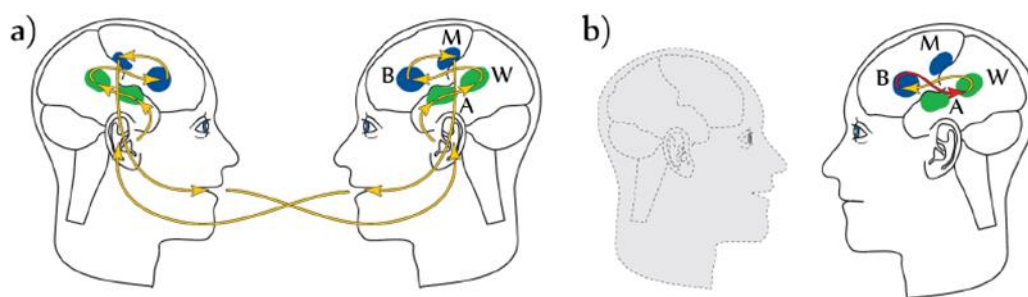


Figura 1.1. a) Proceso de habla articulada entre dos sujetos b) Proceso de habla imaginada. W: área de Wernicke, B: área de Broca, A: área auditiva, M: corteza motora.

Hasta el momento, el concepto de SI ha sido explicado con base en el modelo clásico de generación de lenguaje basado en las áreas de Broca y Wernicke. Sin embargo, recientemente

han surgido nuevos descubrimientos que no han sido incluidos y explorados del todo en habla imaginada, tal es el caso del modelo de Hickok-Poeppel en [38] también llamado modelo de doble flujo que se ilustra en la Figura 1.2. La diferencia de este modelo con el modelo clásico es que existen más áreas cerebrales involucradas en el proceso de escucha y articulación de lenguaje, lo que hace más preciso al modelo en cuanto a la generación de lenguaje.

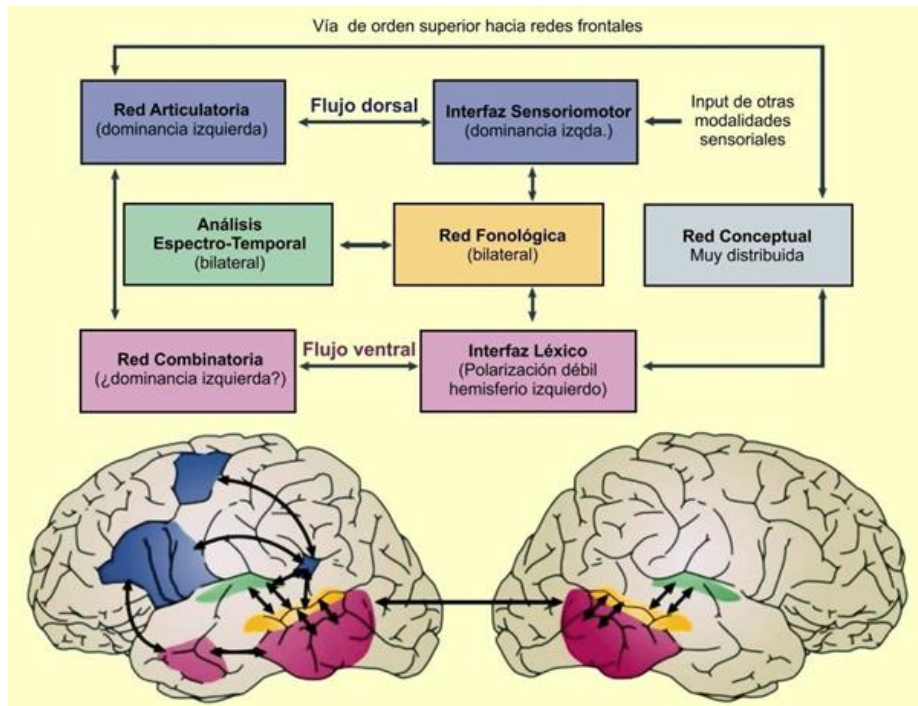


Figura 1.2 Modelo de doble flujo de procesamiento de habla Hickok-Poeppel [38].

El proceso de procesamiento de lenguaje por el modelo Hickok-Poeppel se basa en flujos de información ventral y dorsal. Los flujos ventrales se relacionan con el procesamiento de señales auditivas para la comprensión, y se ubican en estructuras en las porciones medio y superior del lóbulo temporal (regiones azules de la Figura 1.2). Los flujos dorsales, envuelven la traducción de señales acústicas en representaciones articuladas, y se ubican en las regiones del lóbulo temporal posterior y el lóbulo frontal posterior (zonas rosas de la Figura 1.2). De esta forma, en el modelo de Hickok-Poeppel se sugiere que los flujos ventrales son bilaterales (hemisferio derecho e izquierdo), mientras que los dorsales son dominantes en el hemisferio izquierdo, lo que difiere con el modelo clásico donde el lenguaje solo se procesa en el hemisferio izquierdo. Las regiones verdes y amarillas de la Figura 1.2 son regiones tempranas compuestas

por redes fonológicas. La región verde está asociada con características espectro temporales, mientras que la amarilla se relaciona con la representación de fonemas [38].

Con los modelos anteriormente mencionados, el habla imaginada se ha estudiado desde un punto de vista lingüístico y de inteligencia artificial. En cuanto a la lingüística, el habla imaginada se categoriza de acuerdo con la estructura del lenguaje articulado: vocales, fonemas y palabras [39], [40]. El enfoque de inteligencia artificial que se ha utilizado hasta el momento es la detección de la estructura del habla imaginada por medio de modelos de aprendizaje de máquina en señales EEG [41]. En este sentido, se exploró el estado del arte en búsqueda de información relacionada con habla imaginada en EEG, en el que se encontraron cinco trabajos que se desarrollaron y liberaron las bases de datos que se conformaron en dichos trabajos. De este número, cuatro de ellas no cuentan con un nombre oficial, por lo que en esta tesis se nombran de acuerdo con el autor o el lugar donde se desarrolló el trabajo. Las bases de datos recopiladas son:

1. EEG DaSalla [42]
2. KaraONE [43]
3. EEG Arizona [44]
4. EEG Argentina [45]
5. INAOE-IMAGINEDSPEECH Dataset [46]

Las bases de datos se encuentran disponible para su descarga y cada una de ellas cuenta con una descripción acerca de su composición. Además, cada una fue creada con diferentes propósitos, por lo que es importante detallar la composición de cada una de ellas. En las siguientes subsecciones se describe la composición e información de cada una de las bases encontradas.

1.2.1 EEG DaSalla

La base de datos EEG DaSalla fue diseñada por Charles S. DaSalla et al. [42], y tiene como objetivo crear un *framework* para la comunicación de BCI utilizando un esquema de control

natural al incluir actividad neuronal relacionada con la producción del habla. La creación de la base de datos está fundamentada en los potenciales evocados relacionados al habla (SRP por sus siglas en inglés), los cuales son activaciones de la corteza motora al realizar imaginación de vocales (una parte del habla imaginada), siendo la vocal la unidad mínima que compone a una sílaba o fonema.

La base de datos EEG DaSalla se compone de señales EEG de 3 sujetos de prueba de 26-29 años (2 hombres y 1 mujer) que realizan el habla imaginada de las vocales /a/, /u/ y se incluye un estado de reposo que se denomina estado de control, donde el sujeto se concentra en una cruz que aparece en un monitor. Las vocales fueron seleccionadas de acuerdo con la diferencia de activación en los músculos durante la producción del habla articulada. Es decir, para pronunciar la /a/ se necesita abrir la boca mientras que para la /u/ se necesita redondear los labios. El estado de reposo tiene como objetivo el no generar imaginaciones, sino actividad meramente de concentración.

El protocolo de registro de señales se ilustra en la Figura 1.3 y consiste en un proceso de 3 etapas que se muestran en un monitor para indicar instrucciones al sujeto de prueba. La primera etapa tiene como objetivo preparar al sujeto de pruebas en las tareas de imaginación, y para ello, se inicia con un sonido *beep*, e inmediatamente se muestra una cruz en el monitor. La cruz se mantiene en un tiempo aleatorio dentro del rango de 2-3 segundos, para que cualquier componente de Potenciales Evocados Relacionados a Eventos (ERP por sus siglas en inglés) no interfiera cuando se realice la imaginación y prevenir que el sujeto se predisponga o prediga la siguiente etapa en el proceso. En la segunda etapa, se pueden presentar 3 posibles escenarios en el monitor durante 2 segundos. Uno puede ser una silueta de labios abiertos para la imaginación de /a/, otra se refiere a labios cerrados para /u/, y la otra es la misma cruz para el estado de control. Posteriormente, en la última etapa se utiliza para indicar un descanso y se presenta durante 3 segundos la pantalla en blanco, y con ello finaliza el protocolo. La realización del protocolo en las 3 etapas se le denominará sesión.

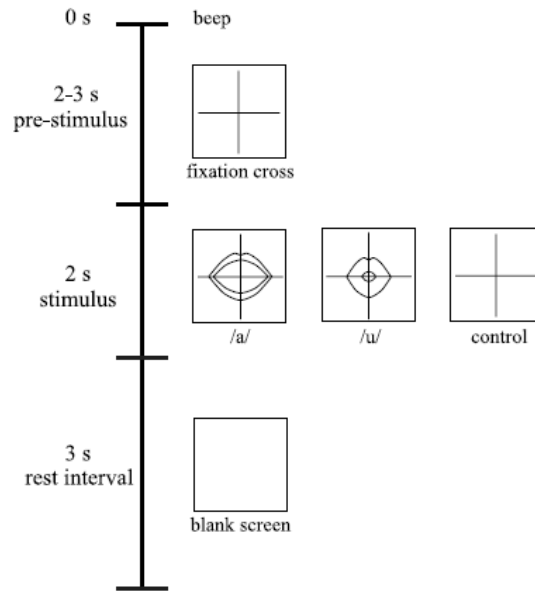


Figura 1.3. Protocolo de grabación para SI [42].

Las señales EEG se registraron con el sistema BioSemi ActiveTwo con 64+8 electrodos activos de Ag/AgCl y una frecuencia de muestreo de 2048 Hz. Se utilizó la “headcap” BioSemi para la posición de 64 electrodos en el cuero cabelludo de acuerdo con el sistema internacional 10-20.

Cada sujeto realizó 150 sesiones, 50 para cada tarea de imaginación. Las señales EEG son procesadas con un filtro pasa bandas con un rango de cortes de frecuencia de 1-45 Hz. Para la selección de la información relacionada con la pronunciación de cada una de las tareas imaginadas, se realizó una selección de épocas con una duración de 3 s (1 s de pre-estímulo y 2 de estímulo). De tal manera que para cada sujeto se extraen 50 épocas de cada tarea y se expresarán mediante E_g^i , donde i indica el número de época y g la tarea a la que corresponde. Los renglones y columnas de la matriz E_g^i corresponden a canales EEG y muestras respectivamente.

Luego de registrar las señales EEG y realizar un filtrado para llevar a cabo la extracción de épocas, se utiliza Patrones Espaciales Comunes (CSP por sus siglas en inglés) con el fin de obtener características basadas en la maximización de varianza entre dos grupos para hacer más fácil la clasificación de señales EEG. Este proceso está dado por la siguiente ecuación:

$$\bar{C}_g = \frac{1}{\gamma} \sum_{i=1}^{\gamma} \frac{E_g^i (E_g^i)^T}{\sum_i (E_g^i (E_g^i)^T)} \quad 1.1$$

$\bar{C}_g$ son las matrices de covarianza promediadas y normalizadas por época de cada imaginación. Mientras que γ es el número de sesiones de cada imaginación. Las dos matrices resultantes se suman para producir una suma de covarianzas C_c , la cual es factorizada en sus eigenvectores, dados por las siguientes ecuaciones:

$$\begin{aligned} C_c &= \bar{C}_1 + \bar{C}_2 \\ C_c &= V_c \lambda_c V_c^T \end{aligned} \quad 1.2$$

V_c es la matriz de *eigenvectores* y λ_c la matriz identidad. Posteriormente, se utiliza la transformada lineal *withening* para ecualizar las varianzas en un *eigen*-espacio y está dada por la siguiente ecuación:

$$\varpi = \sqrt{\lambda_c^{-1}} V_c \quad 1.3$$

La transformada es aplicada a las dos matrices de covarianza de la siguiente manera:

$$S_g = \varpi \bar{C}_g \varpi^T \quad 1.4$$

$$S_1 = V_c \lambda_1 V_c^T \quad S_2 = V_c \lambda_2 V_c^T \quad 1.5$$

Finalmente, se define una matriz de proyección $P = (V_c^T \varpi)^T$ donde las columnas de P^{-1} son los patrones espaciales comunes, los cuales pueden verse como vectores EEG en el tiempo, y descomponer cada época tal que:

$$Z_g^i = P E_g^i \quad 1.6$$

Z_g^i son considerados como los vectores de características. Los autores de la base de datos señalan que solamente los primeros y últimos 2 filtros espaciales son necesarios para la clasificación. De esta manera, los archivos que se descargan en la base de datos están compuestos por los Z_g^i .

y se encuentra disponible en http://www.brainliner.jp/data/brainliner-admin/Speech_Imagery_Dataset.

1.2.2 KaraONE

KaraONE es una base de datos creada por Shunan Zhao et al. [43] y está compuesta por señales de 3 modalidades: EEG, movimientos faciales y audio. Esta base de datos se creó con el objetivo de estudiar la acústica, el movimiento muscular y la actividad cerebral en la producción del habla imaginada o articulada. Para su elaboración, participaron 12 sujetos de prueba con edad media de 27.4 años (4 mujeres y 8 hombres). Para el caso de la grabación de señales EEG, se generan 132 señales por sujeto con los fonemas de /iy/, /uw/, /piy/, /tiy/, /m/, /m/ y /n/, más cuatro palabras derivadas de la lista de Kent de pares fonemas-similares (/pat/, /pot/, /knew/, /gnaw/) [47]. El conjunto de fonemas y palabras fueron seleccionados con el fin de mantener un número relativamente uniforme de consonantes nasales y bilabiales, vocales y fonemas producidos con sonido y sin sonido.

Para adquirir las señales, el sujeto debe permanecer en una silla cómoda donde se le coloca el dispositivo Quick-cap de Neuroscan. Luego de ello, el sujeto fija la mirada en un monitor y deberá producir la menor cantidad de movimientos musculares, mientras se muestran los comandos de los fonemas a lo largo de 30-40 minutos. El protocolo de adquisición de señales se compone de 3 etapas:

1. El sujeto permanece en un estado de reposo durante 5 segundos para relajarse y despejar su mente.
2. Aparece el comando en forma de texto con el fonema/palabra que deberá de producir. Luego se da un tiempo de 2 segundos para permitir que el sujeto mueva o acomode sus articulaciones en caso de haberse movido.
3. Se realiza la imaginación del fonema a producir sin realizar sonidos o movimientos de articulaciones durante 5 segundos.

4. Finalmente, en la etapa 4 se realiza el habla articulada del fonema y con esto termina un intento del protocolo de cada fonema o palabra imaginada.

Para el análisis de la base de datos de KaraOne se han implementado diferentes metodologías en el reconocimiento de palabras y fonemas. Por ejemplo, el análisis que se propuso por parte de los autores Shunan Zhao et al. [43] está basado en la extracción de características estadísticas de media, mediana, desviación estándar, varianza, mínimos, máximos, y entropía espectral, energía, entre otras. Se recurre a encontrar características estadísticas para estudiarlas por medio de correlación, lo cual brinda información acerca de la similitud de actividad producida en diferentes áreas cerebrales.

En la etapa de clasificación, se han propuesto diferentes tipos de categorización, tales como clasificación de presencia y ausencia de palabras y fonemas que contienen producción bilabial, nasal, vocal frontal-alta, vocal trasera-alta y detección de consonante o vocal. Dichas categorías se toman como base en esta tesis, considerando solamente la utilización de señales EEG para investigar la actividad cerebral producida en la imaginación de palabras y fonemas con diferentes características.

1.2.3 EEG Arizona

EEG Arizona es una base de datos creada por Chuong H. Nguyen et. al [44] con el propósito de contribuir al desarrollo de BCI como una opción de comunicación basada en habla imaginada en personas con discapacidad de habla y escucha. La composición de la base de datos consta de las señales EEG de 15 sujetos con edades de 22-32 años (11 masculinos y 4 femeninos) con 3 tipos de habla imaginada: palabras cortas, largas y vocales. Las palabras cortas que se incluyen en la base de datos son /in/, /out/ y /up/, mientras que las palabras largas son /cooperate/ e /independent/. Además, con el fin de evaluar el efecto del sonido, se incluyeron tres fonemas, los cuales son /a/, /i/ y /u/. Para ello, cada uno de los sujetos realiza de 1-3 sesiones de habla imaginada, donde cada sesión comprende la imaginación de 100 intentos por palabra o fonema que se muestran aleatoriamente. Nguyen et. al detallan la implementación de un protocolo de grabación para la obtención de señales EEG. Sin embargo, en el presente proyecto de tesis se intentó replicar el método propuesto por Nguyen et. al, sin embargo, la composición de la base

de datos no coincide con lo reportado en la literatura, por lo que no se consideró esta base de datos en la tesis.

1.2.4 EEG Argentina

EEG Argentina fue creada por Pressen et al. con el objetivo de proveer una base de datos accesible y libre a la comunidad científica para brindar un punto de partida a los investigadores y permitir el desarrollo de nuevos algoritmos [45]. EEG Argentina contiene señales EEG con habla imaginada de 15 sujetos con 25 años promedio (8 hombres, 7 mujeres). Esta base está compuesta por dos categorías de habla imaginada: vocales y palabras. El primer grupo corresponde a las vocales en el idioma español de /a/, /e/, /i/, /o/ y /u/, las cuales fueron seleccionadas de acuerdo con la estacionalidad acústica de su sonido y simplicidad. El segundo grupo corresponde a las palabras de /arriba/, /abajo/, /derecha/, /izquierda/, /adelante/ y /atrás/. Estas palabras se seleccionaron ya que se asocian como posibles comandos para el control de un dispositivo externo en un sistema BCI. Para las sesiones de registro de señales EEG, cada sujeto es colocado en una silla cómoda frente a un monitor que presenta los comandos que indican las vocales o palabras que se deben imaginar bajo un protocolo que se compone de un proceso de 12 segundos por imaginación.

Las señales fueron registradas con el dispositivo de adquisición 18-Channel Grass®, utilizando 6 electrodos húmedos de Ag/AgCl colocados de acuerdo con el sistema internacional 10/20 con una frecuencia de muestreo de 1024 Hz. Los resultados de la creación y análisis de la base de datos EEG Arizona por medio de características extraídas por *Wavelet* de un canal y cinco niveles de descomposición en [45] son de aproximadamente 20% y 16% para el reconocimiento de vocales y diferentes comandos respectivamente.

Para comprobar el estado y fiabilidad de la base de datos, se replicó en esta tesis el método reportado, sin embargo, los resultados de experimentación no coincidían con lo reportado. Bajo un estudio de las señales que se realizó en el presente proyecto de tesis, al igual que con la base de datos EEG Arizona, se observó una mala documentación de la composición de las señales, por lo que se descartó para ser utilizada en este proyecto de tesis.

1.2.5 INAOE-IMAGINEDSPEECH Dataset

EEG INAOE es una base de datos creada por Torres et al. [48] con el objetivo de reconocer comandos de habla imaginada en español para el control de sistemas BCI. Esta base de datos se compone de señales EEG del habla imaginada de 27 sujetos cuyo idioma materno es el español. Las señales se obtienen con el dispositivo Emotiv EPOC+ con 14 electrodos húmedos. Para las evaluaciones del método propuesto, se utilizan 4 electrodos localizados en las posiciones F7, FC5, T7 y P7. Para esta base de datos se seleccionaron las palabras en español para imaginar /arriba/, /abajo/, /derecha/, /izquierda/, /seleccionar/. El protocolo de grabación consiste en sentar a la persona en una silla cómoda frente a un monitor, y la mano derecha sobre un ratón de computadora. La función del este último es mandar marcadores al equipo de cómputo que indican el inicio y fin de la imaginación de alguna de las 5 palabras. Para cada palabra se realizan 33 imaginaciones de forma aleatoria. Además, esta base de datos no brinda más información acerca del protocolo de adquisición que se siguió para adquirir las señales de habla imaginada.

La base de datos está disponible para su descarga mediante una solicitud donde se explica el objetivo que se tiene al estudiar dicha base de datos. Para la experimentación, se reciben las señales de 7 sujetos solamente con el propósito de generar los primeros experimentos con la base de datos. Cabe mencionar que es una base de datos bastante documentada, ya que cada uno de los archivos está descrito de acuerdo con su composición y propósito. Sin embargo, no se encuentra disponible para su descarga libre a diferencia de las otras bases de datos.

1.3 Estado del arte de habla imaginada

Mediante un estado del arte de la literatura, existen diversos trabajos en la literatura que se han desarrollado para adquirir y procesar señales de habla imaginada. Algunos de los ejemplos de trabajos encontrados se presentan en la Tabla 1.1, la cual incluye información de tipo de clasificación se realiza (palabras, fonemas o vocales), el desempeño obtenido, número de clases, y el año en que se realizó dicho trabajo. La Tabla 1.1 reúne resultados de clasificación dado que los trabajos que se muestran, solamente incluyen la el desempeño de clasificación como métrica de comparación.

Tabla 1.1 Resumen de otros trabajos relacionados con habla imaginada.

Algoritmo	Clasificación	Desempeño y número de clases				Año
		2	3	4	>4	
CSP y SVM [42]	Fonemas: /a/, /u/	78%	-	-	-	2009
K-nearest neighbor [49]	Sílabas: /ba/, /ku/	61%	-	-	-	2010
Análisis discriminante lineal [50]	Fonemas: /a/, /e/, /i/, /o/, /u/	72%	-	-	-	2016
SVM y distancia Riemanniana [44]	Fonemas: /a, /i/, /u/ Palabras: ‘in’, ‘cooperate’	49.2% 80.1%	-	-	-	2017
SVM y red neuronal [51]	Palabras: ‘yes’, ‘no’ Palabras hindi: ‘haan’, ‘na’	75.4%	-	-	-	2017
Transformada de <i>Wavelet</i> y red neuronal [52]	Palabras: ‘yes’, ‘no’	67%	-	-	-	2017
SVM [53]	Sílabas: /iy/, /piy/, /tiy/, /diy/, /uw/, /m/, /n/ Palabras: ‘pat’, ‘pot’, ‘knew’, ‘gnaw’	-	-	33.33%	-	2018
Red neuronal convolucional [54]	Números en inglés de 0-9	-	-	-	92.01%	2018
Red neuronal jerárquica [55]	Sílabas: /iy/, /piy/, /tiy/, /diy/, /uw/, /m/, /n/ Palabras: ‘pat’, ‘pot’, ‘knew’, ‘gnaw’	83.42%	-	-	66.2%	2019
Red neuronal recurrente [56]	Fonemas: /a/, /e/, /i/, /o/, /u/ Palabras: ‘yes’, ‘no’, ‘left’, ‘right’	- -	- -	- 99.38%	98.61% -	2019
Red neuronal recurrente [57]	Palabras con la estructura de vocal-consonante-vocal	72%	-	-	-	2019
Red Perceptrón Multicapa (MLP) [58]	Palabras: ‘left’, ‘right’, ‘up’, ‘down’	85.57%	-	49.77%	-	2020

Como se puede observar en la Tabla 1.1, los desempeños varían de acuerdo con el número de clases y el año de elaboración del proyecto. Sin embargo, los desempeños pueden verse favorecidos dependiendo el tipo de señales que se utilizan para el reconocimiento del habla. En algunos casos, señales EEG y de audio son utilizadas, y por esa razón se obtienen desempeños altos.

El habla imaginada ha sido estudiada bajo diferentes esquemas de procesamiento para diferentes aplicaciones [59]–[63]. En esta sección se aborda el tema y se explican diversas metodologías utilizadas en el reconocimiento de vocales, fonemas y palabras imaginadas. Dado

que el presente proyecto de tesis se centra en el estudio habla imaginada en señales EEG, se acotará el estado del arte a aquellos métodos que utilizan señales EEG.

Para la extracción de características de habla imaginada se ha utilizado técnicas como CSP [42], [64], extracción de características estadísticas [43], [50], y la transformada Wavelet [65]. En algunos casos, la extracción de características es utilizada por modelos de redes neuronales. En este sentido, se han utilizado redes neuronales [64], [66], máquinas de soporte vectorial (SVM por sus siglas en inglés) y análisis discriminante lineal (LDA por sus siglas en inglés) [42], [50]. De los métodos que demuestran mayor eficiencia se encuentran las redes neuronales, razón por la cual serán utilizadas en el presente proyecto de tesis y serán explicadas en las siguientes secciones.

CAPÍTULO 2. ESTUDIO DE VOCALES IMAGINADAS

En esta sección se presenta un estudio de la base de datos de DaSalla para lograr el reconocimiento de patrones producidos en la imaginación del habla. Primero se estudia la información relacionada con el habla que existe en la literatura para conocer las bases de fonología que permiten diferenciar la producción de dos vocales. Posteriormente, se realiza un análisis de las redes neuronales profundas CNN y CapsNet para conocer cómo procesan las señales EEG de dos vocales imaginadas de la base de datos de EEG DaSalla. Este análisis tiene el fin de abordar el fenómeno de habla imaginada en su nivel más básico que es la imaginación de vocales para obtener el diseño de un modelo de reconocimiento, ya sea de tipo CNN o CapsNet. Asimismo, se busca que los modelos puedan brindar información acerca del patrón con el que se está clasificando la imaginación de vocales para encontrar características que apoyen en la clasificación de fonemas y palabras.

Después de una revisión de información en la literatura relacionada con la producción de vocales, se encontró la teoría consonante vocal, por lo que será incluida y detallada en la sección 2.1. En la sección 2.2 se describe el modelo sCNN compuesto por redes neuronales convolucionales. En la sección 2.3 se detalla el modelo de sCapsNet basado en redes neuronales tipo cápsula y en la sección 2.4, se reporta los resultados de reconocimiento de la base de datos EEG DaSalla.

2.1 Fisiología relacionada con la fonología de vocales y consonantes

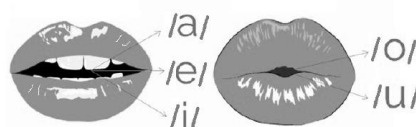
Para este estudio se toma como base la información encontrada en la literatura de la generación de vocales imaginadas en la producción de sonidos de acuerdo con su articulación. En la Figura 2.1 se muestra la generación de vocales imaginadas teniendo en cuenta la posición de la lengua y labios para formar el diagrama de articulación de la Figura 2.2.

Para comprender las vocales desde un punto de vista fonológico se utiliza el diagrama de articulación [67] de la Figura 2.2, del cual se pueden determinar diferencias específicas en la producción de distintas vocales. El diagrama muestra las posiciones de la lengua en el paladar para generar distintos sonidos. De acuerdo con esta diferencia fonológica, relacionada con la

posición de labios y lengua se distinguen las vocales. En cuanto a la imaginación de vocales, DaSalla et al. [42] concluyó que en la imaginación de las vocales /a/ y /u/ existe la producción de SRP. Estos potenciales se presentan alrededor de los 350-400 ms después de haberse presentado la imaginación. Sin embargo, en el caso de la producción de fonemas compuestos por vocales y consonantes, los SRP de cada letra se traslapan entre sí, lo que dificulta el reconocimiento de cuales vocales y consonantes fueron imaginadas [68]. Los SRP se asocian a potenciales evocados endógenos, y se relacionan con el potencial Bereitschaftspotential (PB) [69].



a) Posición de los labios



b) Posición de los labios

Figura 2.1 Posición de labios y lengua en la generación de vocales.

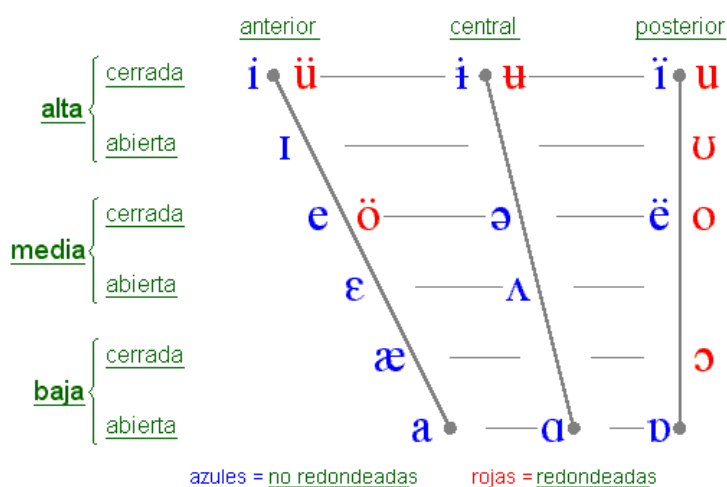


Figura 2.2 Diagrama de articulación de vocales.

El PB [69] es un potencial cortical negativo e inicia con 1-1.5 segundos de anticipación al inicio de la articulación de lenguaje. Wohler et. al [70] indica que, para movimientos musculares existe un potencial de preparación en la zona cortical, representado por el potencial PB. En el caso de movimientos orales el PB puede verse afectado por la naturaleza lingüística según sea lo que se pronunciará. En el habla articulada de repeticiones de palabras, se puede asociar a un PB en el vértice craneal de mayor amplitud que en el caso de habla imaginada, el cerebro genera actividad diferente en la producción de vocales y consonantes, se generan potenciales endógenos relacionados con eventos (ERP por sus siglas en inglés) [71], [72]. Para la pronunciación o imaginación de vocales, se generan potenciales con deflexiones positivas posterior a la pronunciación o imaginación. En el caso de la pronunciación de consonantes y vocales, dichos potenciales se traslapan [72], [73].

La teoría consonante-vocal indica que, los fonemas desencadenan un ERP negativo posterior en la producción de consonantes y un ERP negativo anterior en la producción de vocales [68]. Estos ERP pueden codificarse como ERPs endógenos presentes con diferentes picos, amplitudes, latencias y polaridades. Mediante esta teoría, se utilizarán las redes neuronales CNN y CapsNet para estudiar los patrones de reconocimiento que se utilizan para lograr reconocer vocales imaginadas. A continuación, se describen a detalle cada uno de los modelos utilizados.

2.2 Modelos de aprendizaje profundo para el análisis de vocales imaginadas

De acuerdo con el estado del arte [12], [14], [24], [41], [42], [44], [49]–[54], [56]–[58], [74]–[86], las vocales, fonemas y palabras imaginadas se han procesado con redes neuronales, entre las cuales destacan las redes neuronales convolucionales en el reconocimiento de vocales, fonemas y palabras imaginadas por sus altos desempeños de reconocimiento. Por ello, en esta sección se describen las redes neuronales que fueron seleccionadas en este trabajo de tesis para experimentación.

2.2.1 Redes neuronales convolucionales

Esta red se propuso en el año 1998, y son redes profundas de aprendizaje supervisado que se componen de varias capas convolucionales, y su funcionamiento se asemeja a la corteza

visual primaria del cerebro cuando identifica características visuales y lograr el reconocimiento de patrones. Fueron desarrolladas inicialmente para el área de visión por computadora, pero han sido extensamente utilizadas en aplicaciones que involucran procesamiento de voz, visión por computadora [87], [88], procesamiento de señales [89]–[92], etc. La estructura de una CNN para clasificación se muestra en la Figura 2.3, la cual se compone por capas convolucionales seguidas de capas de submuestreo y por último una capa totalmente conectada que es similar a una red MLP [93].

Para el ejemplo mostrado en la Figura 2.3, se tienen las entradas $x(n,m)$ con una dimensión $m \times m \times r$, donde m se refiere al largo y ancho de la imagen y r al número de canales que se utilizan, que para el caso del ejemplo $r=3$ por la representación RGB de la imagen. El análisis del funcionamiento de la arquitectura presentada se hará con base en la arquitectura de la Figura 2.3.

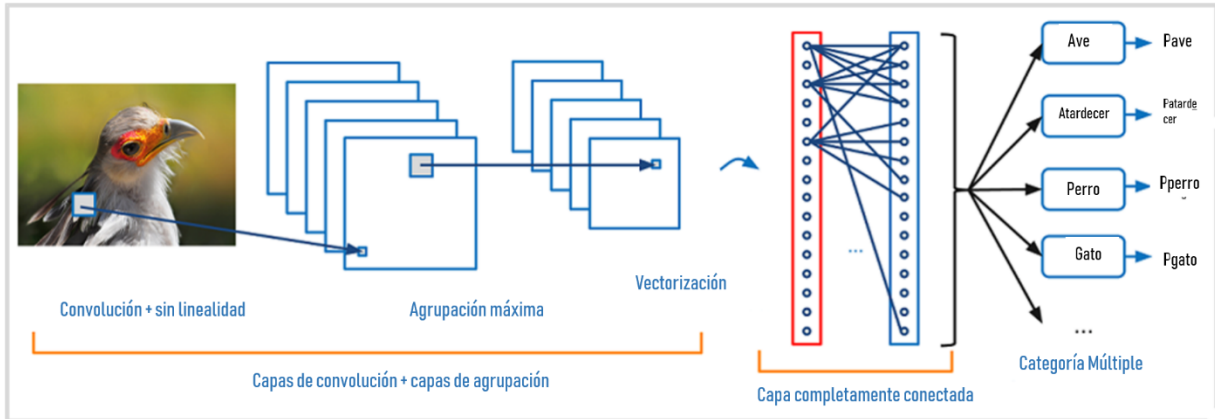


Figura 2.3 Modelo de una red neuronal CNN.

En cada capa convolucional se localizan k filtros lineales de tamaño $n \times n \times q$, donde $n < m$ y $q \leq r$. Los filtros están compuestos por los pesos $W_k(n,n)$ y se convolucionan con la imagen de entrada y la polarización b_k para generar mapas de características c_k , cada una de tamaño $m - n - 1$. c_k está dado por una función de activación f y se caracteriza por la siguiente ecuación:

$$c_k = f(W_{qk}(n,n) * x_r(m,m) + b_k) \quad 2.1$$

Cada c_k se procesa en la capa de submuestreo donde se realiza una reducción dimensional para disminuir la cantidad de información utilizada en la red. Para lograr esto, se pueden realizar dos operaciones que dividen el mapa de características en k submapas. La primera operación consiste en reducir su número de elementos a uno tomando como único valor el máximo de ellos, de tal manera que se tendría como resultado un nuevo mapa de características con dimensiones $\frac{z}{2} \times \frac{z}{2}$. La segunda operación consiste en realizar un promediado de elementos en vez de tomar el valor máximo. Estas operaciones son definidas por las siguientes ecuaciones:

$$cm_k = \left\{ \arg \max_k (c_k) \parallel \frac{1}{k} \sum c_k \right\} \quad 2.2$$

donde se toma el máximo de cada mapa de características c_k o se realiza un promedio de este. De esta manera, se genera una representación cm_k del mapa original c_k con menor dimensión. Ahora bien, se ha explicado el funcionamiento de una capa convolucional, y en el caso de la Figura 2.3, se tienen solamente dos capas de este tipo, pero pueden ser duplicadas en mayor cantidad. Finalmente, cada mapa de características cm_k es propagado a la capa totalmente conectada y tiene una estructura similar a las redes MLP, lo cual está dado por la siguiente ecuación:

$$y_r = f \left(\sum w_{o_k} \cdot cm_k + bo_k \right) \quad 2.3$$

donde r corresponde al número de neuronas involucradas en esta capa, las cuales determinan la cantidad de clases a la salida. w_o son los pesos asociados a la capa totalmente conectada y bo_k su polarización.

2.2.2 Redes neuronales tipo cápsula

Las redes cápsula o CapsNet fueron propuestas en el año 2017 por Sabour et al. [94], y al igual que las redes CNN parten del proceso de visión humano. Estas redes proponen básicamente en agrupar las neuronas de cada capa en conjuntos denominadas “cápsulas”. Donde cada grupo se puede relacionar a un tipo específico de entidad, como un objeto, o una parte del objeto según sea el caso de la entrada [95]. Por ende, la salida de una cápsula es un vector que

está compuesto por la actividad de cada neurona de la misma cápsula y es llamado vector de actividad, y a través de este es por el que se accede a la información relacionada a la entidad, ya sea, bordes, posición, orientación, tamaño, etc.

El funcionamiento de la red se basa en que solo se activarán las cápsulas asociadas a ciertas características de la entrada. Por ejemplo, en el caso de los rostros, una cápsula puede estar asociada a la representación de los labios, de tal manera que se obtendrá información en un vector de su morfología. Consecuentemente, aquellas cápsulas que respondan ante características en específico de la entrada representan activaciones y cada una de ellas direcciona su información a una cápsula clasificadora de una capa superior para transferir su información.

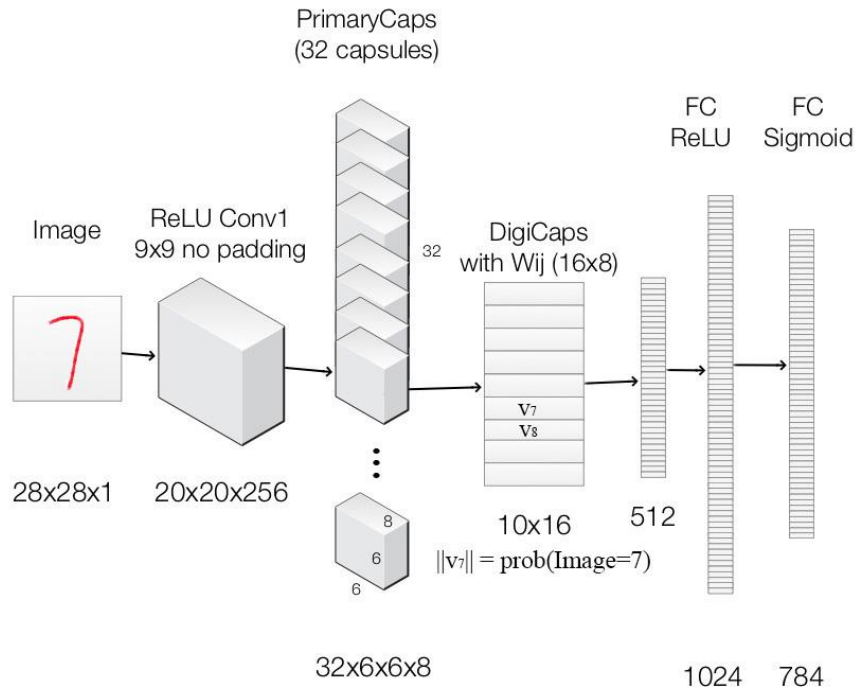


Figura 2.4 Arquitectura de CapsNet simple para el reconocimiento de dígitos [94].

La arquitectura de una CapsNet se compone de dos etapas: codificación y decodificación. La etapa de codificación está dividida en capas de: convolución, cápsulas primarias y cápsulas reconocedoras de dígitos (DigiCaps). La etapa de decodificación está compuesta solamente por capas totalmente conectadas. En la Figura 2.4 se muestra un ejemplo de una arquitectura de una

CapsNet utilizada para el reconocimiento de dígitos en imágenes. Este ejemplo está compuesto por una capa convolucional, una capa de cápsulas primarias, una capa de DigiCaps y tres capas totalmente conectadas.

En la capa ReluConv1 del ejemplo en la Figura 2.4, se realiza una convolución estándar como la ecuación 2.1. En este caso se utilizan 256 *kernels* de 9x9, y se genera una salida de 20x20x256, lo que conforma el mapa de características de salida de la capa convolucional.

Una cápsula se define como la agrupación de neuronas. La capa de cápsulas primarias toma como entrada el mapa de características de salida de la capa ReluConv1, para generar una matriz de 8 dimensiones. Ahora bien, la forma de calcular los vectores de salida de una cápsula se realiza por medio de una función no lineal llamada *squashing* que garantiza que aquellos vectores con módulo pequeño se reduzcan a una magnitud casi nula, y los vectores con módulo grande a una longitud inferior a 1. La función *squashing* está dada por la siguiente ecuación:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad 2.4$$

donde v_j es el vector de salida de la cápsula j y s_j es la entrada a la cápsula.

La entrada s_j es una suma ponderada sobre todos los “vectores de predicción” $\hat{u}_{ji}$ de las cápsulas de capas inferiores, los cuales son producidos mediante la multiplicación de la salida u_i de una cápsula en una capa inferior por una matriz de pesos W_{ij} y está dado por la siguiente ecuación:

$$s_j = \sum_i c_{ij} \hat{u}_{ji}, \quad \hat{u}_{ji} = W_{ij} u_i \quad 2.5$$

donde c_{ij} son los coeficientes de acople los cuales son determinados por el proceso de enrutamiento iterativo dinámico [94]. De tal manera que, la suma de los coeficientes entre la cápsula i y todas las cápsulas en las capas superiores suman uno para representar probabilidades de que una entidad en la entrada exista. Los coeficientes son determinados por una función “softmax de enrutamiento”, cuyos valores iniciales b_{ij} son las probabilidades previas de registro de que la cápsula i debiera de estar acoplada a la cápsula j [94].

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad 2.6$$

En el ejemplo de la Figura 2.4, se utilizan *kernels* de 8x32 en la capa de cápsulas primarias para generar 32 cápsulas de 8 dimensiones (8D), siendo 8 salidas que se agruparon para formar una cápsula. De igual forma, se utilizan *kernels* de 9x9 con *stride* de 2 y no *padding*, con el fin de reducir la dimensión espacial de 20x20 de la etapa anterior a 6x6. De esta forma, la composición de las cápsulas en la capa *PrimaryCaps* será de 32x6x6x8.

En la siguiente capa de DigiCaps se utiliza a la matriz W_{ij} de pesos con dimensiones 16x8 para convertir las cápsulas de 8D a cápsulas de 16D en cada clase j con la ecuación 1.11 y así calcular s_j y v_j . Dado que esta arquitectura está diseñada para el reconocimiento de dígitos en imágenes, existen 10 clases. Por lo tanto, la salida de la capa DigiCaps será de 10x16. Finalmente, en la etapa de decodificación, por medio de 3 capas de totalmente conectada, se obtiene como salida una imagen de dimensiones 28x28.

2.2.3 Optimizadores y funciones de aprendizaje

Los modelos de redes neuronales profundas han tenido un auge en el procesamiento de imágenes y video como lo son las CNN y Redes Neuronales Recurrentes (RNN por sus siglas en inglés). Sin embargo, es importante destacar que el proceso de entrenamiento de estas redes se basa en la teoría de optimización de un gradiente, un concepto fundamental en el aprendizaje de máquina. Desde el punto de vista de un gradiente, los métodos de optimización pueden ser divididos en:

1. Optimización de primer orden: son aquellos representados por los métodos de gradiente estocástico y sus variantes.
2. Alto orden: tales como el método optimizador de Newton [96], y suelen ser llamados así por su rápida convergencia.
3. Heurísticos: comúnmente representado por el método de *Coordinate descent* [97].

En el caso de los optimizadores de primer orden, se basan en el método de gradiente descendente (GD), que consiste en que los parámetros se actualizan iterativamente en la dirección opuesta al gradiente de la función de costo para lograr su convergencia. A continuación, se asume que la función $f_{\theta}(x)$ es la función de aprendizaje, $L(\theta)$ la función de pérdida y θ el parámetro a optimizar, el objetivo de gradiente descendente es minimizar la función de pérdida por medio de:

$$L(\theta) = \frac{1}{2N} \sum_{i=1}^N (y_i - f_{\theta}(x_i))^2, \quad 2.7$$

$$f_{\theta}(x) = \sum_j^D \theta_j x_j \quad 2.8$$

donde N es el número de muestras de entrenamiento, D es el número de características de entrada, x_i es una variable independiente con $x_i = (x_{i1}, \dots, x_{iD})$ para $i=1, \dots, N$ y y_i es el objetivo. Para realizar la convergencia, se realizan lo siguiente en cada iteración:

1. Obtener la derivada de $L(\theta)$ en θ_j para obtener el gradiente correspondiente:

$$\frac{\partial L(\theta)}{\partial \theta_j} = -\frac{1}{N} \sum_{i=1}^N (y^i - f_{\theta}(x^i)) x_j^i \quad 2.9$$

2. Actualizar cada θ_j en una dirección negativa al gradiente para minimizar a función de costo

$$\theta'_j = \theta_j + \eta \cdot \frac{1}{N} \sum_{i=1}^N (y^i - f_{\theta}(x^i)) x_j^i \quad 2.10$$

El método GD es simple de implementar, pero converge lentamente cuando la variable es cercana a la solución, de tal manera que se necesitarán de más iteraciones para lograr el aprendizaje. Además, la complejidad O del método en cada iteración estará dada por el número de muestras N y la dimensión D de la variable x , de tal manera que se define como $O = DN$.

Para disminuir el costo computacional, emerge un método llamado Gradiente Descendente Estocástico (SGD por sus siglas en inglés). El método SGD utiliza una muestra aleatoriamente para actualizar el gradiente en cada iteración, en vez de calcular directamente el valor del gradiente. El costo computacional de SGD es independiente del número de muestras y puede alcanzar convergencia más rápida que GD ya que reduce el tiempo de actualización de parámetros. Sin embargo, tiene la desventaja ante GD de que no puede ser utilizado en aprendizaje en línea. Su función de costo está dada por:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (y^i - f_{\theta}(x^i))^2 = \frac{1}{N} \sum_{i=1}^N \text{cost}(\theta, (x^i, y^i)) \quad 2.11$$

Si una muestra i aleatoria es seleccionada en SGD, la función de perdida será $L^*(\theta)$:

$$L^*(\theta) = \text{cost}(\theta, (x^i, y^i)) = \frac{1}{2} (y^i - f_{\theta}(x^i))^2 \quad 2.12$$

De tal manera que para la actualización se utiliza solamente una muestra aleatoria, en vez de todas las muestras en cada iteración:

$$\theta' = \theta + \eta (y^i - f_{\theta}(x^i)) x^i \quad 2.13$$

Debido a que SGD utiliza una muestra aleatoria para la actualización, el costo computacional está dado solamente por $O(D)$, la cual es la longitud de la muestra x . Finalmente, SGD incrementa la velocidad de alcanzar la convergencia. Sin embargo, uno de los problemas de este método es que la dirección del gradiente oscila debido a la selección aleatoria de muestras y, además, la varianza de gradientes la dirección está polarizada (a diferencia de GD que siempre minimiza hacia el valor optimo en dirección negativa al gradiente). Aunque SGD es de los métodos más utilizados, el gradiente determinista puede causar que se caiga en un mínimo local en problemas multimodales. Asimismo, existen algunos detalles que pueden causar problemas a la hora de utilizar este optimizador, como lo es elegir el coeficiente de aprendizaje adecuado. A un coeficiente pequeño, se dará la convergencia a largo plazo, mientras que a un coeficiente

grande se quedará oscilando en cierto valor de la función. Comúnmente este valor se ajusta empíricamente según sean los resultados y características del conjunto de muestras.

Ante los problemas acerca de los mecanismos para elegir el coeficiente de aprendizaje óptimo, cómo alcanzar mayor velocidad de aprendizaje y convergencia, se agrega un *momentum* a SGD. Este concepto parte de la idea física basada en la inercia de los objetos se propone. Este mecanismo se sustenta en preservar la influencia de la actualización de dirección de la iteración anterior. El algoritmo de *momentum* introduce la variable v conocida como velocidad, la cual representa la dirección y grado de movimiento, lo cual se expresa por:

$$v = \eta \cdot \left(-\frac{\partial L(\theta)}{\partial \theta} \right) + v^{old} \cdot mtm \quad 2.14$$

donde mtm es el factor de *momentum* que generalmente vale entre 0 y 1 y v^{old} la velocidad de la iteración anterior.

El algoritmo SGD con *momentum* desempeña un papel importante en reducir la oscilación de convergencia cuando el aprendizaje es prolongado [98]. Sin embargo, seleccionar su parámetro mtm también representa un problema al igual que el parámetro de aprendizaje η . Si el momentum es muy pequeño, es difícil de obtener una mejora en la velocidad de convergencia, y si es grande, se brincaré entre mínimos locales sin encontrar el mínimo global.

Además, existe el método de optimización gradiente descendiente por pequeños lotes (MSGD), el cual utiliza b muestras para actualizar parámetros en cada iteración. Lo anterior reduce la varianza de gradientes y hace posible una convergencia más estable, lo que ayuda a mejorar la velocidad de optimización.

2.2.4 Algoritmos de entrenamiento en las redes cápsula

Las redes cápsula se presentaron inicialmente en el 2011 con el título de “transformación de auto-encoders” por Hinton et al. [99] con el objetivo de reconocer poses de un objeto. Sin embargo, el entrenamiento del modelo resultó ser deficiente, dado que había problema en la

comunicación de cápsulas entre capas. Ante lo anterior, se han propuesto 2 algoritmos de entrenamiento para solucionar el problema de conexión entre cápsulas:

1. Enrutamiento dinámico iterativo entre cápsulas [94]
2. Matrices de cápsulas basadas en enrutamiento de expectation-maximization [100]

Sabour et al. [94] proponen la primera modificación al concepto de redes cápsula, con la implementación del algoritmo de entrenamiento de “enrutamiento dinámico iterativo entre cápsulas” donde las cápsulas direccionan su información a cápsulas de capas superiores por medio de semejanzas en las entidades representadas con los vectores de actividad de cada cápsula. Este algoritmo fue previamente explicado en la subsección 2.2.2, el cual es el más utilizado para el entrenamiento de redes neuronales tipo cápsula.

Hinton et al. [100] proponen la modificación al modelo de Sabour [100], donde utilizan las cápsulas para determinar la presencia de una entidad en la entrada, y una matriz de 4x4 en la salida de cada cápsula para representar la relación entre esa entidad y sus características. Al igual que en el modelo de Sabour, se utilizan coeficientes de acoplamiento a la salida de cada capa para determinar hacia qué cápsula en capas superiores se mandará la información. La modificación propuesta por Hinton radica en el uso de Expectation Maximization (EM) [101], donde cada coeficiente se actualiza iterativamente. Hasani et al. utilizan de primera instancia el modelo de EM para determinar el número de iteraciones optimas con el fin de reducir el tiempo de entrenamiento y el error de clasificación [102].

2.3 Redes profundas para vocales imaginadas

Con la información presentada en la sección 2.2, que incluye modelos profundos promisorios para ser evaluados con el habla imaginada y métodos de optimización, se presentan a continuación dos modelos que se utilizaron en el presente proyecto de tesis para el reconocimiento de vocales imaginadas.

2.3.1 Modelo de red neuronal sCNN

Se ha observado en el estado del arte de habla imaginada que el aprendizaje profundo ha tenido un auge en los últimos años y, además, ha demostrado ser de utilidad para incrementar el desempeño de reconocimiento en señales EEG [103]. Sin embargo, los modelos presentados en los antecedentes no brindan información acerca de lo que significa estos desempeños, la investigación hasta el momento ha estado enfocada en probar herramientas para la clasificación de habla imaginada. Por ello, en esta tesis se muestra el desarrollo de un modelo CNN para generar conocimiento del fenómeno de vocales imaginadas. El estudio se enfoca en examinar la base de datos EEG DaSalla utilizando los filtros convolucionales para buscar características que puedan ser de utilidad en el reconocimiento de vocales imaginadas, para luego ser utilizadas en el reconocimiento de palabras y fonemas.

En esta sección se utiliza un modelo básico de redes tipo CNN, el cual será llamado sCNN (CNN relacionado a *speech*). La Figura 2.5 muestra su arquitectura, la cual fue seleccionada por su similitud con las redes cápsula. sCNN se compone por cinco capas ($l = 1, \dots, 5$): que corresponden a una capa de entrada, dos capas convolucionales, dos capas completamente conectadas y una capa de salida.

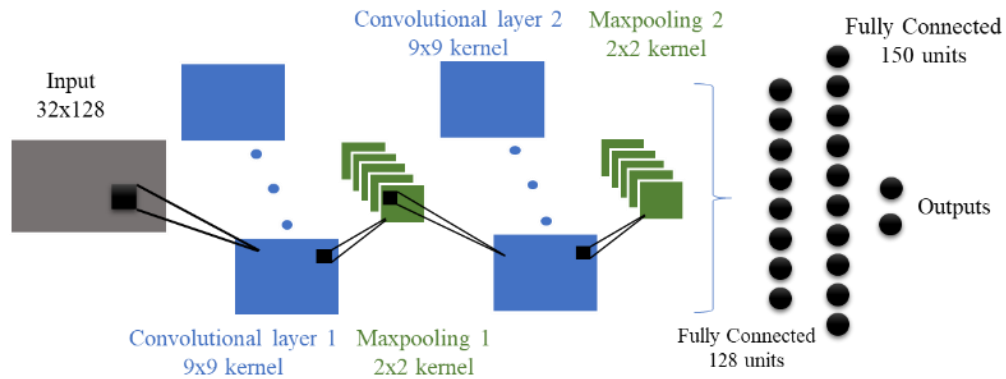


Figura 2.5 Modelo de red neuronal sCNN.

La entrada de la red es la señal EEG CSP proyectada en las matrices Z_g^i de la base de datos de DaSalla que se presentó en la sección 1.2.1. Las capas $l = 2$ y $l = 3$ son las convoluciones dadas por:

$$c_{lk} = f(W_{lk} * Z_g^i + b_{lk}), \quad l = \{1, 2\} \quad 2.15$$

donde c_{lk} son los mapas de características generados por la convolución entre la entrada Z_g^i y k número de filtros convolucionales formados por un conjunto de pesos W_{lk} y una polarización b_{lk} . Cada c_{lk} tiene una función de activación f que en este caso se utilizó una unidad de rectificación lineal (ReLU) [104]. Ambas capas tienen 256 filtros con un *kernel* de 9x9 y *stride* de 2. La función ReLU fue seleccionada porque su derivada siempre es igual a 1 si la entrada es mayor a cero [105] y previene el problema de desvanecimiento de gradiente [106]. Después de cada ReLU, se encuentra el proceso de *max-pooling*. Sea z_{pqk} un elemento en una ventana (kernel) por cada c_{lk} de tamaño $H \times H$, de tal forma que el proceso estará dado por:

$$u_{ijk} = \max_{p,q \in c_{lk}} z_{pqk} \quad 2.16$$

donde u_{ijk} es la salida, y p, q son los índices para cada elemento de c_{lk} .

Las capas $l = 4$ y $l = 5$ y $l = 6$ forman un Perceptrón Multicapa (MLP) con tres capas, cuya entrada es la salida u_{ijk} de agrupación máxima de la segunda capa convolucional. El MLP está compuesto por 128 neuronas en la entrada, 150 neuronas en la capa intermedia y finalmente, la salida tiene 2 neuronas para definir la clasificación de una de las dos tareas presentadas en la entrada. Las neuronas del MLP tienen una función de activación ReLU y su número se definió empíricamente puesto que el objetivo de esta tesis no se centra en encontrar los parámetros óptimos para lograr un mayor desempeño en la clasificación.

2.4 Modelo de red neuronal sCapsNet

Las redes neuronales tipo cápsula fueron seleccionadas para los experimentos de esta tesis porque brindan una idea de los patrones que se detectan en la entrada, específicamente, este tipo de redes son invariantes a la traslación, escala y rotación en el procesamiento de imágenes. Esto podría aportar información a la detección de patrones de habla imaginada. Además, las CapsNet se consideran en clasificación de vocales imaginadas porque han mostrado buen desempeño en tareas de clasificación de señales EEG [107]–[109] y, además, son una mejora de las redes neuronales CNN [110] debido a la invarianza anteriormente mencionada. Para este experimento,

se muestra la arquitectura del modelo en la Figura 2.6, el cual se denomina CapsVI (Cápsulas para Vowel Imagery) compuesto por cuatro capas ($l=1, \dots, 4$). Dichas capas son la capa convolucional, capa de cápsulas de características, capa de cápsulas clasificadoras y la última capa es una red totalmente conectada. Esta arquitectura fue seleccionada debido a que se pretende estudiar las vocales, y bajo este esquema se reúne información para crear un conocimiento acerca de las características que son útiles para el reconocimiento de vocales.

La capa convolucional extrae patrones de la entrada, en este caso de las señales CSP de la base de datos de DaSalla. Tal y como se mencionó en la subsección 2.2.2, las cápsulas de características consisten en un agrupamiento de neuronas, en las que su salida representa la probabilidad de que una característica de las señales de EEG esté presente. En este caso, las propiedades son el objeto de interés para ser estudiado y encontrar los patrones que permiten el reconocimiento de vocales imaginadas. Las cápsulas clasificadoras se encargan de generar vectores que representan la probabilidad de que una entrada corresponda a una clase.

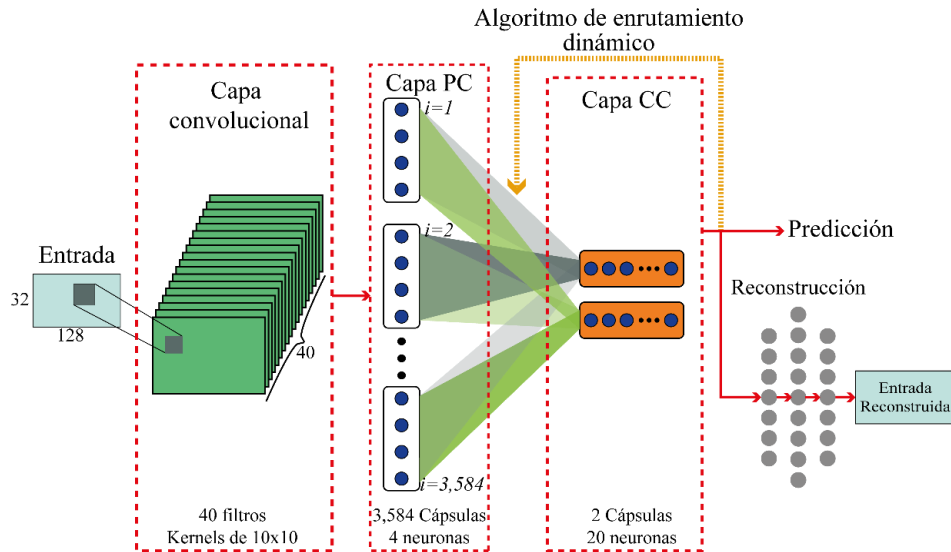


Figura 2.6 Modelo de red neuronal CapsVI.

La capa convolucional $l = 1$, es una capa convolucional tradicional dada por (2.1) que se utiliza para detectar características primarias en la entrada. En $l = 1$ se utiliza un conjunto de 40 filtros con una dimensión de 10x10. Este número y dimensiones se determinaron por medio de un análisis en los entrenamientos, en los que se observaban los mapas de características

generados por la operación convolucional, y, se observó que la mayoría de ellos eran similares, por lo que podemos concluir que la red se entrenaba para encontrar las mismas características.

La capa de cápsulas de características $l = 2$ tiene como función producir combinaciones de los patrones generados en la capa convolucional para generar características de alto nivel que representen entidades en la entrada. Esta capa está compuesta por 32 cápsulas que realizan la convolución con la función de activación *squashing*. PrimaryCaps tiene 256 filtros de 3×3 con un *stride* de 2, *padding* de cero. La tercera capa $l = 3$ se llama "ClassCap", y tiene dos cápsulas con un tamaño de 1×4 cada una. Formalmente, para todas las capas de cápsulas, la operación de salida viene dada por la ecuación 2.5 donde se genera la entrada s_j la j cápsula, la cual está dada por la suma ponderada de los vectores de predicción con los coeficientes de acoplamiento c_{ij} . Estos últimos son generados mediante el algoritmo de *iterative dynamic routing algorithm* [94] y son parámetros de entrenamiento tales como pesos. u_j es la salida de la cápsula de una capa anterior y W_{ij} son las matrices de pesos con la función de acoplar las salidas de capas anteriores y la entrada de cápsulas de capas superiores. Después del cálculo de la salida de cada cápsula, la función no lineal *squashing* de la ecuación 2.4 se utiliza para normalizar la magnitud de los vectores en valores cercanos a 0 y cercanos a 1. Esta función es utilizada porque en sCapsNet se utiliza el módulo del vector para representar la probabilidad de existencia de que una entidad exista en la entrada, de tal manera que esta función expresa sus valores basados en la magnitud en términos de probabilidad con valores entre 0 y 1.

El decodificador está compuesto por tres capas totalmente conectadas con 1024, 2048 y 4096 neuronas respectivamente. Estas capas tienen como función reconstruir la señal EEG basada en la salida de la capa ClassCaps para ser utilizada como una función de regularización.

2.5 Experimentos y resultados

Los experimentos fueron desarrollados en una computadora *Dell Precision Tower 5810* con GPU NVIDIA 970x, procesador Intel E5-1603 y una memoria RAM de 8 Gb de 2133 Mhz. En cuanto a los entrenamientos de los modelos, sCNN fue entrenado con el algoritmo gradiente descendente estocástico [111], 25 épocas, con un tamaño de lote de uno y un coeficiente de aprendizaje de 0.0025. sCapsNet se entrenó mediante el algoritmo de enrutamiento dinámico

entre cápsulas [94] con 20 épocas, un tamaño de lote de uno, y $\alpha=0.0005$ para la función de pérdida. La base de datos DaSalla fue particionada para las etapas de entrenamiento, validación y prueba en un 60%, 20% y 20% respectivamente. Se realizó validación cruzada *k-fold* [112], [113] con cinco validaciones en ambos modelos propuestos. Asimismo, cada modelo se entrenó por sujeto para cada clasificación.

Para las tareas de clasificación es importante recordar y aclarar las clases que brinda la base de datos. DaSalla utiliza clasificaciones entre vocales y un estado de control, de tal manera que 3 clases serán definidas: /a/:control, /u/:control y /a:/u/. Las primeras de ellas diferencian entre el habla imaginada de las vocales “a” y “u” un estado de control, la segunda hace distinción entre vocales. Estas clasificaciones binarias están dadas debido a que en DaSalla las señales EEG disponibles están procesadas mediante CSP de dos clases, y por ende se realiza la proyección de las señales EEG en matrices de proyección por medio de filtros espaciales que maximizan la varianza entre 2 clases.

2.5.1 Resultados de reconocimiento de sCNN y CapsNet

Las Tablas 2.1 y 2.2 muestran el desempeño de los modelos propuestos en la clasificación para la etapa de validación. De acuerdo con la Tabla 2.1, sCNN alcanza un mejor promedio de desempeño con el sujeto 1, debido a que este modelo presenta el mejor desempeño en las clases /a/:control y /u/:control con este mismo sujeto. Adicionalmente, este modelo tiene los mejores resultados para el sujeto 3 y la clase /a:/u/. Mientras que sCapsNet obtiene los mejores resultados en la clase /a:/u/ con todos los sujetos con un 89.3% de exactitud.

Tabla 2.1 Resultados de clasificación de sCNN.

sCNN	/a/:control	/u/:control	/a:/u/	Promedio
Sujeto 1	78%	76%	66%	73.3%
Sujeto 2	64%	74%	59%	65.6%
Sujeto 3	58%	60%	74%	64%
Promedio	66.6%	70%	66.3%	67.63%

Tabla 2.2 Resultados de clasificación de sCapsNet.

sCapsnet	/a/:control	/u/:control	/a/:/u/	Promedio
Sujeto 1	61.47%	65%	89%	71.8%
Sujeto 2	60%	66%	91%	72.3%
Sujeto 3	64%	63%	88%	71.6%
Promedio	61.82%	64.6%	89.3%	71.93%

El pico SRP para la vocal /a/ se presenta alrededor de los 379 ms después de la tarea de imaginación y 355 ms para /u/ y en el estado de control no se genera ningún pico [42]. De esta manera, sCNN determina la presencia de SRP's por medio de operaciones convolucionales y predice si existe una vocal en la señal EEG. sCapsNet modela la presencia de entidades por medio de cápsulas que son vectores. De acuerdo con los experimentos, se encontró que el número de neuronas en las cápsulas podrá ser asociado con el modelado de entidades presentes en la entrada. En el caso de la imaginación de /a/ o /u/, sCapsNet pudo claramente modelar y clasificar las vocales utilizando 2 cápsulas con 4 neuronas cada una. El número de cápsulas fue definido porque se implementaron 2 clases, mientras que el número de neuronas se definió empíricamente de acuerdo con los resultados. Sin embargo, en el caso de la comparación de estado de control, las señales no cuentan con una estructura a comparación de las señales de EEG generadas al imaginar vocales, por lo que las señales de EEG del estado de control no tienen definido un patrón para ser modelado por medio de vectores de actividad en cápsulas. Esto se produce el resultado de la Tabla 2.2 donde el promedio de desempeño para /a/:/u/ es mayor que las vocales contra control.

2.5.2 Comparación con otros trabajos

En la elaboración de la base de datos DaSalla [42] se utilizó un SVM no lineal con un *kernel* de base radial para el reconocimiento de vocales. Las comparaciones entre modelos se llevaron a cabo en términos de desempeño, dado que es la métrica utilizada en esta base de datos para reportar resultados. De acuerdo con la Tabla 2.3, sCapsNet obtuvo los mejores resultados. Adicionalmente, las clasificaciones con sCapsNet entre los tres sujetos reflejan resultados similares, mientras que con SVM presentan un comportamiento sobresaliente solo en el caso del sujeto 1. Este comportamiento podrá ser atribuido a que las cápsulas modelan los patrones

de SRP representados por el vector de actividad de sus neuronas. Estas características pueden ser representadas sin importar su tiempo y rotación en la que se presenten. En otras palabras, las cápsulas pueden modelar clases basándose solamente en sus entidades. Por otra parte, un SVM simple utiliza un hiperplano para separar dos conjuntos de datos, y dado el hecho de que este clasificador utiliza las diferencias y varianza entre clases, SVM obtiene mejores resultados en la clasificación de vocales vs estado de control, ya que estas dos clases son las que presentan mayor varianza entre clases, a diferencia de la clase /a:/u/ donde la varianza entre clases es menor que en las otras clases. Este comportamiento es contrario a los resultados de sCapsNet, lo cual se muestra en la Tabla 2.3.

Tabla 2.3 Comparación entre modelos con sCNN y sCapsNet.

	sCNN	sCapsNet	LDA [114]	SVM [42]
Sujeto 1	73.3%	71.8%	78.33%	78%
Sujeto 2	65.6%	72.3%	75%	68%
Sujeto 3	64%	71.6%	72.5%	68%
Promedio	67.63%	71.9%	75.27%	71.33%

Como se puede observar en la Tabla 2.3, LDA y SVM tienen los mejores resultados en el caso del sujeto 1, lo cual se puede atribuir a que los métodos fueron diseñados considerando las señales del sujeto 1, de tal forma que se puede considerar una especialización del modelo propuesto para este sujeto. Por otra parte, sCapsNet fue entrenada considerando las señales de todos los sujetos, que finalmente obtiene menor varianza entre el reconocimiento de los sujetos.

Es importante destacar que en el caso de sCapsNet, se obtuvieron resultados en la validación cruzada de hasta un 91% de desempeño, mientras que sCNN y SVM obtienen resultados por debajo del 82%.

CAPÍTULO 3. PROPUESTA DE MODELO BASADO EN CAPSNET

En este capítulo se describe el modelo propuesto para el análisis de señales EEG con habla imaginada para el reconocimiento de vocales y posteriormente, utilizarlo para reconocer algunos fonemas y palabras. Este modelo tiene como base los conocimientos adquiridos en el reconocimiento de vocales mediante el análisis de los SRP. Por ello, el objetivo de este capítulo es describir el modelo que permite encontrar patrones a posteriori que puedan ser utilizados como información a priori en el reconocimiento de palabras y fonemas en el idioma inglés.

En la sección 3.1 se presenta la arquitectura del modelo propuesto y en la sección 3.2 su funcionamiento y modelo de aprendizaje.

3.1 Arquitectura del modelo propuesto

La arquitectura del modelo propuesto se divide en las siguientes capas: convolución, cápsulas de características, cápsulas clasificadoras y una red totalmente conectada. La capa de convolución extrae diferentes patrones de características en la entrada. En cuanto a las capas de cápsulas, cada cápsula representa una característica de la información en la entrada, lo que permite que la red pueda aprender componentes que conforman la entrada y por medio de la medición de la relación entre cómo se presentan las características en la entrada, la red encuentra características invariantes al ruido, posición en el tiempo, orientación y rotación. La capa de cápsulas clasificadoras es entrenada para generar vectores de patrones prototipo que representan las clases.

Una de las ventajas que tienen las redes cápsula es que tienen la habilidad de generalización en bases de datos con pocas muestras, a diferencia de otras redes neuronales como CNN o Deep Belief Networks [94], [115]. Consecuentemente, en relación con el tiempo de entrenamiento, las redes cápsula necesitan menor tiempo de entrenamiento que otros modelos de redes neuronales como CNN.

Otra de las ventajas de las redes neuronales tipo cápsula está relacionada con el hecho de que las capas convolucionales en CNN funcionan como filtros que encuentran características en

diferentes niveles de resolución, mientras que las cápsulas encuentran patrones que describen la relación entre características en la entrada.

El primero de los modelos de redes neuronales tipo cápsula utilizó un algoritmo de entrenamiento llamado “enrutamiento por acuerdo”, el cual envía información entre cápsulas de diferentes capas y es el algoritmo más popular en la literatura para implementar el concepto de cápsula. El enrutamiento por acuerdo es el reemplazo de la operación de pooling en CNN porque se elimina la operación de reducción de tamaño en los mapas de convolución y se realiza un agrupamiento en los escalares de los mapas para formar las cápsulas. Además de éste, existen otras variaciones para implementar el proceso de envío de información entre cápsulas propuestos por Hinton et al. [99]. El primero de ellos se denomina *transforming auto-encoders*, y el segundo se denomina *expectation maximization routings*.

En la Figura 3.1 se muestra la arquitectura del modelo propuesto, esencialmente es similar al modelo utilizado en el capítulo 2 de esta tesis. Sin embargo, para lograr el reconocimiento de vocales imaginadas existen diferencias en:

1. Cambio del número de neuronas de cada capa para las cápsulas.
2. Diferente tamaño de kernel en las capas convolucionales.
3. Análisis del optimizador para la reconstrucción de la entrada.

Las redes neuronales tipo cápsula fueron consideradas en el problema de habla imaginada porque estos modelos reportan buen desempeño en problemas relacionados con señales de EEG [107], [108], y, además son una mejora de las redes neuronales CNN ya que muestran un mejor desempeño al lograr una generalización de datos mayor en bases de datos más pequeñas.

Al modelo propuesto se le denomina CapsVI y está compuesto por una capa convolucional, una capa de cápsulas primarias (PC), una capa clasificadora denominada “*Class caps*” (CC) y 3 capas totalmente conectadas que sirven como regularizador en el entrenamiento de la red al reconstruir la matriz de entrada a partir del vector de salida de la capa CC. Cada capa será descrita en las siguientes subsecciones.

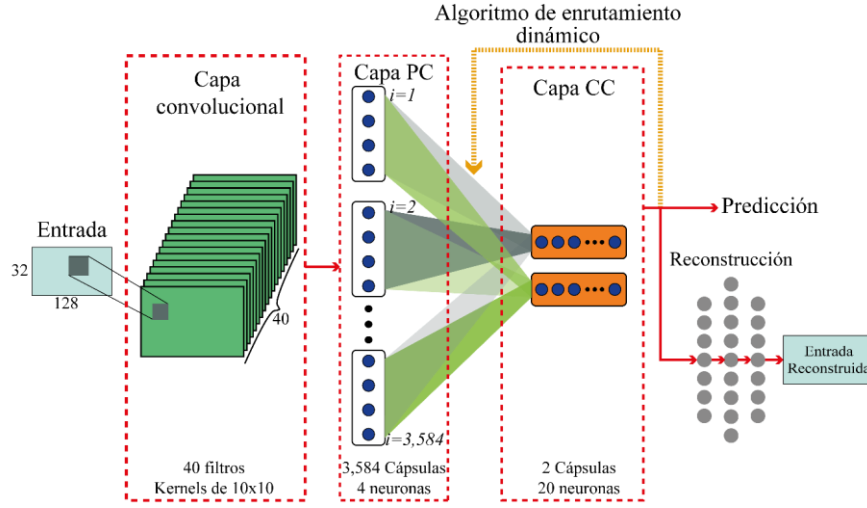


Figura 3.1 Arquitectura del modelo propuesto CapsVI.

3.2 Capa de entrada y capa convolucional

La entrada de la red está compuesta por señales EEG en forma de una matriz de $n \times m$ píxeles, a esta entrada se le denominará X_t y alimenta a la capa convolucional compuesta por 40 filtros compuestos por *kernels* de 10×10 , sin *padding* y con un *stride* de 1. La operación que se realiza en esta capa está dada por la siguiente ecuación:

$$c_t(n, m) = f(W_\tau(n, m) * X_t(n, m) + \beta_t) \quad 3.1$$

donde se convoluciona la entrada X_t con W_τ que son los *kernels* y β_t es la polarización, mientras que f es la función de activación. t corresponde al índice del número de señal y τ es la profundidad de los *kernels*. En este caso se utiliza la unidad de rectificación lineal ReLU como función de activación f . En esta capa se extraen mapas c_t que contienen características abstractas a través de la convolución para ser alimentadas en la siguiente capa y formar las cápsulas por medio de la agrupación de neuronas.

3.2.1 Capa de cápsulas primarias (PC)

Para la creación de la primera capa de cápsulas se utiliza una capa convolucional de cápsulas. Se le denomina así debido a que esta capa agrupa valores escalares resultado de la convolución entre los mapas de características de la primera capa convolucional y los *kernels* de esta nueva capa, para formar cápsulas. El proceso de convolución está dado por la ecuación 3.1 a diferencia

que los filtros son $W_\tau = 32$ de dimensión 9×9 y se convolucionan con c_t de la capa anterior con *strides* de dos y sin *padding*. Luego de calcular la convolución, se realiza un reacomodo de los mapas de características $c_t(n,m) \in \mathbb{R}^{4 \times 3584}$, lo que significa que la capa PC tiene 3585 cápsulas con 4 neuronas cada una. El reacomodo se define como:

$$\phi_i, i = 1, \dots, 1584, v = 1, \dots, 4 \quad 3.2$$

donde ϕ son las cápsulas e i su índice, v es el índice de las neuronas en cada cápsula.

La salida de cada cápsula es rectificada con una función de activación no lineal de tipo *squashing* utilizada para normalizar la magnitud de los vectores en valores cercanos a 0 y 1. Esta función de activación está definida por la Ecuación 2.4. Donde u_i es la salida de la cápsula i en la capa PC y mi es el vector de las cápsulas que se va a rectificar. Esta función es utilizada porque el vector de salida de las cápsulas representa la probabilidad de existencia de una entidad presente en la entrada. La salida u_i es denominada vector de actividad, el cual codifica la información de la entrada en forma de parámetros de instanciación.

3.2.2 Capa clasificadora CC

Tal y como se mencionó en la subsección anterior, cada cápsula i de la capa PC tiene un vector de salida u_i , el cual es alimentado a todas las cápsulas j de la capa CC, de tal forma que cada cápsula j recibirá a u_i y calculará el producto entre esta entrada y una matriz de pesos ϖ_{ij} mediante la siguiente ecuación:

$$\hat{u}_{j|i} = \varpi_{ij} u_i \quad 3.3$$

El vector $\hat{u}_{j|i}$ resultante es denominado vector de predicción y es interpretado como la transformación de la entidad representada por la cápsula i por la cápsula j de la capa CC. El vector de predicción representa la contribución de la cápsula i a la clase representada por la cápsula j .

El siguiente paso en la capa CC está dado por el cálculo entre $\hat{u}_{ji}$ y los coeficientes de acoplamiento c_{ij} en cada una de las cápsulas j para obtener el “acuerdo” entre cápsulas de las capas PC y CC. Por ejemplo, si la cápsula $i=1$ está relacionada con la cápsula $j=2$, el acuerdo entre ellas será alto y significa que lo que representan esas cápsulas son relevantes entre ellas. Consecuentemente, los coeficientes c_{ij} serán maximizados, de otra forma si no tienen relación los coeficientes serán minimizados.

Tal y como puede verse en la Figura 3.1, los coeficientes son representados mediante las líneas de color verde y gris. Las líneas de color verde con más color son aquellos coeficientes maximizados de las cápsulas 1 y 3584 relacionadas con la cápsula $j=2$ de la capa CC. Por otra parte, las líneas de color gris de la cápsula 2 representan un alto acuerdo con la cápsula $j=1$ de la capa CC.

Los coeficientes c_{ij} son determinados por el algoritmo de enrutamiento dinámico entre cápsulas, el cual es la base de funcionamiento de este tipo de redes neuronales. Una suma ponderada s_j de todos los vectores de predicción para todas las cápsulas j en la capa CC es calculada mediante:

$$s_j = \sum_i c_{ij} \hat{u}_{ji} \quad 3.4$$

Donde s_j es la salida de la cápsula j en la capa CC. Por último, esta salida es rectificada mediante la función *squashing* para obtener la salida v_j de la capa CC. Lo anterior se realiza mediante la ecuación 2.4.

En párrafos anteriores se ha hablado de los coeficientes de acoplamiento c_{ij} pero no se ha descrito cómo se obtienen. Estos coeficientes se utilizan para asegurar que la predicción de la cápsula i en la capa PC está relacionada con la cápsula j de la capa CC. Los coeficientes son dados mediante la ecuación:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_t \exp(b_{it})} \quad 3.5$$

El un enrutamiento dinámico hace varias iteraciones donde los coeficientes se apoyan de una variable auxiliar b_{ij} que es modificado en cada iteración utilizando el producto punto entre vectores de predicción $\hat{u}_{ji}$ y la salida de la cápsula v_j . El producto punto es la distancia euclidiana de dos vectores y el coseno del ángulo entre sus proyecciones. En otras palabras, lo que se mide es la proyección de un vector sobre otro, entre más proyección se tenga, mayor será el acuerdo entre cápsulas.

Ya que la función *squashing* expresa valores de probabilidad entre 0 y 1 en el vector de actividad, si y solo si una entidad está presente en la entrada, las cápsulas para la clase t tendrán un vector de instanciación maximizado. Para permitir el reconocimiento de diferentes clases, se utiliza una función de margen de error separado, definido para cada cápsula expresado por la siguiente ecuación:

$$M_t = T_t \max(0, m^+ - \|v_t\|)^2 + \lambda(1 - T_t) \max(0, \|v_t\| - m^-)^2 \quad 3.6$$

Donde $T_t=1$ si y solo si la clase t está presente (objetivo o etiqueta) y los valores de m^+ y m^- son los márgenes de error que se tendrán siendo valores entre 0.1 y 0.9. El factor de escalamiento λ es un coeficiente para detener el aprendizaje en los casos de clases ausentes. Este valor está definido por 0.5. La magnitud de los vectores de salida de la capa clasificadora define la clasificación del modelo.

El interés del modelo en este capítulo son los vectores de salida s_j , los cuales contienen los patrones de reconocimiento codificados para una clasificación binaria. Mediante una capa de reconstrucción se estudiarán y decodificarán los patrones para obtener información acerca de la imaginación de palabras y fonemas.

3.2.3 Capa de reconstrucción

Esta capa está compuesta por un decodificador de 3 capas totalmente conectadas con 1024, 2048 y 4096 con el objetivo de reconstruir el arreglo de entrada a partir de los patrones prototipo aprendidos en la capa CC, y encontrar los valores de los *kernels* W_{tp} a través de los

optimizadores Adam y *backpropagation*. La reconstrucción se utiliza solamente durante el entrenamiento y está basado en el modelo clásico de CapsNet.

De tal forma que las capas totalmente conectadas reconstruyen la señal de EEG pocesada con CSP de las cápsulas clasificadoras v_t . El entrenamiento del modelo comienza al reducir la suma de cuadrados entre $Y_t(n,m)$, y la entrada de la red $X_t(n,m)$, lo cual está dado por la siguiente ecuación:

$$L_t = \alpha \sqrt{\frac{1}{N} \sum_{t=1}^N (Y_t(n,m) - X_t(n,m))^2} \quad 3.7$$

donde L_t es el error cuadrático medio, α es un factor de escalamiento que asegura que el margen de error en las predicciones domine el entrenamiento. El error total ι_t es obtenido mediante la suma del margen de error Mg_t y el error del regularizador L_t , definido por la siguiente ecuación:

$$\iota_t = Mg_t + L_t \quad 3.8$$

En la Figura 3.2 se puede observar el grafo del modelo propuesto para mejorar su comprensión en el proceso de entrenamiento, y de forma resumida se presenta la información presentada en las subsecciones 3.1.1-3.1.4.

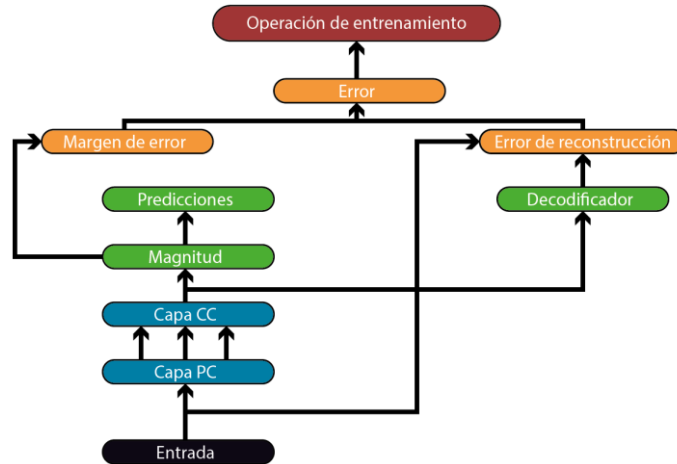


Figura 3.2 Grafico del modelo de entrenamiento para CapsVI.

3.3 Resultados de clasificación

Para la clasificación se utiliza una clasificación binaria, teniendo los siguientes tres grupos de clasificaciones: /a/:control, /u/:control y /a/:u/ para cada uno de los sujetos. El optimizador utilizado es el tipo “Adaptive Moment Estimation” (*Adam*), el cual calcula factores de aprendizaje adaptivos para cada parámetro [116].

En los experimentos se aplica una prueba de validación cruzada con cinco validaciones, 10 épocas en cada validación, cada una con un lote de 1. Para darle validez estadística a los experimentos, se utilizó la validación cruzada, implementada a través del kit de *sklearn* [112]. La división de la base de datos para los conjuntos de entrenamiento, prueba y validación fue de 56%, 20% y 24% respectivamente. En las Tablas 3.1, 3.2 y 3.3 se muestran los resultados de cada validación (CV) con los datos de cada uno de los sujetos.

Tabla 3.1 Resultados de desempeño en la validación cruzada del sujeto 1 con CapsVI.

Validación	/a/:control	/u/:control	/a/:u/
CV1	86.95%	79.16%	87.50%
CV2	91.66%	95.83%	100.00%
CV3	100.00%	100.00%	91.66%
CV4	100.00%	87.50%	100.00%
CV5	100.00%	100.00%	100.00%
Promedio	95.72% $\pm$ 6	92.50% $\pm$ 9	95.83% $\pm$ 6

Tabla 3.2 Resultados de desempeño en la validación cruzada del sujeto 2 con CapsVI.

Validación	/a/:control	/u/:control	/a/:u/
CV1	70.83%	54.16%	95.83%
CV2	91.66%	75.00%	100.00%
CV3	100.00%	91.66%	95.83%
CV4	100.00%	100.00%	100.00%
CV5	91.66%	100.00%	100.00%
Promedio	90.83% $\pm$ 14	84.16% $\pm$ 20	98.33% $\pm$ 2

Tabla 3.3 Resultados de desempeño en la validación cruzada del sujeto 3 con CapsVI.

Validación	/a/:control	/u/:control	/a/:u/
CV1	95.83%	88.33%	95.83%
CV2	95.83%	95.83%	100.00%
CV3	100.00%	75.00%	91.66%
CV4	95.83%	87.50%	100.00%
CV5	95.83%	100.00%	100.00%
Promedio	96.66% $\pm$ 2	99.33% $\pm$ 6	97.50% $\pm$ 4

El mejor resultado promedio por sujeto es un 94.68% $\pm$ 2 del sujeto 1 para todas las clases. Por otra parte, los resultados de clasificación por clase, el mejor de ellos fue un 99.33% $\pm$ 6. Con estos resultados se observa que el modelo genera mejores resultados cuando la imaginación en la clasificación corresponde a comparación de dos vocales.

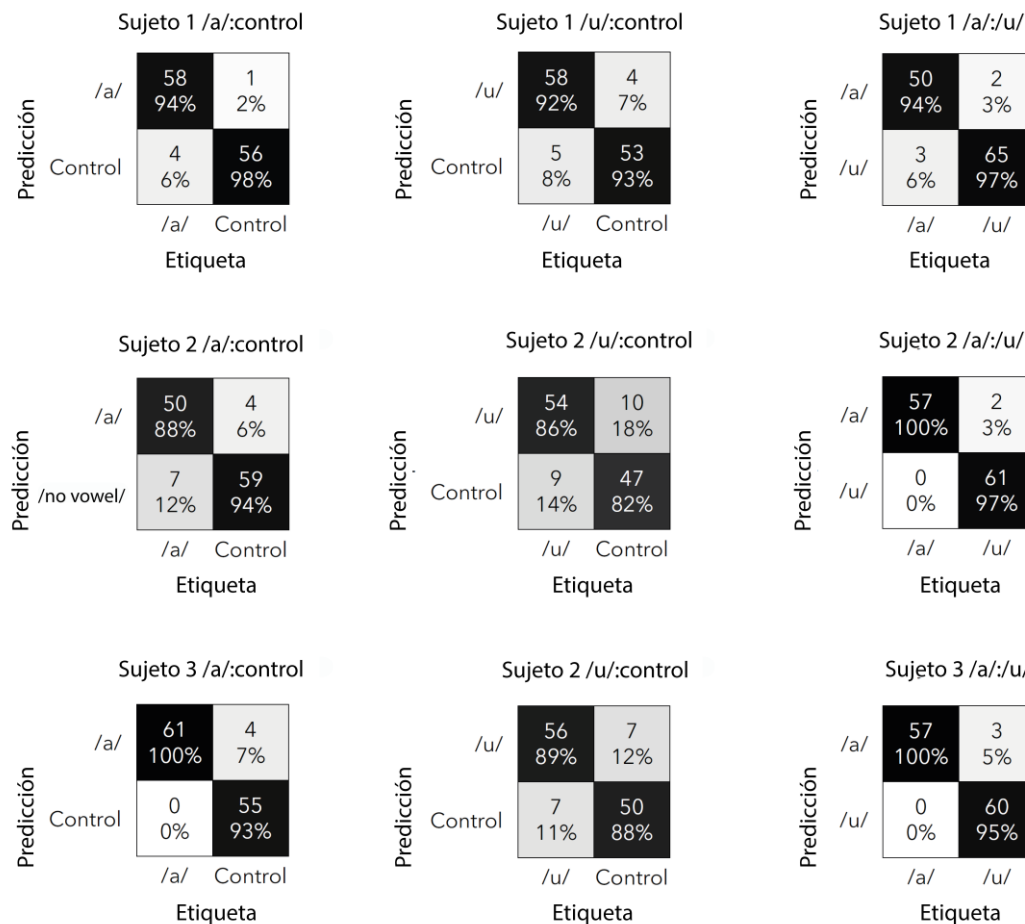


Figura 3.3 Matrices de confusión para la clasificación de vocales imaginadas.

La Figura 3.3 muestra las matrices de confusión por sujeto. En el caso del sujeto 1 se generan 19 falsos positivos y negativos en total, y los mayores errores ocurren cuando las clases /a/ y /u/ son clasificadas. En el caso del sujeto 2 se generan 32 falsos positivos y negativos, siendo la clasificación /u/:control la clase en la que más falsos se generan. En el caso del sujeto 3 se reportan 21 falsos positivos y negativos, donde se generan con mayor frecuencia en la clasificación de /u/:control. Los resultados muestran que la clasificación de la clase /a:/u/ es más invariante a la habilidad de los sujetos al imaginar vocales que al relajarse en la clase de control.

La Figura 3.3 muestra las matrices de confusión para CapsVI, donde en el caso de la clasificación del sujeto 1, se generan 19 errores en total, y los mayores errores de clasificación se generan cuando se realiza la clasificación /a/ y /u/ contra el estado de control. Para el sujeto 2 se generan 32 errores en total, y la mayor confusión se presenta cuando se clasifica /u/:control. Las matrices de confusión para el sujeto 3 reporta 21 errores de clasificación, siendo los mayores en el caso de la clasificación /u/:control. Todos los sujetos obtienen los mejores resultados de clasificación en la clase /a:/u/, y en el caso del sujeto 1 se obtiene el mejor desempeño promedio en todas las clases. Los resultados en las matrices de confusión demuestran que la clasificación de la clase /a:/u/ es más invariante a la habilidad de los sujetos de imaginar vocales que la distinción en /a/:control y /u/:control.

3.3.1 Comparación con otros trabajos

En esta subsección se presenta la comparación de desempeños CapsVI con modelos empleados para la clasificación de la base de datos de DaSalla. Asimismo, se incluyen los modelos desarrollados en el capítulo 2. Los modelos incluyen un SVM implementado por DaSalla et al. [42], una red neuronal convolucional y una red de tipo cápsula desarrollada por Macías et al. [64] y el último modelo basado en LDA fue desarrollado por Idrees et al. [114]. En esta comparación, se utilizan los modelos mencionados de acuerdo con la información presentada en la literatura. Es importante mencionar, que estos modelos solamente incluyen resultados de exactitud, por lo que la comparación está de acuerdo con este esquema.

Tabla 3.4 comparación de resultados con otros trabajos de la literatura.

	sCNN[64]	sCapsNet[64]	SVM [42]	LDA [114]	CapsVI
Sujeto 1	73.3%	71.8%	78%±5	78.33%	94.68%±2
Sujeto 2	65.6%	72.3%	68%±7	75%	91.11%±6
Sujeto 3	64%	71.6%	68%±12	72.5%	94.16%±4
Promedio	67.63%	71.93%	71.33%	75.27%	93.32%±4

La métrica utilizada en DaSalla es la métrica de exactitud. De acuerdo con la Tabla 3.4, CapsVI obtiene los mejores resultados en el promedio de cada uno de los sujetos y el promedio de todos los sujetos. Adicional a esto, todos los modelos obtienen sus mejores resultados con el sujeto 1, lo cual puede ser atribuido a que las señales de EEG de este sujeto tienen más separación de clases entre ellas, gracias a ello, varios modelos reportan los mejores resultados con este sujeto.

La razón de que CapsVI obtuvo los mejores resultados se puede deber a que este tipo de red neuronal puede identificar características presentes en la entrada entre cada intento de actividad EEG que se generó en la base de datos, y son representadas por grupos de neuronas con vectores de actividad, las cuales son invariables al tiempo en el que se presenten las características a lo largo de la señal EEG. Aunado a esto, esta red tiene ventajas sobre los otros modelos, ya que el tipo de *kernels* que utiliza está optimizado para las señales que se presentan en este caso.

3.3.2 Análisis de los *kernels* en la capa convolucional

En esta subsección se presenta un análisis del efecto de los *kernels* sobre las imágenes de entrada mediante la convolución. Consecuentemente los valores de los *kernels* son analizados para determinar el tipo de características que se obtienen en esta capa. En la Figura 3.4 se muestra un ejemplo de las matrices de entrada para el habla imaginada de la vocal /a/ y el estado de control.

Como se puede observar en la Figura 3.4, los patrones generados para cada caso de habla imaginada son diferentes, sin embargo, este patrón no se repite en la base de datos. Sin embargo, los desempeños de clasificación muestran que el modelo de red neuronal es capaz de obtener información que caracteriza a las señales de diferentes clases de la base de datos. La pregunta

en este punto es cuáles son esas características, por lo que se propone un análisis de los valores de los *kernels* y la imagen de convolución resultante. En la Figura 3.5 se muestran algunos ejemplos de las imágenes resultantes en esta capa para el habla imaginada de la vocal /a/.

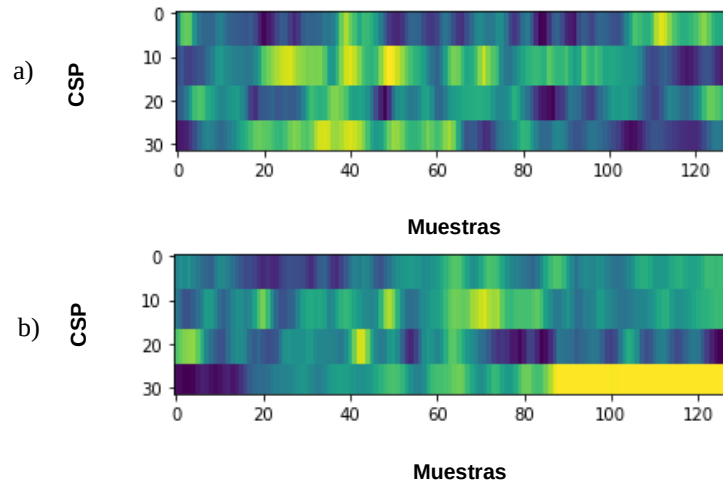


Figura 3.4 Señales de entrada al modelo CapsVI. a) Habla imaginada de /a/. b) Estado de control.

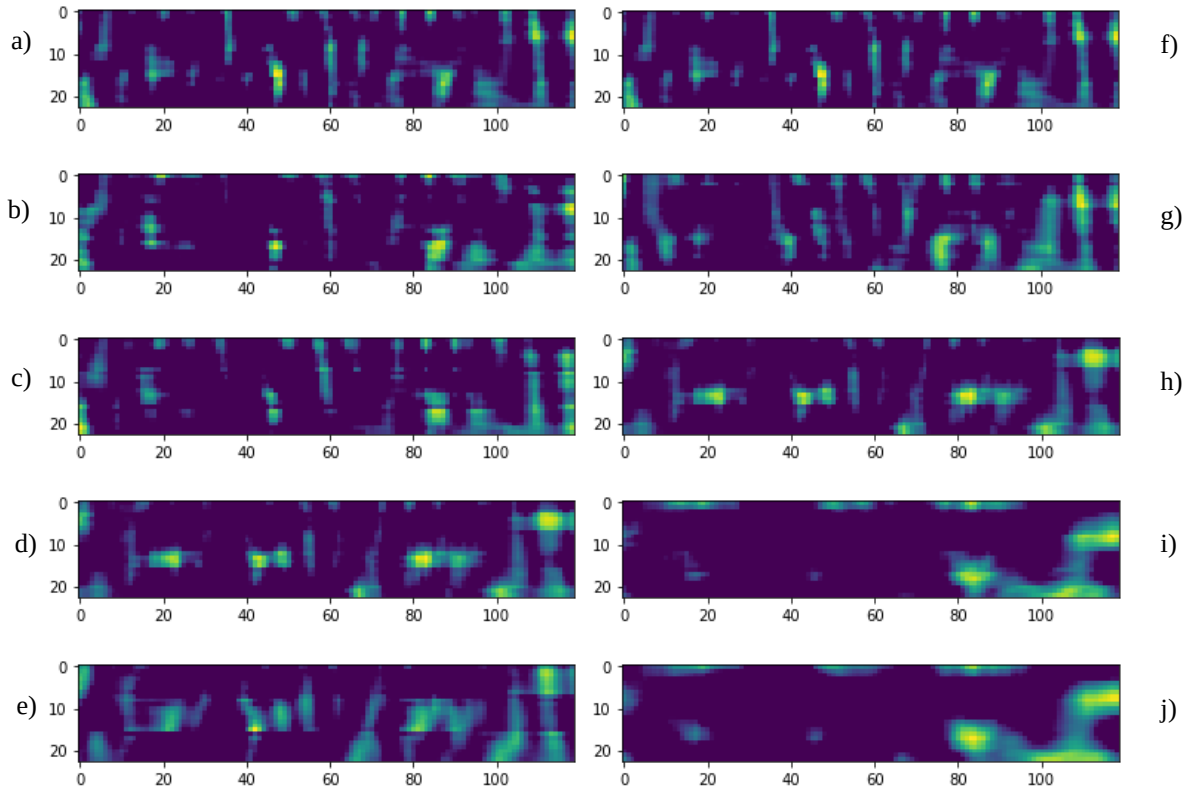


Figura 3.5 Resultado de la convolución para el habla imaginada de la vocal /a/.

Se encontró que la convolución en esta capa extrae como características los valores de la señal en cuanto a profundidad. En otras palabras, los *kernels* son sensibles a los picos, valles y diferenciales positivo y negativo de la señal. Sin embargo, bajo un análisis visual de las imágenes, por ejemplo, las imágenes i) y j) de la Figura 3.5 tienen un grado de similitud entre ellas, las cuales fueron relacionadas a los valles en la señal original. Por esta razón, se propone un análisis de los valores en los *kernels* para determinar si algunos de ellos están aprendiendo las mismas características que otros con base en su similitud. Para ello se obtienen los valores de los *kernels* y se mide el grado de similitud con una técnica de agrupamiento llamada “*Spectral clustering*”, la cual se basa en medir la distancia euclidiana de dos matrices y con ello realizar agrupaciones. Después de la agrupación se determinó que existían 12 grupos de *kernels* similares, para luego modificar el modelo de red neuronal de 40 a 12 *kernels* en la capa convolucional. Los resultados de exactitud de este proceso se muestran en las Tablas 3.5, 3.6, 3.7 y 3.8.

Tabla 3.5 Desempeño promedio de validación cruzada con 12 *kernels* para el sujeto 1.

Sujeto 1						
Validación #	1	2	3	4	5	Promedio
Clase 1	94.44%	92.59%	96.29%	96.29%	92.59%	94.44%±2
Clase 2	92.85%	94.64%	94.64%	91.07%	91.07%	92.85%±2
Clase 3	96.36%	94.54%	96.36%	96.42%	96.42%	96.02%±1
Promedio	94.55%	93.92%	95.76%	94.59%	93.36%	94.44%±1

Tabla 3.6 Desempeño promedio de validación cruzada con 12 *kernels* para el sujeto 2.

Sujeto 2						
Validación #	1	2	3	4	5	Promedio
Clase 1	80.35%	85.71%	87.50%	91.07%	89.28%	86.78%±4
Clase 2	69.64%	87.50%	82.14%	83.92%	87.50%	82.14%±7
Clase 3	85.71%	83.92%	85.71%	82.14%	91.07%	85.71%±3
Promedio	78.57%	85.71%	85.12%	85.71%	89.28%	84.88%±4

Tabla 3.7 Desempeño promedio de validación cruzada con 12 *kernels* para el sujeto 3.

Sujeto 3						
Validación #	1	2	3	4	5	Promedio
Clase 1	89.28%	91.07%	92.85%	94.64%	91.07%	91.78%±2
Clase 2	87.50%	85.71%	83.92%	83.92%	82.14%	84.64%±2

Clase 3	87.50%	94.64%	91.07%	94.64%	91.07%	91.78%±3
Promedio	88.09%	90.47%	89.28%	91.07%	88.09%	89.40%±1

Tabla 3.8 Desempeño promedio de validación cruzada con 12 *kernels* para todos los sujetos.

	/a/:control	/u/:control	/a/:u/	Promedio
Sujeto 1	94.44%±2	92.85%±2	96.02%±1	94.44%±1
Sujeto 2	86.78%±4	82.14%±7	85.71%±3	84.88%±2
Sujeto 3	91.78%±2	84.64%±2	91.78%±3	89.40%±3
Promedio	91%±3	86.54%±4	91.17%±3	89.57%±2

Tal y como se puede observar los resultados muestran una mejora en el aprendizaje al determinar que existían 12 grupos con *kernels* similares. Esto se puede comparar con los resultados de sCNN. La razón de esto es que ahora el modelo de red neuronal está aprendiendo justo lo que necesita aprender para poder clasificar las señales de habla imaginada para los 3 sujetos en el caso de las 3 clases presentes.

CAPITULO 4. RESULTADOS DE RECONOCIMIENTO CON EL MODELO PROPUESTO

En este capítulo se describe el análisis de la base de datos “KaraOne”, el cual tiene como objetivo poder llegar a una caracterización de las señales de habla imaginada de cada sujeto. Para ello se utiliza una metodología diseñada a partir del análisis de varianzas ANOVA [117], una selección de canales y el modelo CapsVI con adaptaciones para la clasificación de ausencia/presencia de palabras y fonemas, el cual se denomina CapsK-SI [118].

El reconocimiento de señales de SI, específicamente fonemas y palabras, es una estrategia mental que ha sido explorada menos que otras estrategias populares como potenciales evocados y imaginiería motora puesto que su estudio es más reciente. Además, las señales de SI son más complejas que aquellas que contienen potenciales evocados e imaginiería motora [58]. El objetivo de esta sección es estudiar y analizar características estadísticas y propiedades de señales EEG de SI para lograr la identificación de fonemas y palabras imaginadas. Con base en este análisis, se propone un método de clasificación de actividad cerebral derivado de la categorización de fonemas y palabras de acuerdo con las diferencias fonológicas entre ellos.

En la sección 4.1 se detalla la composición de la base de datos, en la sección 4.2 la selección de canales, en la sección 4.3 se detalla el procesamiento y selección de características basado en ANOVA, y, finalmente en la sección 4.4 se describe al modelo CapsK-SI.

4.1 Base de datos

KaraONE es una base de datos creada por Shunan Zhao et al. [4] y está compuesta por señales de 3 modalidades: EEG, movimientos faciales y audio. Estas señales son grabadas mientras se realiza un protocolo del habla.

Esta base de daos consta de señales de 12 sujetos de prueba con edad media de 27.4 años (4 mujeres y 8 hombres). Para el caso de EEG, se generan 132 señales por sujeto con los fonemas de /iy/, /uw/, /piy/, /tiy/, /m/, /m/ y /n/, más cuatro palabras derivadas de la lista de pares fonemas-similares de la lista de Kent (/pat/, /pot/, /knew/, /gnaw/) [8]. Para adquirir las señales, el sujeto debe permanecer en una silla cómoda donde se le coloca el dispositivo Quick-cap de Neuroscan

y se inicializa el dispositivo Microsoft Kinect para las señales de audio y video. Luego de ello, el sujeto deberá fijar la mirada en un monitor y producir la menor cantidad de movimientos musculares tal cual sea posible, mientras se muestran los comandos de los fonemas a lo largo de 30-40 minutos.

El protocolo de adquisición de señales se compone de cuatro etapas:

1. El sujeto permanece en un estado de reposo durante cinco segundos, tiempo en el que deberá de relajarse y despejar su mente.
2. El “estado de estimulación” es la etapa mediante la cual aparece el comando en forma de texto con el fonema/palabra que deberá de producir, con una duración de dos segundos para permitir que el sujeto mueva u acomode sus articulaciones en caso de haberse movido.
3. Posteriormente a la etapa dos, se realiza la imaginación del fonema a producir sin realizar sonidos o movimientos de articulaciones durante cinco segundos. Esta etapa es la encargada de realizar el habla imaginada.
4. Finalmente, se realiza el habla articulada del fonema y con esto termina el protocolo de cada sílaba o palabra imaginada.

Al final de estas cuatro etapas, se termina el protocolo de grabación. Cada sujeto realiza 132 protocolos. Cada 40 protocolos, el sujeto tiene un tiempo de descanso. De acuerdo con las diferencias fonológicas entre cada comando (fonema y palabra) es posible definir grupos de clasificaciones binarias. Las diferencias entre los comandos se ubican en el lugar y la manera de articulación en la producción del sonido en el habla articulada. El lugar de articulación se refiere a dónde las articulaciones de la boca se colocan para producir un sonido, por ejemplo, el contacto de los dos labios genera un sonido bilabial. La manera de articulación se refiere a si el sonido es producido sin o con obstrucciones. En la Tabla 4.1 se muestra la categorización de consonantes de acuerdo con el lugar y la manera de articulación en el idioma inglés. Se puede observar que según las articulaciones y la obstrucción de aire se clasifican en diferentes categorías.

Tabla 4.1 Clasificación de consonantes de acuerdo con el lugar y la manera de articulación.

		Lugar de articulación					
		Bilabial	Labiodental	Dental	Alveolar	Palatal	Velar
Manera de articulación	Oclusiva	/p/, /b/		/t/, /d/			/k/, /g/
	Fricativa		/f/		/s/	/j/	/x/
	Africada					/tʃ/	
	Nasal	/m/			/n/		
	Laterales				/l/		
	Percusivas				/r/		
	Vibrantes				/r/		

Tabla 4.2 Propuesta de categorización de palabras y fonemas para KaraOne.

		Comando										
		/iy/	/uw/	/piy/	/tiy/	/diy/	/m/	/n/	/pat/	/pot/	/gnaw/	/knew/
Clase	BL	-	-	+	-	-	-	-	-	+	+	-
	NS	-	-	-	-	-	-	-	+	-	-	+
	C/V	2	2	1	1	1	1	1	1	1	1	1
	/iy/	+	-	-	-	-	-	-	-	-	-	-
	/uw/	-	+	-	-	-	-	-	-	-	-	-

En el caso de los fonemas y palabras de la base de datos de KaraOne, no solo se producen sonidos de consonantes, por lo que se tomó la información de la Tabla 4.1 para proponer una categorización de los fonemas y palabras que se incluyen. En la Tabla 4.2 se muestra la propuesta utilizada para la clasificación del presente proyecto de tesis. Los símbolos + indica presencia, - indica ausencia en las clases, en el caso de los números 1 y 2 indican pertenencia a la primera o segunda letra de la clase respectivamente. Para la clasificación Bilabial (BL), /piy/, /pot/ y /pat/ son consideradas con presencia de BL porque su producción envuelve los dos labios de la boca (lugar de articulación). La categoría Nasal (NS) es cuando el aire viaja solo por la nariz sin alguna obstrucción [119], [120]. Consecuentemente, la clase NS está presente en /n/, /gnaw/ y /knew/. En el caso de /m/, se considera como parte BL y NS, por lo que para el presente estudio se excluye de las clases. La clase de Consonante o Vocal (C/V), se describe a una vocal como aquella que se produce sin ninguna obstrucción de aire, de lo contrario será consonante. Las clases presencia y ausencia de /iy/ y /uw/ es la detección de las vocales en las señales EEG.

4.2 Selección de canales

La base de datos KaraOne fue elaborada con 62 electrodos activos, y en algunos trabajos comúnmente se utilizan todos los electrodos para analizar las señales [44], [76], [121]–[123]. Sin embargo, se conoce que la producción del habla sugiere que tanto el habla articulada, los susurros y el habla imaginada tienen procesos similares generando actividad en el área frontal del cerebro [124], [125]. La presente tesis se basa en el modelo de Indefrey [126], el cual describe una organización funcional de producción de palabras donde se encontró que los componentes principales en el cerebro para la producción del habla son mayormente en el hemisferio izquierdo del cerebro. Además, incluyen un proceso de cómo es que el habla se produce, las cuales son:

1. Preparación conceptual
2. Concepto léxico: con duración de 175- 250 ms, proceso desarrollado en la sección media del giro temporal medio izquierdo.
3. Recuperación de lema, selección de lema y recuperación de código fonológico: este proceso es llevado a cabo en las secciones posteriores del giro temporal medio y superior en el hemisferio izquierdo con una duración de 250-330 ms
4. Silabificación: este proceso se observa en el giro frontal inferior temporal izquierdo con una duración de 330-455 ms.
5. Codificación fonética: llevada a cabo en el giro frontal izquierdo con una duración de 455 ms.
6. Articulación: llevada a cabo en el área motora, es la finalización del proceso.

Por esta razón, se realiza la selección de canales asociando las áreas cerebrales a los procesos de la decodificación de habla imaginada descrita por Panachakel et al. [76], la morfología del lenguaje en el cerebro detallada por Marslen-Wilson et. al [25] y un estudio realizado por Alderson-Day et. al que describe las activaciones cerebrales al realizar conversaciones de habla imaginada [3]. La selección arroja los resultados mostrados en la Tabla 4.3. La localización de las áreas que se describen en la Tabla 4.3 se muestran en la Figura 4.1, donde cada área se relaciona con una función fisiológica. En el caso del habla, el modelo sugiere a las áreas de

Broca, Wernicke y asociación límbica como las principales responsables de su producción. Los canales que corresponden a estas áreas son: F7, C5, FC3, FC1 y F5.

Tabla 4.3 Selección de canales y su localización en áreas cerebrales.

Canal	Localización
C4	Giro postcentral
FC3	Corteza premotora
FC1	Corteza premotora
F5	Giro frontal Inferior
C3	Giro postcentral
F7	Área de Broca
FT7	Giro temporal Inferior
CZ	Giro postcentral
P3	Lóbulo superior parietal
T7	Giro temporal medio Corteza auditiva secundaria
C5	Área de Wernicke Corteza auditiva primaria

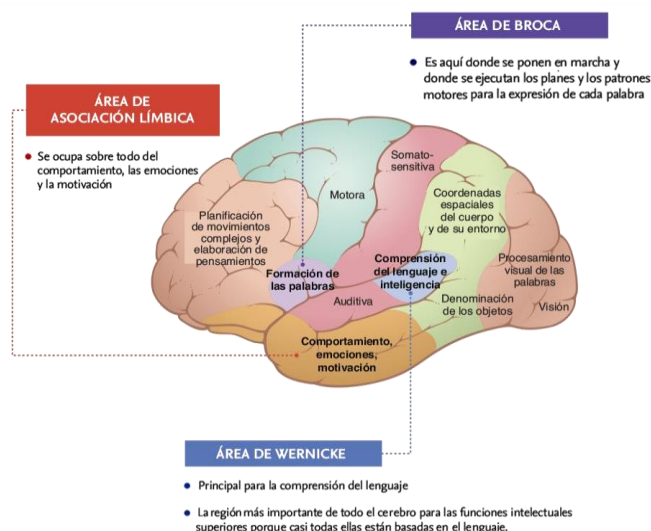


Figura 4.1 Áreas cerebrales asociadas al habla.

4.3 Selección de características

De acuerdo con la literatura, la mayoría de los trabajos resuelven el problema de reconocimiento de habla imaginada utilizando una metodología basada en el preprocesamiento de la señal de EEG, extracción de características y clasificación. Sin embargo, dado que las

señales SI son más complejas que otras señales de EEG y se requiere de un análisis y estudio de las características y propiedades extraídas para reconocer comandos en las señales EEG de SI.

De acuerdo con Islam et. al [123], la frecuencia juega un papel esencial en SI porque existe evidencia de que las bandas theta y delta son útiles en el análisis de habla articulada en unidades lingüísticas y prosódicas [127], [128]. Sin embargo, otros autores como Zhao et. al [43] utilizaron la base de datos de KaraOne y presentan un método de extracción de características estadísticas, el cual demuestra que este tipo de características pueden ser utilizadas para identificar comandos en señales EEG de SI. Por lo tanto, en el presente proyecto de tesis se utilizan características estadísticas para la etapa de extracción de características.

Para la extracción de características se ha encontrado el uso de características estadísticas en esta base de datos. Tales características son media, varianza, desviación estándar, pico máximo, pico mínimo, etc. En esta sección se realiza un análisis de las características de entrada por cada sujeto y para cada época por canal de las diferentes palabras y fonemas imaginados que componen a la base de datos.

Para la etapa de extracción de características utiliza un preprocesamiento de señales que se define como el agrupamiento de épocas por cada canal correspondiente al tiempo donde se realizó la imaginación del habla. Al conjunto de épocas se define como E_{cij} , y se describe como las i épocas de la señal EEG para cada c canal y j hace relación al fonema o palabra imaginada. En la extracción de características, para cada época i se obtienen las características: área positiva (área bajo la curva con valores positivos), área negativa (área bajo la curva con valores negativos), área total (suma de las áreas, áreas negativa y positiva), área absoluta total (suma de las áreas, áreas negativa y positiva), amplitud máxima (valor máximo negativo o positivo), latencia (número de muestras desde el inicio donde se encuentra la amplitud máxima), amplitud absoluta, radio latencia/amplitud, radio latencia/amplitud absoluta, valor pico a pico (diferencia entre los valores máximo y mínimo), pendiente pico-pico (relación entre el valor pico-pico y la ventana temporal pico-pico), ventana de tiempo pico a pico, cruces por cero (momentos en que la señal EEG pasa por cero), densidad de cruces por cero (radio entre cruces por cero y ventana de tiempo pico a pico), desviación estándar varianza, media, mediana, moda redondeada un

decimal (valor del valor más valor más común de la señal EEG redondeado a un decimal), modo redondeado a dos decimales (valor del valor más común en la señal del EEG redondeado a dos decimales), modo redondeado a tres decimales (valor del valor más común en la señal de EEG redondeado a tres decimales), frecuencia media (un centroide del espectro), moda y frecuencia mediana.

Estas características se consideraron por el esquema de características estadísticas presentado por Zhao et. al [43] y por la teoría consonante vocal [68], la cual indica que, en la producción de fonemas y palabras, los fonemas producen ondas con negatividad posterior, mientras que las vocales generan negatividad anterior. Esta negatividad puede ser observada como un Potencial Relacionado a Eventos (ERP por sus siglas en inglés) endógeno con diferentes picos, amplitud, latencia y polaridad. También, todos estos ERP's se relacionan con la actividad cerebral y representan importantes características como frecuencia, el tiempo entre picos y otras características estadísticas.

Al conjunto de características se le denotará como la matriz Fc y antes de la siguiente etapa se medirá si los grupos de características por canal tienen un comportamiento normal a través de las diferentes épocas obtenidas. Esto debido a que posteriormente se utiliza un análisis de varianzas entre grupos y es necesario contar con una distribución normal. Para ello, se utiliza la prueba *Shapiro Wilk* donde se define una hipótesis nula:

$$H_0 : [Fc_1, Fc_2, \dots, Fc_n] \rightarrow Fc \approx N(\mu, \sigma) \quad 4.1$$

donde se asume que el conjunto Fc tiene una media μ y una desviación estándar σ normal. Para aceptar o rechazar la hipótesis nula se define un análisis estadístico dado por:

$$W = \frac{\left(\sum_{i=1}^n a_i Fc_i \right)^2}{\sum_{i=1}^n (Fc_i - \bar{x})^2} \quad 4.2$$

donde los coeficientes a están dados por los valores estadísticos ordenados definidos por:

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-2} m}} \quad 4.3$$

Donde V son las matrices de covarianza y finalmente se acepta H_0 si $W > 0$. Si los conjuntos de características para cada canal y época son normales, estos se incluyen para el estudio de la relación de características y canales. El estudio presentado de selección de características se lleva a cabo por medio de ANOVA, el cual es un análisis de varianzas de grupos donde se contrastan las siguientes hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \dots = \mu_k \quad H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \dots \neq \mu_k \quad 4.4$$

donde la hipótesis nula es aquella que determina que las medias de grupos son iguales entre sí, mientras la hipótesis alternativa es aquella que indica que al menos una media de los grupos es diferente. Para comprobar estas hipótesis se obtienen los valores de significancia dados por la suma de cuadrados por grupos (SC_T) y entre grupos (SC_{EG}) de características definido por:

$$SC_T = \left(\sum x_1^2 + \sum x_2^2 + \dots + \sum x_k^2 \right) - \frac{\left(\sum x_1 + \sum x_2 + \dots + \sum x_k \right)^2}{N} \quad 4.5$$

$$SC_{EG} = \left[\frac{\left(\sum x_1 \right)^2}{k} + \frac{\left(\sum x_2 \right)^2}{k} \dots \frac{\left(\sum x_k \right)^2}{k} \right] - \frac{\left(\sum x_1 + \sum x_2 + \dots + \sum x_k \right)^2}{N} \quad 4.6$$

y con estos valores obtener el valor de significancia intragrupo (SC_{IG}) definido por la siguiente ecuación:

$$SC_{IG} = SC_T - SC_{EG} \quad 4.7$$

Finalmente, con los valores de significancia se seleccionan las características por canal en las que su comportamiento es normal, y además tienen un comportamiento similar a lo largo de las épocas. Un ejemplo de la selección de características se puede observar en la Tabla 4.4, donde se muestran algunas de las características del sujeto MM05.

Para cada fonema y palabra imaginada, se seleccionan 30 características en total, las cuales conforman el vector de características $P_i(n)$. Es importante destacar que, para cada sujeto, el ranking de características cambia de acuerdo con los resultados de ANOVA. Este proceso selecciona las principales para cada sujeto por lo que se adapta al reconocimiento de SI. Para

probar la eficiencia de la selección de características por medio del análisis de varianzas, se realizó un estudio que consiste en realizar una selección aleatoria de características de los canales en vez de utilizar ANOVA. Finalmente, se observó con la selección aleatoria las métricas como exactitud, precisión y reconocimiento fueron más bajas que aquellos utilizando ANOVA. También, se realizó un análisis para saber cuáles son las características que más relevancia presentan. El análisis consiste en realizar un conteo de frecuencia de selección de cada característica a través de todos los sujetos. Los resultados del análisis se pueden observar en la Tabla 4.5. El análisis concluye que las características más comunes son valor pico-pico, área negativa, varianza, desviación estándar, media, área total, área total absoluta, latencia, cruces por cero y frecuencia media.

Tabla 4.4 Características seleccionadas en la base de datos de KaraONE.

Canal	Característica
F7	Área total absoluta
F7	Área positiva
F5	Frecuencia media
F7	Desviación estándar
F5	Cruces por cero
C3	Desviación estándar
FT7	Frecuencia media
C3	Amplitud Absoluta
C3	Varianza
F7	Área negativa

El conjunto de características seleccionado por ANOVA es utilizado como vector de características a clasificar por medio de CapsK-SI. La metodología de clasificación se realiza por grupos de clases de acuerdo con la Tabla 4.2. Con base en los criterios de diseño de CapsVI se realizaron modificaciones para el análisis de palabras y fonemas. El diseño de las capas de CapsVI se mantiene igual. Sin embargo, la entrada, neuronas y *kernels* se definen como:

1. Entrada: son las características estadísticas de la señal EEG en un ranking por medio de ANOVA. El vector de características $P_t(n)$ con dimensiones 1×30 para cada época tiene baja resolución por los cálculos convolucionales. Por lo tanto, $P_t(n)$ es redimensionado creando un nuevo arreglo de características, utilizando aumentación

- de datos que consiste en duplicar los renglones de características. Consecuentemente, el vector de datos resulta en un arreglo de dimensiones de 30x30.
- Número de neuronas en las cápsulas: el diseño de CapsVI indica que CapsK-SI debe de tener ocho neuronas en la capa PC y 16 neuronas en la capa CC de acuerdo con [109].
 - Tamaño de *kernels* en la capa convolucional y capa PC: el diseño de CapsVI muestra que CapsK-SI debe de tener 10 *kernels* con dimensiones de 9x9 en la primera capa convolucional y 160 con la misma dimensión en la capa PC.

Tabla 4.5 Frecuencia de ocurrencia de cada característica en los sujetos utilizando ANOVA.

Característica	Sujeto												Frecuencia
	MM05	MM08	MM09	MM10	MM11	MM12	MM14	MM16	MM18	MM19	MM20	MM21	
Área positiva	1				1	3	3	2	1	3		2	16
Área negativa	5		11					5	1	1		5	28
Área total	4			7	4			1		5			21
Área absoluta total	3					4	2	3	2	2		4	20
Amplitud máxima					3			1	3			2	9
Latencia		3		1	3		2	3	2	2	4		20
Amplitud absoluta								1	1		11	5	18
Radio de latencia/amplitud	1	1			1		2	1	2		2		10
Radio absoluto latencia/amplitud				1	1		2	1	3	3			11
Valor pico pico	1		10			11	3		1	1		5	32
Pendiente de pico a pico		4			2		2						8
Ventana de tiempo de pico a pico	1	2		3	1		3		2				12
Cruces por cero	1	7		6	1					1	4		20
Densidad de cruces por cero		2			3		2				1		8
Desviación estándar	2		4			6	1	2	1	3		5	24
Varianza	2		5			6	2	2	2		7	2	28
Media	5			7	5			1	2	5			25
Mediana	3			2			2	4	2	1			14
Moda redondeada un decimal										1			1
Moda redondeada dos decimales													0
Moda redondeada tres decimales													0

Frecuencia media	1	5		1	2					1			10
Frecuencia moda		5		1	2					1			9
Frecuencia mediana		1		1	1		4	3	5		1		16

En la Figura 4.2 se muestra la arquitectura de CapsK-SI, donde la arquitectura de CapsVI se preserva, pero la entrada e hiper-parámetros se modifican. Los cálculos que se realizan en el modelo son los mismos que se presentan en el capítulo 3.

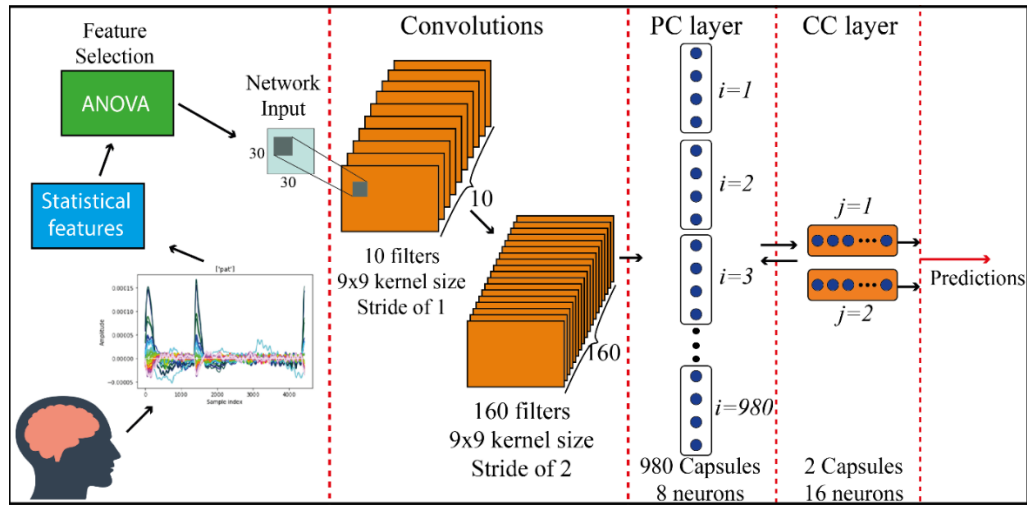


Figura 4.2 Arquitectura del modelo CapsK-SI.

4.4 Resultados de clasificación de CapsK-SI

Para evaluar los resultados del rendimiento del modelo, se utilizan las métricas por clase de *accuracy* (ACC), precisión (PR), sensibilidad (SS), especificidad (SC), coeficiente de correlación de *Matthew's* y *balanced accuracy* (MCC). Todas estas métricas se obtienen con los Verdaderos Negativos (TN por sus siglas en inglés), Verdaderos Positivos (TP por sus siglas en inglés), Falsos Negativos (FN por sus siglas en inglés) y Verdaderos Negativos (VN) de las matrices de confusión generadas. De esta forma, ACC es obtenida mediante la ecuación 1.8 dividiendo la suma de TP y TN entre la suma total de predicciones del modelo. La PR es la razón de TP con todas las predicciones positivas, tal y como se muestra en la ecuación 1.9. La SS es obtenida mediante la ecuación 1.10 y consiste en la razón de TP con la suma de TP y FN para conocer la influencia de los FN en las predicciones del modelo. La SC es obtenida mediante la ecuación 1.11 mediante la razón de TN y todas las predicciones incorrectas. El MCC es calculado con la

ecuación 1.12. BACC es calculada mediante el promedio de SS y SC. BACC se utiliza cuando existe un desbalance de clases.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad 4.8$$

$$PR = \frac{TP}{TP + FP} \quad 4.9$$

$$SS = \frac{TP}{TP + FN} \quad 4.10$$

$$SC = \frac{TN}{TN + FP} \quad 4.11$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad 4.12$$

CapsK-SI se entrena por sujeto para cada una de las clasificaciones binarias. Consecuentemente, la partición de las señales EEG para cada sujeto es de 80% para entrenamiento y 20% para prueba. En cada una de las iteraciones de la validación cruzada, un 25% del entrenamiento se utiliza para conformar el conjunto de validación. Al final de cada época, la ACC del conjunto de prueba es reportada. De tal forma que, durante la evaluación, se utiliza esta métrica para el reporte de resultados.

A continuación, se muestran los resultados de la clasificación de tareas binarias con CapsK-SI definidas en la Tabla 4.2. La Tabla 4.6 muestra los resultados para la clase BL, la Tabla 4.7 para la clase NS, la Tabla 4.8 para la clase C/V, la Tabla 4.9 para presencia/ausencia de /iy/, y, Tabla 4.10 para presencia/ausencia de /uw/ (se incluye la desviación estándar multiplicada por 100 en el caso de las Tablas 4.6-4.10). Dado que se realiza una clasificación binaria, debe tenerse en cuenta la relación de las instancias. En el caso de la clasificación BL y NS hay un 27,27% de presencia y un 72,72% de ausencia. Para la clasificación C/V es del 81,81% para las consonantes y del 18,18% para las vocales. En el caso de las vocales /iy/ y /uw/, la presencia es del 9,09% y la ausencia del 90,90%.

Tabla 4.6 Resultados de la clasificación bilabial de KaraONE.

Bilabial										
Sujetos			Predicción		ACC	PR	SS	SC	MCC	BACC
			Negativo	Positivo						
MM05	Etiqueta	Negativo	94	2	97.73% ± 3	0.95	0.97	0.99	0.94	0.98
		Positivo	1	35						
MM08	Etiqueta	Negativo	90	5	92.37% ± 5	0.86	0.86	0.95	0.81	0.90
		Positivo	5	31						
MM09	Etiqueta	Negativo	93	3	81.82%±10	0.83	0.42	0.82	0.50	0.62
		Positivo	21	15						
MM10	Etiqueta	Negativo	93	3	96.97%±5	0.92	0.97	0.99	0.93	0.98
		Positivo	1	35						
MM11	Etiqueta	Negativo	94	2	95.45%±3	0.94	0.89	0.96	0.88	0.92
		Positivo	4	32						
MM12	Etiqueta	Negativo	95	1	76.52%±6	0.86	0.17	0.76	0.31	0.46
		Positivo	30	6						
MM14	Etiqueta	Negativo	89	7	80.30%±4	0.71	0.47	0.82	0.46	0.65
		Positivo	19	17						
MM16	Etiqueta	Negativo	90	6	89.39%±3	0.82	0.78	0.92	0.73	0.85
		Positivo	8	28						
MM18	Etiqueta	Negativo	92	4	96.21%±3	0.90	0.97	0.99	0.91	0.98
		Positivo	1	35						
MM19	Etiqueta	Negativo	96	0	95.45%±5	1.00	0.83	0.94	0.89	0.89
		Positivo	6	30						
MM20	Etiqueta	Negativo	95	1	98.48%±2	0.97	0.97	0.99	0.96	0.98
		Positivo	1	35						
MM21	Etiqueta	Negativo	94	2	90.15%±9	0.93	0.69	0.90	0.74	0.79
		Positivo	11	25						
Promedio					90.90%±7	0.89	0.75	0.92	0.76	0.83

En el caso de los resultados de clasificación presencia/ausencia de BL mostrados en la Tabla 4.6, la exactitud media para todos los sujetos es del 90.90% ± 7, alcanzando el 98.48% ± 2 para el sujeto MM20 y un mínimo de 76.52% ± 6 para el sujeto MM12. La relación de clases en la categoría BL es del 72.72% para la ausencia de bilabial y del 27.27% para la presencia. A continuación, CapsK-SI aprende la diferencia entre las clases de presencia/ausencia, también en todos los sujetos la precisión media para cada sujeto es superior al 72.72%.

Para la clasificación de presencia/ausencia de NS tomando en cuenta los resultados mostrados en la Tabla 4.7, la exactitud media en todos los sujetos es del 90,15% ± 8, alcanzando el 98.48% ± 2 en el caso del sujeto MM10 y un mínimo del 74.24% ± 12 para el sujeto MM09. De acuerdo con los resultados mostrados en la Tabla 4.7 en la clasificación C/V, el método de

CapsK-SI alcanza una precisión media del $94.01\% \pm 6$, alcanzando el $100.00\% \pm 0$ para el sujeto MM19 y un mínimo de $82.71\% \pm 7$ para el sujeto MM12.

Tabla 4.7 Resultados de la clasificación NS de KaraONE.

Nasal										
Sujetos			Predicción		ACC	PR	SS	SC	MCC	BACC
			Negativo	Positivo						
MM05	Etiqueta	Negativo	95	1	92.42%±5	0.96	0.75	0.91	0.81	0.83
		Positivo	9	27						
MM08	Etiqueta	Negativo	92	4	93.94%±4	0.89	0.89	0.96	0.85	0.92
		Positivo	4	32						
MM09	Etiqueta	Negativo	88	8	74.24%±12	0.56	0.28	0.77	0.25	0.52
		Positivo	26	10						
MM10	Etiqueta	Negativo	96	0	98.48%±2	1.00	0.94	0.98	0.96	0.96
		Positivo	2	34						
MM11	Etiqueta	Negativo	94	2	93.94%±5	0.94	0.83	0.94	0.84	0.89
		Positivo	6	30						
MM12	Etiqueta	Negativo	95	1	77.27%±5	0.88	0.19	0.77	0.34	0.48
		Positivo	29	7						
MM14	Etiqueta	Negativo	94	2	89.39%±8	0.92	0.67	0.89	0.72	0.78
		Positivo	12	24						
MM16	Etiqueta	Negativo	93	3	95.45%±3	0.92	0.92	0.97	0.89	0.94
		Positivo	3	33						
MM18	Etiqueta	Negativo	96	0	96.97%±5	1.00	0.89	0.96	0.92	0.92
		Positivo	4	32						
MM19	Etiqueta	Negativo	94	2	93.94%±4	0.94	0.83	0.94	0.84	0.89
		Positivo	6	30						
MM20	Etiqueta	Negativo	96	0	85.61%±10	1.00	0.47	0.83	0.63	0.65
		Positivo	19	17						
MM21	Etiqueta	Negativo	94	2	90.15%±6	0.93	0.69	0.90	0.74	0.79
		Positivo	11	25						
Promedio					90.15%±8	0.91	0.70	0.90	0.73	0.80

Por lo tanto, en la clasificación C/V con los resultados de exactitud media, podemos decir que CapsK-SI separa los grupos de clasificación clases. Sólo para el sujeto MM12, la precisión de $82.71\% \pm 7$ es menor al 81.81% de la relación de clases, donde podemos concluir que la varianza entre las características de clase no es suficiente para clasificar correctamente a este sujeto. Para el caso de la clasificación de presencia/ausencia de /iy/ de la Tabla 4.9, CapsK-SI alcanza un $94.32\% \pm 3$ de precisión media, y el valor máximo es para el sujeto MM10 con un $99.24\% \pm 2$.

Tabla 4.8 Resultados de la clasificación C/V de KaraONE.

C/V										
Sujetos			Predicción		ACC	PR	SS	SC	MCC	BACC
			Negativo	Positivo						
MM05	Etiqueta	Negativo	107	0	97.71%±6	1.00	0.88	0.97	0.92	0.92
		Positivo	3	21						
MM08	Etiqueta	Negativo	105	2	95.42%±6	0.91	0.83	0.96	0.84	0.90
		Positivo	4	20						
MM09	Etiqueta	Negativo	101	7	84.09%±9	0.59	0.42	0.88	0.41	0.65
		Positivo	14	10						
MM10	Etiqueta	Negativo	107	1	91.67%±7	0.93	0.58	0.91	0.70	0.75
		Positivo	10	14						
MM11	Etiqueta	Negativo	105	3	96.21%±5	0.88	0.92	0.98	0.87	0.95
		Positivo	2	22						
MM12	Etiqueta	Negativo	105	3	82.71%±7	0.63	0.20	0.84	0.28	0.52
		Positivo	20	5						
MM14	Etiqueta	Negativo	108	0	97.73%±3	1.00	0.88	0.97	0.92	0.92
		Positivo	3	21						
MM16	Etiqueta	Negativo	105	3	96.21%±3	0.88	0.92	0.98	0.87	0.95
		Positivo	2	22						
MM18	Etiqueta	Negativo	108	0	95.52%±6	1.00	0.77	0.95	0.85	0.86
		Positivo	6	20						
MM19	Etiqueta	Negativo	108	0	100.00%±0	1.00	1.00	1.00	1.00	1.00
		Positivo	0	24						
MM20	Etiqueta	Negativo	107	1	93.18%±5	0.94	0.67	0.93	0.76	0.80
		Positivo	8	16						
MM21	Etiqueta	Negativo	106	2	97.73%±3	0.92	0.96	0.99	0.93	0.97
		Positivo	1	23						
Promedio					94.01%±6	0.89	0.75	0.95	0.78	0.85

Para la clasificación de presencia/ausencia de /uw/ CapsK-SI alcanza un $94.19\% \pm 3$ de precisión media. En este caso, el sujeto MM10 también obtiene la mejor clasificación con un $99.24\% \pm 2$. Se puede observar que para el sujeto MM10 se obtuvo una exactitud mayor en todas las tareas de clasificación. Por último, queríamos encontrar los patrones que codifican las cápsulas CC para reconocer la presencia de BL y NS y las diferencias entre imaginar una consonante y una vocal.

Los resultados demuestran aprendizaje por el modelo de red neuronal en la clasificación fonológica. De este modo se observa que existe un patrón de habla imaginada que está siendo reconocido. Para comprobar esto, se utilizan los vectores de activación s_j de las cápsulas clasificadoras en el modelo de red neuronal. Dado que cada uno de los escalares de los vectores está asociado con un canal en el arreglo de entrada, por lo que, utilizando esta información, de

acuerdo con los valores de activación de cada escalar se utilizan para producir mapas cerebrales en las categorías de bilabial, nasal y consonante vocal.

Tabla 4.9 Resultados de la clasificación /iy/ de KaraONE.

/iy/										
Sujetos			Predicción		ACC	PR	SS	SC	MCC	BACC
			Negativo	Positivo						
MM05	Etiqueta	Negativo	116	4	94.70%±3	0.69	0.75	0.97	0.69	0.86
		Positivo	3	9						
MM08	Etiqueta	Negativo	120	0	96.97%±2	1.00	0.67	0.97	0.80	0.82
		Positivo	4	8						
MM09	Etiqueta	Negativo	120	0	93.18%±5	1.00	0.25	0.93	0.48	0.59
		Positivo	9	3						
MM10	Etiqueta	Negativo	119	1	99.24%±2	0.92	1.00	1.00	0.96	1.00
		Positivo	0	12						
MM11	Etiqueta	Negativo	117	3	96.97%±2	0.79	0.92	0.99	0.83	0.95
		Positivo	1	11						
MM12	Etiqueta	Negativo	119	1	90.91%±6	0.50	0.08	0.92	0.18	0.50
		Positivo	11	1						
MM14	Etiqueta	Negativo	116	4	94.70%±6	0.69	0.75	0.97	0.69	0.86
		Positivo	3	9						
MM16	Etiqueta	Negativo	120	0	93.94%±4	1.00	0.33	0.94	0.56	0.64
		Positivo	8	4						
MM18	Etiqueta	Negativo	118	2	93.18%±3	0.71	0.42	0.94	0.51	0.68
		Positivo	7	5						
MM19	Etiqueta	Negativo	119	1	93.94%±4	0.83	0.42	0.94	0.56	0.68
		Positivo	7	5						
MM20	Etiqueta	Negativo	117	3	90.91%±7	0.50	0.25	0.93	0.31	0.59
		Positivo	9	3						
MM21	Etiqueta	Negativo	120	0	93.18%±3	1.00	0.25	0.93	0.48	0.59
		Positivo	9	3						
Promedio					94.32%±3	0.80	0.51	0.95	0.59	0.73

Tabla 4.10 Resultados de la clasificación /uw/ de KaraONE.

/uw/										
Sujetos			Predicción		ACC	PR	SS	SC	MCC	BACC
			Negativo	Positivo						
MM05	Etiqueta	Negativo	120	0	91.67%±3	1.00	0.08	0.92	0.28	0.50
		Positivo	11	1						
MM08	Etiqueta	Negativo	120	0	93.94%±8	1.00	0.33	0.94	0.56	0.64
		Positivo	8	4						
MM09	Etiqueta	Negativo	119	1	93.18%±5	0.80	0.33	0.94	0.49	0.64
		Positivo	8	4						
MM10	Etiqueta	Negativo	120	0	99.24%±2	1.00	0.92	0.99	0.95	0.95
		Positivo	1	11						
MM11	Etiqueta	Negativo	119	1	96.97%±5	0.90	0.75	0.98	0.81	0.86
		Positivo	3	9						
MM12	Etiqueta	Negativo	118	2	90.15%±2	0.33	0.08	0.91	0.13	0.50
		Positivo	11	1						
MM14	Etiqueta	Negativo	118	2	93.94%±6	0.75	0.50	0.95	0.58	0.73
		Positivo	6	6						
MM16	Etiqueta	Negativo	120	0	96.21%±5	1.00	0.58	0.96	0.75	0.77
		Positivo	5	7						
MM18	Etiqueta	Negativo	120	0	94.70%±6	1.00	0.42	0.94	0.63	0.68
		Positivo	7	5						
MM19	Etiqueta	Negativo	119	1	94.70%±3	0.86	0.50	0.95	0.63	0.73
		Positivo	6	6						
MM20	Etiqueta	Negativo	117	3	89.39%±4	0.25	0.08	0.91	0.10	0.50
		Positivo	11	1						
MM21	Etiqueta	Negativo	120	0	96.21%±5	1.00	0.58	0.96	0.75	0.77
		Positivo	5	7						
Promedio					94.19%±3	0.82	0.43	0.95	0.55	0.69

4.5 Generación de mapas de actividad cerebral

Con base en el hecho de que las cápsulas codifican las características presentes en la entrada [115] y en el hecho de que la capa CC codifica las características estadísticas de las señales EEG, decidimos analizar la relación entre las cápsulas de la capa CC con la actividad cerebral, puesto que cada característica proviene de un canal relacionado con una ubicación específica de la corteza cerebral. Dado que la capa de reconstrucción decodifica los patrones que se crearon para clasificar utilizando los vectores de actividad de las cápsulas CC, podemos utilizar la reconstrucción como un decodificador y crear una matriz de valores que representan las características aprendidas, tal y como se muestra en la Figura 4.3. De esta forma, se alimenta a CapsK-SI con todos los vectores de características para cada sujeto y cada grupo de clasificación para realizar predicciones.

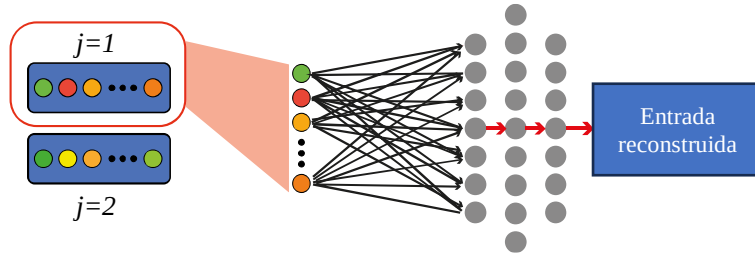


Figura 4.3 Proceso para generación de mapas cerebrales.

Después, sumamos todas las matrices de salida $R_{\varpi,t}(n,m)$ (presencia/ausencia) para cada predicción de la red, donde ϖ es el número de arreglo, t es la clase y $\hat{R}$ es el arreglo promedio resultante con dimensiones n,m . Esto está descrito por la siguiente ecuación:

$$\hat{R}_t(n,m) = \frac{1}{M_t} \sum_{\varpi=1}^{M_t} R_{\varpi,t}(n,m) \quad 4.13$$

Los valores de la matriz de la entrada reconstruida $\hat{R}$ en su dimensión n , representa el valor de la característica de entrada y su respectivo canal, resultado de la decodificación de los vectores de predicción de las cápsulas. Estos valores se utilizan para construir un gráfico que describa la actividad cerebral en términos del valor de reconstrucción obtenido de la capa CC para cada canal. Las Figuras 4.4, 4.5, 4.6 muestran los resultados de la actividad cerebral en las categorías BL, NS y C/V, respectivamente. Los sujetos están ordenados según su precisión media en orden descendente a medida que disminuye su precisión.

Los mapas de presencia de BL incluyen las activaciones cerebrales de todos los sujetos del conjunto de datos en función de su orden de presentación. La Figura 4.4 muestra las similitudes de activación en los sujetos de las dos primeras filas, que son los sujetos que obtuvieron un rendimiento del 95.47% hacia arriba. En particular, se observa que en la producción de palabras y fonemas bilabiales se producen activaciones en los canales F7, T7, FC3 y FC1, relacionados con el área de Broca, y la corteza auditiva secundaria y la corteza premotora. Asimismo, el canal C5, asociado al área de Wernicke y al córtex auditivo primario, también muestra activaciones en algunos sujetos. Se observa que el reconocimiento del BL se basa en diferencias en su producción, principalmente en el movimiento muscular. Se observa que el patrón que la red

codifica para el habla imaginada BL se encuentra principalmente en el área premotora, el área auditiva y el área de Wernicke.

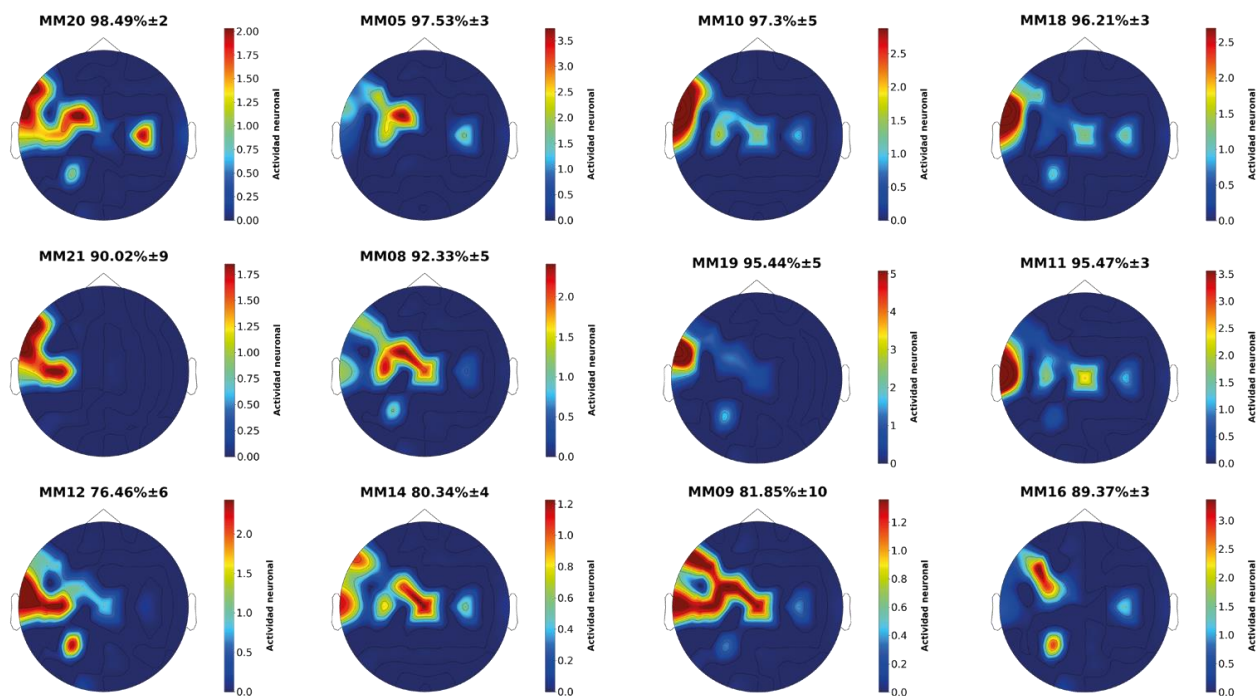


Figura 4.4 Activaciones cerebrales en la imaginación de palabras y fonemas bilabiales.

La Figura 4.5 representa el patrón codificado por la red en las imágenes NS. Puede observarse que las principales activaciones se encuentran en el área de Broca y la circunvolución frontal inferior (F7 y F5), la circunvolución postcentral (C4) y la circunvolución temporal inferior (FT7). Los sujetos con rendimientos superiores al 90,14% presentan una actividad consistente en los canales F7 y C4. Por lo tanto, se observa que la actividad de interés podrá encontrarse en el área de Broca y en la circunvolución postcentral. A diferencia de la imaginación BL (que implica principalmente el movimiento de ambos labios y, en consecuencia, el área premotora), en la imaginación NS sólo se activa el área de Broca en la que se forma la palabra, ya que los movimientos musculares son menores que en BL.

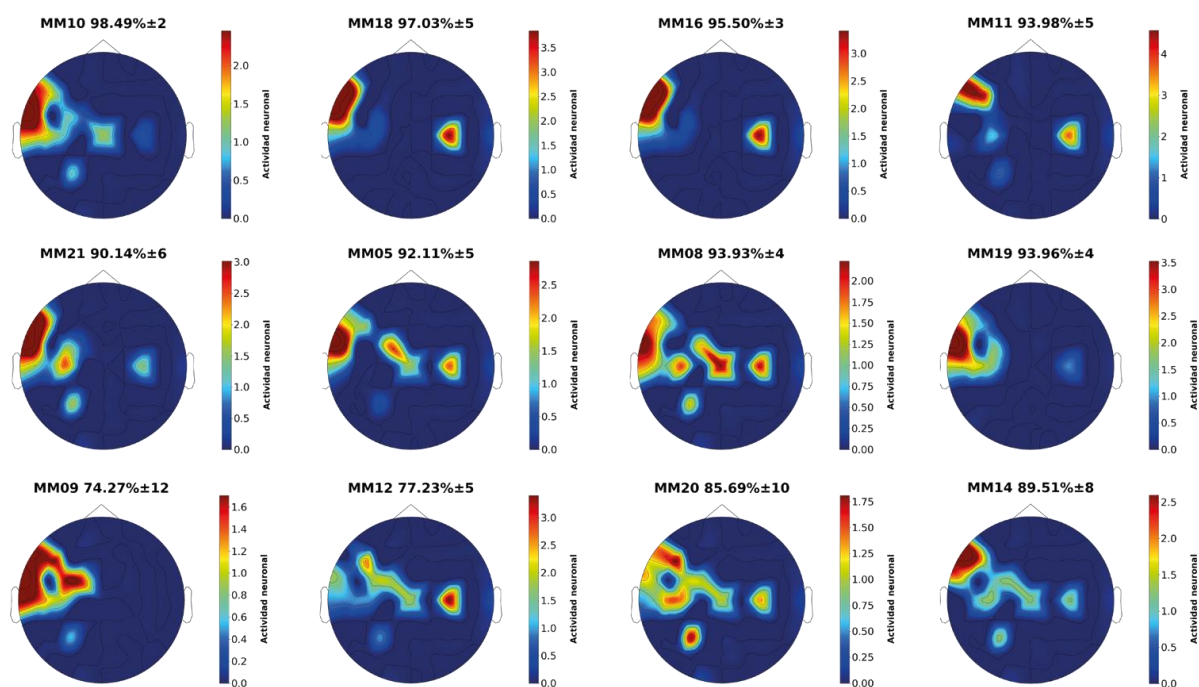


Figura 4.5 Activaciones cerebrales en la imaginación de palabras y fonemas nasales.

La Figura 4.6 muestra que las principales diferencias en la producción de consonantes y vocales son la actividad producida en el área premotora de los canales FC1 y FC3, el área de Wernicke del canal C5 y el giro postcentral de los canales Cz y C3. La actividad para las consonantes imaginarias en estas áreas es mayor porque su producción es más compleja que la de las vocales imaginarias.

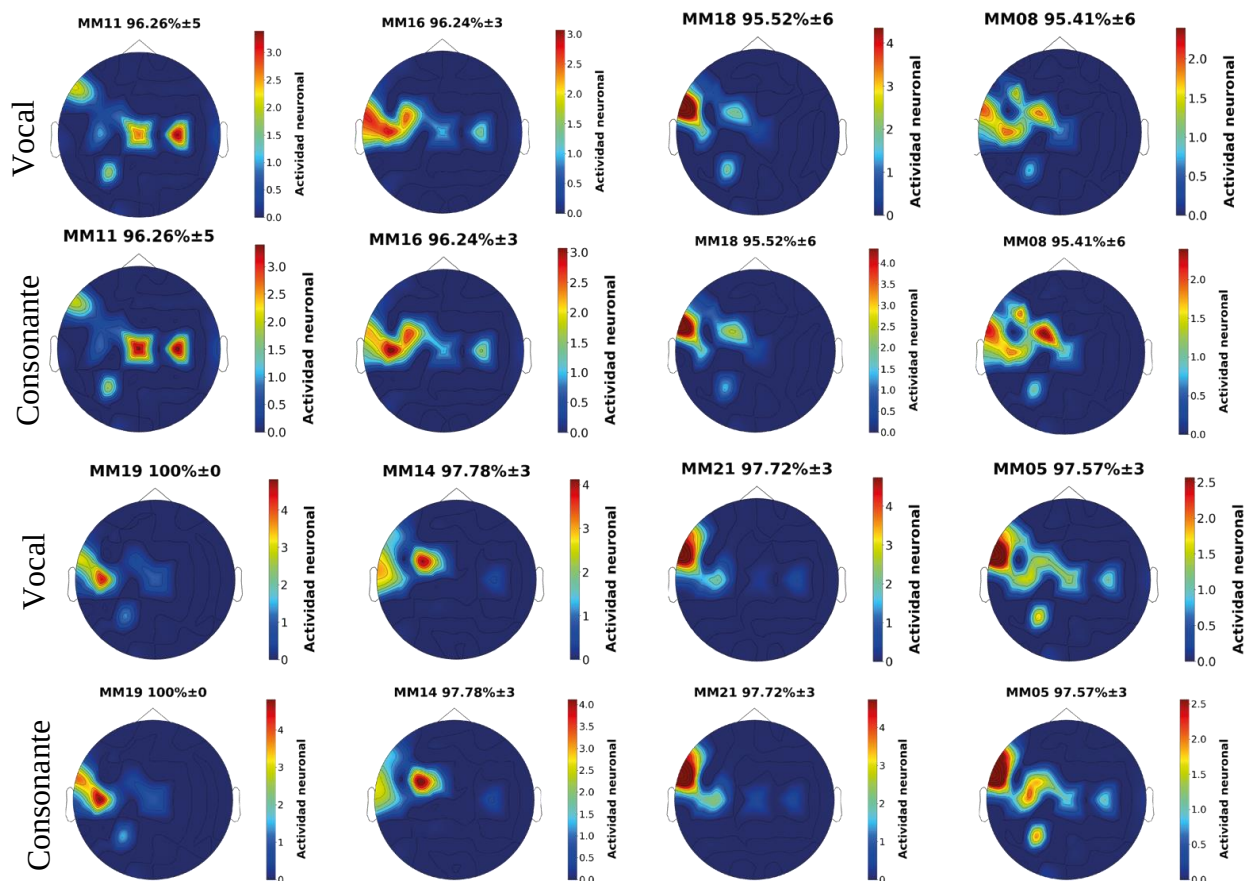


Figura 4.6 Activaciones cerebrales en la imaginación de consonantes y vocales.

En el caso de la producción bilabial de la Figura 4.4 se puede concluir lo siguiente:

- La corteza premotora (FC3, FC1) muestra mayor actividad en conjunto con la corteza auditiva primaria (T7) y el área de Wernicke (C5).
- Asimismo, el área de Broca (F7, F5) genera actividad en la producción de bilabial.

En el caso de la producción nasal de la Figura 4.5 se puede concluir lo siguiente:

- El área de Broca (F7 y F5) muestra mayor actividad.
- El área cerebral del giro postcentral (C4) y el giro inferior temporal (FT7) muestran en la producción de fonemas y palabras nasales.

En el caso de la producción de consonantes y vocales de la Figura 4.6, se puede concluir lo siguiente:

- La principal diferencia está en la actividad producida en el área premotora (FC1, FC3), en el área de Wernicke (C5) y en el giro postcentral (C3).

4.6 Discusión

De acuerdo con el análisis de selección de características, las propiedades estadísticas más comunes de la morfología y la frecuencia de las señales de imágenes del habla. Las características de media, área negativa, área total, área total absoluta, varianza y desviación estándar tienen la información estadística necesaria para describir los patrones morfológicos de cada clase. Además, las características de latencia de valores pico a pico, cruces por cero y frecuencia media (centroide del espectro) tienen la información estadística necesaria para representar los patrones de frecuencia de cada clase.

CapsK-SI aprende las características $P_t(\eta, I)$ de las señales de entrenamiento con *backpropagation* y enrutamiento dinámico entre cápsulas. Estos algoritmos de entrenamiento generan vectores prototipo en las cápsulas CC con información de morfología y frecuencia de la bilabial, nasal, consonante, vocal y fonemas a partir de señales de imágenes del habla. Estos prototipos se comparan con las características estadísticas que describen la morfología y la frecuencia de cada señal de entrada de EEG de los conjuntos de prueba y validación con las ecuaciones 3.3 y 3.4. La comparación de los prototipos aprendidos en las cápsulas CC es la principal diferencia entre CapsK-SI y los otros métodos que reportan resultados con KaraOne en la literatura. Además, CapsK-SI presenta las mejores precisiones de la literatura con 90.9% ± 7 para BL, 90.15% ± 8 para NS, 94.01% ± 6 para C/V, 94.33% ± 3 para /iy/ y 94.33% ± 3 para /uw/.

Según los mapas de la sección 4.5, la imaginación de los sonidos está relacionada con la distribución espacial de la magnitud de los rasgos de morfología y frecuencia. Esto significa que las magnitudes de los rasgos son mayores en la producción del lenguaje que en el resto de las áreas cerebrales. La distribución de la magnitud de los rasgos es similar entre sujetos, pero diferente entre sonidos imaginados.

CAPITULO 5. COMPARACIÓN DE RESULTADOS UTILIZANDO CAPSK-SI

En este capítulo se presenta una comparativa del rendimiento del modelo CapsK-SI con otros trabajos que han estudiado la base de datos de KaraOne. Esta comparativa está basada en la métrica de exactitud promedio puesto que es la métrica con la que los trabajos de la literatura utilizan para compararse. Además, no proveen ninguna otra métrica con la que se pudieran hacer comparaciones objetivas. Para tener comparaciones bajo un mismo esquema de comparación se seleccionaron los siguientes:

1. Los trabajos deberán de incluir las clasificaciones binarias de ausencia y presencia de BL, NS, C/V, /iy/, y /uw/. Esto es porque el presente proyecto de tesis busca desarrollar un método que encuentre actividad interpretable que pudiera ser relacionada con las imaginaciones de palabras y fonemas en estas categorías.
2. Electrodo: el número y la localización de los electrodos deberá ser especificada.
3. Las etapas del procesamiento de las señales de EEG deberán estar claramente reportadas.

Los métodos a comparar son un clasificador cuádruple SVM utilizado en [43], K-Vecinos Más Cercanos (KNN por sus siglas en inglés) basados en la distancia Riemanniana [122], una red neuronal convolucional con un autoencoder profundo [55], una red neuronal de extremo a extremo [129], y una red de correlación [130]. También implementamos dos modelos de red neuronal profunda preentrenados conocidos como VGG-16 [131] y ResNet-50 [132] para contribuir aún más a las comparaciones. Estos modelos se seleccionaron porque VGG-16 tiene mejor precisión y configuraciones de convolución para buscar características que otras redes, como LeNet [133] y AlexNet [134]. El modelo ResNet-50 [132] fue seleccionado porque tiene altas precisiones debido al aprendizaje residual. Las redes VGG-16 y ResNet-50 fueron implementadas utilizando la librería de aplicaciones Keras, donde se encuentran pre-entrenadas. Posteriormente, se añadió una capa densa como última ambas redes. Las capas por defecto incluidas en los modelos originales se configuraron como no entrenables, es decir, solo se utilizaron los modelos como extractores de características. Utilizamos el optimizador ADAM con una tasa de aprendizaje de 0.001 (se seleccionaron los valores preestablecidos del optimizador), 200 épocas, y un *batch* de 1 para cada entrenamiento ya que se observó que el

overfitting prevalecía a medida que el entrenamiento llegaba a 100 épocas. Se utilizaron las mismas entradas que aquellas utilizadas para el modelo CapsK-SI. La Tabla 5.1 muestra que CapsK-SI obtiene los mejores resultados en todas las categorías de clasificación.

Tabla 5.1 Comparación de resultados del modelo CapsK-SI con modelos de la literatura.

Método	Sujetos	BL	NS	C/V	/iy/	/uw/
[43]	8	56.64%	63.5%	18.08%	59.6%	79.16%
[43]	8	-	-	≈89.27%	-	≈80.73%
[122]	8	69.08%	72.10%	86.52%	75.27%	83.98%
[55]	14	75.55%	73.45%	85.23%	73.30%	81.99%
[129]	14	53%	47%	25%	53%	74%
betaL [130]	14	71.21%	71.12%	76.52%	73.48%	79.55%
betaP [130]	14	75.64%	76.5%	89.39%	80.30%	82.58%
VGG-16 [131]	12	74.62%	74.31%	83.02%	91.55%	91.41%
ResNet-50 [131]	12	74.33%	73.66%	82.88%	91.10%	90.79%
CapsK-SI	12	90.90% ± 7	90.15% ± 8	94.01% ± 6	94.33% ± 3	94.19% ± 3

De acuerdo con los resultados obtenidos en la Tabla 5.1, una posible explicación de las exactitudes obtenidas son que el clasificador cuádruple SVM [43] obtuvo un rendimiento inferior porque utilizó una correlación de Pearson entre 65,835 características para seleccionar 100 características. Dada la dimensionalidad del espacio y la deficiente selección de características para la complejidad de las imágenes del habla, el SVM no pudo clasificar correctamente las señales EEG. El método [122] se basa en KNN utilizando la distancia Riemanniana para comparar matrices de densidad espectral de correntropía. La distancia se utiliza como una característica, y el modelo puede reconocer el rendimiento promedio de 90.25% entre dos palabras haciendo combinaciones entre las cuatro palabras (*gnaw*, *knew*, *pat* y *pot*) en el conjunto de datos. Sin embargo, la exactitud disminuyó considerablemente en la clasificación de ausencia y presencia de BL, NS, C/V, /iy/, y /uw/ (Tabla 5.1), esto podrá deberse a que, aunque las señales de EEG se transformaron en matrices de correlación de densidad, la única información para clasificar se basa en las distancias, resultando en una separación no

eficaz entre clases cuando se incluyen las once clases. El método presentado en [55] utiliza la covarianza cruzada de canales y una CNN para clasificar. Las características consideradas como entrada al modelo son las matrices de covarianza, y sus resultados son similares pero inferiores a los de [122]. La exactitud media de CapsK-SI es superior a la de [55] porque la CNN utilizada no pudo codificar las relaciones entre todos los canales (a diferencia de CapsK-SI que utiliza el producto punto para medir proyecciones de vectores y comparar su similitud), considerando sólo los canales que contribuyen a los fenómenos físicos de las imágenes del habla. El método de [129] propone un modelo basado en un Boltzmann restringido, que transforma las señales EEG del habla imaginada y hablada con tres redes. Los resultados reportados son inferiores a los métodos de la literatura debido a que se realiza una alta mezcla de señales EEG (habla imaginada y hablada) a través de las redes. Se puede observar, dados sus resultados, que tanto la extracción de características como los métodos de clasificación no se llevan a cabo para caracterizar las señales de EEG en un dominio adecuado para generar información que permita clasificar correctamente. En cuanto al método [130], Sharon et al. realizaron dos etapas principales de extracción de características basadas en el análisis de correlación canónica. Presentaron sus resultados del modelo de referencia (*betaL*) y el modelo de mejor rendimiento (*betaP*). Los resultados son inferiores a CapsK-SI porque el método se basa en el análisis de correlación canónica (CCA), y este método está sesgado en las fases de las señales de EEG, que se ha demostrado en el presente documento que no es suficiente para clasificar con características como sólo fase. Los resultados de VGG-16 y ResNet-50 son similares, pero no lograron un alto reconocimiento en comparación con CapsK-SI porque estos modelos necesitan conjuntos de datos más largos para lograr la generalización.

6. CONCLUSIONES

En este trabajo se analizan las propiedades de las señales de EEG del habla imaginada de EEG para encontrar patrones y características que permitan reconocer vocales fonemas y palabras. Para ello se propone un método de clasificación basado en una red neuronal de cápsulas, en el caso del análisis de vocales imaginadas se realizó el estudio por medio de un modelo propuesto llamado CapsVI, para luego analizar palabras y fonemas de acuerdo con agrupaciones basadas en diferencias fonológicas utilizando el modelo propuesto y realizar modificaciones para generar un modelo nuevo llamado CapsK-SI con el fin de obtener las características y canales útiles para el reconocimiento de patrones de palabras y fonemas imaginadas.

La arquitectura de CapsVI se basa en convolución, cápsulas de características, cápsulas de clases y capas totalmente conectadas. CapsVI se diseñó con una metodología basada en calcular el número y tamaño de los *kernels* de convolución con técnicas de correlación, además de encontrar la dimensión de las cápsulas basándose en los patrones SRP que CapsVI aprendió durante el entrenamiento. Dichos patrones son características de la actividad cerebral de la imaginación de vocales invariantes a la composición y al ruido de la señal EEG. Los experimentos se desarrollaron con el conjunto de datos de DaSalla [42] debido a que tiene muestras de tres clases, dos de vocales imaginadas y un estado de control. En cuanto a las vocales, se utilizan /a/ y /u/, que, de acuerdo con la fonética, están situadas en esquinas opuestas de producción sonidos. Con esta premisa, se presenta la conclusión en los párrafos siguientes.

En primer lugar, la varianza presentada en las señales CSP es suficiente para encontrar patrones para las vocales /a/, /u/ y estado de control en la clasificación por pares. Por lo tanto, esos patrones se clasifican con CapsVI, alcanzando un $98.33\% \pm 2$ como máximo y un $84.16\% \pm 20$ como promedio mínimo de exactitud en la validación cruzada. Estos rendimientos superan las exactitudes reportadas en la literatura.

En segundo lugar, una posible explicación a los resultados sobresalientes es que la arquitectura de CapsVI encuentra las características relevantes del CSP en las capas de convolución, y las cápsulas encuentran patrones descriptivos de /a/, /u/ y estado de control. El vector de predicción $U_{j/i}$ genera una primera clasificación que encuentra varias categorías de los diferentes patrones generados por las características /a/, /u/ y estado de control de la CSP encontradas durante el

entrenamiento. Además, con los coeficientes b_{ij} se encuentran diferentes proyecciones de U_{ji} con las clases /a/, /u/ y estado de control; las cápsulas generan resultados que son invariantes a la amplitud, la traslación o el ruido. Este resultado construye mejores patrones descriptivos que se comparan con los patrones prototipo aprendidos en las cápsulas de clase, consiguiendo un alto rendimiento. Estos resultados se deben principalmente al enrutamiento dinámico entre cápsulas, ya que puede encontrar las diferentes proyecciones de las clases de /a/, /u/, y estado de control, con el patrón de entrada. Entonces, el enrutamiento dinámico entre cápsulas ayuda a encontrar características invariantes a la amplitud, la traslación o el ruido, que son fáciles de comparar con los patrones prototipo aprendidos de las cápsulas de la capa CC los cuales son una representación comprimida de la entrada.

En el caso del análisis de fonemas y palabras con CapsK-SI, se realizó un análisis mediante ANOVA unidireccional para estudiar las varianzas entre y dentro de los grupos de características para ponderar todas las características y seleccionar las 30 de entre todas con más varianza. ANOVA selecciona diferentes características en función de la actividad cerebral de cada sujeto. Con este análisis, se codifica y clasifica las propiedades de las señales EEG con CapsK-SI en cinco categorías binarias: presencia y ausencia de BL, NS, C/V, /iy/ y /uw/.

CapsK-SI describe las imágenes del habla con características estadísticas, que son suficientes para categorizar la ausencia y presencia de BL, NS, C/V, y la presencia y ausencia de /iy/, y /uw/. Además, la exactitud media máxima es de $94,19\% \pm 3$, y la mínima de $90,15\% \pm 8$, lo que sugiere que CapsK-SI codifica la información y que existe un patrón de actividad cerebral que permite el reconocimiento de ausencia y presencia de palabras y fonemas. En nuestro caso, encontramos que las características estadísticas describen la actividad cerebral en términos de distribución de energía relacionada con la actividad de las neuronas en la producción del lenguaje.

Según las comparaciones respecto a exactitud, el rendimiento de CapsK-SI es superior al de los métodos descritos en la literatura. Además, se propuso la categorización de palabras y fonemas, que aún no se ha detallado previamente en la comunidad científica. Esta categorización es importante porque no existe un estándar específico de a qué clase corresponde cada una de los fonemas y palabras, lo que genera inconsistencia en los experimentos ya que la mayoría de los autores que trabajan en clasificación fonológica no detallan cómo se realiza la categorización.

CapsK-SI obtuvo una exactitud sobresaliente en la clasificación de patrones de diferentes clases de habla imaginada porque los vectores de actividad de la capa CC generan características que pueden relacionarse a las activaciones cerebrales. Estas características son invariantes a la traslación y la escala, lo que resulta útil para modelar las propiedades de diferentes fonemas y palabras imaginados. A continuación, los vectores de actividad distinguen entre fonemas y palabras en diferentes modalidades. Además, el número de canales de electrodos utilizados en CapsK-SI es menor que en otros métodos de habla imaginada propuestos en la literatura.

Por otra parte, los resultados informan de que las propiedades que describen los vectores de actividad de la capa CC pueden ser utilizados para representar la distribución espacial de la actividad cerebral. De esta manera, CapsK-SI no sólo clasifica con características las tareas imaginadas BL, NS, C/V, /iy/ y /uw/, sino que también es un sistema que permite hacer interpretaciones a posteriori sobre las relaciones de activación neural en diferentes áreas. En el caso de los fonemas bilabiales y la producción de palabras, el córtex premotor, el área de Broca y el córtex auditivo secundario presentan una elevada actividad cerebral, que está relacionada con la producción del habla articulada bilabial porque estas áreas son responsables del movimiento de los músculos. En el caso de la categoría nasal, en la que intervienen menos músculos para su producción, las áreas cerebrales que generan más información para discriminar son el área de Broca, temporal inferior y giro central. En el caso de las consonantes y las vocales imaginadas, la principal diferencia se encuentra en la corteza premotora, el área de Wernicke y la circunvolución postcentral porque la producción de vocales es más sencilla que la de consonantes, dado que no hay obstrucciones orales. Esta información sugiere que la capa CC de CapsK-SI encuentra la distribución espacial de la energía generada por la actividad neuronal para clasificar fonemas y palabras imaginados.

Con estas conclusiones, se presenta la creación de una base de datos descrita en el anexo A, la cual utiliza el dispositivo de adquisición de señales de EEG OpenBCI utilizando 8 electrodos activos en las posiciones que sugieren las conclusiones del hemisferio izquierdo. Finalmente, los resultados fueron inconsistentes debido a que los sujetos requieren de un entrenamiento previo para estandarizar la forma en que se realizan las imaginaciones, ya que algunos de ellos, por ejemplo en la imaginación de la vocal /a/, pueden imaginar la letra “a” gráficamente, otros pueden imaginar el sonido, o, incluso pueden imaginar los labios abiertos de la persona que está realizando

la imaginación. Esto da paso a nuevas investigaciones que han surgido recientemente, tales como visualizaciones imaginadas, la cual consiste en imaginar imágenes o gráficos, lo cual es un área nueva y diferente a habla imaginada.

BIBLIOGRAFÍA

- [1] L. Ríos P. and C. Álvarez D., “Aporte De Los Distintos Métodos Epilepsy Diagnostic : Update in Eeg Contribution,” *Revista Médica Clínica Las Condes*, vol. 24, no. 6, pp. 953–957, 2013.
- [2] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, “EEG-Based Brain-Computer Interfaces Using Motor-Imagery: Techniques and Challenges,” *Sensors*, vol. 19, no. 6, p. 1423, 2019, doi: 10.3390/s19061423.
- [3] B. Alderson-Day, S. Weis, S. McCarthy-Jones, P. Moseley, D. Smailes, and C. Fernyhough, “The brain’s conversation with itself: Neural substrates of dialogic inner speech,” *Soc Cogn Affect Neurosci*, vol. 11, no. 1, pp. 110–120, 2015, doi: 10.1093/scan/nsv094.
- [4] S. N. Abdulkader, A. Atia, and M. S. M. Mostafa, “Brain computer interfacing: Applications and challenges,” *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 213–230, 2015, doi: 10.1016/j.eij.2015.06.002.
- [5] J. A. De la O Serna, M. R. A. Paternina, A. Zamora-méndez, R. K. Tripathy, and R. B. Pachori, “EEG-Rhythm Specific Taylor – Fourier Filter Bank Implemented With O-Splines for the Detection of Epilepsy Using EEG Signals,” *IEEE Sens J*, vol. 20, no. 12, pp. 6542–6551, 2020.
- [6] C. Ouyang, R. Yang, C. Chiang, R. Wu, and L. Lin, “Clinical Neurophysiology EEG autoregressive modeling analysis : A diagnostic tool for patients with epilepsy without epileptiform discharges,” *Clinical Neurophysiology*, vol. 131, no. 8, pp. 1902–1908, 2020, doi: 10.1016/j.clinph.2020.04.172.
- [7] P. Van Mierlo, B. J. Vorderwülbecke, W. Staljanssens, M. Seeck, and S. Vulliémot, “Clinical Neurophysiology Ictal EEG source localization in focal epilepsy : Review and future perspectives,” *Clinical Neurophysiology*, vol. 131, no. 11, pp. 2600–2616, 2020, doi: 10.1016/j.clinph.2020.08.001.
- [8] S. Wang, S. Wang, S. Zhang, and Y. Wang, “Time-ResNeXt for epilepsy recognition based on EEG signals in wireless networks,” *EURASIP J Wirel Commun Netw*, vol. 2020, no. 195, pp. 1–12, 2020, doi: <https://doi.org/10.1186/s13638-020-01810-5>.
- [9] J. A. Mercado, J. I. Fuentes, C. L. Toledo, N. P. Castellanos, and J. Gutiérrez, “Design of a Flexible Platform for prototyping of FES-based Motor Rehabilitation Systems,” *22 Annual*

- Conference of the International Functional Electrical Stimulation Society Conference*, no. August, 2018.
- [10] B. Giesbrecht and J. Garrett, "Electroencephalography," *Reference Module in Neuroscience and Biobehavioral Psychology*, Jan. 2024, doi: 10.1016/B978-0-12-820480-1.00007-3.
 - [11] N. Kaongoen, J. Choi, and S. Jo, "Speech-imagery-based brain-computer interface system using ear-EEG," *J Neural Eng*, vol. 18, no. 1, 2021, doi: 10.1088/1741-2552/abd10e.
 - [12] N. Kaongoen, J. Choi, and S. Jo, "A novel online BCI system using speech imagery and ear-EEG for home appliances control," *Comput Methods Programs Biomed*, vol. 224, 2022, doi: 10.1016/j.cmpb.2022.107022.
 - [13] A. G. Lazcano-Herrera, R. Q. Fuentes-Aguilar, I. Chairez, L. M. Alonso-Valerdi, M. Gonzalez-Mendoza, and M. Alfaro-Ponce, "Review on BCI Virtual Rehabilitation and Remote Technology Based on EEG for Assistive Devices," *Applied Sciences 2022, Vol. 12, Page 12253*, vol. 12, no. 23, p. 12253, Nov. 2022, doi: 10.3390/APP122312253.
 - [14] K. Pai, R. Kallimani, S. Iyer, U. Maheswari, R. Khanai, and D. Torse, "A Survey on Brain-Computer Interface and Related Applications."
 - [15] L. Y. Ganushchak, I. K. Christoffels, and N. O. Schiller, "The use of electroencephalography in language production research: A review," *Frontiers in Psychology*, vol. 2, no. SEP. 2011. doi: 10.3389/fpsyg.2011.00208.
 - [16] K. Värbu, N. Muhammad, and Y. Muhammad, "Past, Present, and Future of EEG-Based BCI Applications," *Sensors*, vol. 22, no. 9. MDPI, May 01, 2022. doi: 10.3390/s22093331.
 - [17] M. F. Mridha, S. C. Das, M. M. Kabir, A. A. Lima, M. R. Islam, and Y. Watanobe, "Brain-computer interface: advancement and challenges," *Sensors*, vol. 21, no. 17. 2021. doi: 10.3390/s21175746.
 - [18] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "EEG signal analysis: a survey.," *J Med Syst*, vol. 34, no. 2, pp. 195–212, 2010, doi: 10.1007/s10916-008-9231-z.
 - [19] S. M. Alarcao and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Trans Affect Comput*, vol. 3045, no. c, pp. 1–20, 2017, doi: 10.1109/TAFFC.2017.2714671.
 - [20] M. Shyu, S. Chen, and S. S. Iyengar, "A Survey on Deep Learning: Algorithms , Techniques ,," vol. 51, no. 5, 2018.

- [21] M. S. Alkoffash, M. J. Bawaneh, H. Muaidi, S. Alqrainy, and M. Alzghool, “A Survey of Digital Image Processing Techniques in Character Recognition,” *International Journal of Computer Science and Network Security*, vol. 14, no. 3, pp. 65–71, 2014.
- [22] A. Mishra and S. K. Majhi, “A comprehensive survey of recent developments in neuronal communication and computational neuroscience,” *J Ind Inf Integr*, 2018, doi: 10.1016/j.jii.2018.11.005.
- [23] X. Gu *et al.*, “EEG-based Brain-Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications,” *IEEE/ACM Trans Comput Biol Bioinform*, pp. 1–22, 2021, doi: 10.1109/TCBB.2021.3052811.
- [24] C. Cooney, R. Folli, and D. Coyle, “Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface,” *iScience*, vol. 8, p. 103, Oct. 2018, doi: 10.1016/J.ISCI.2018.09.016.
- [25] W. D. Marslen-Wilson and L. K. Tyler, “Morphology, language and the brain: The decompositional substrate for language comprehension,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1481, pp. 823–836, 2007, doi: 10.1098/rstb.2007.2091.
- [26] *REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23.^a ed., [versión 23.6 en línea]*. 2023.
- [27] J. P., “Adquisición y desarrollo del lenguaje y la comunicación: una visión pragmática constructivista centrada en los contextos,” *Límite. Revista Interdisciplinaria de Filosofía y Psicología*, pp. 54–66, 2000, [Online]. Available: <https://www.redalyc.org/articulo.oa?id=83600704>
- [28] L. G. Duncan, “Language and Reading: the Role of Morpheme and Phoneme Awareness,” *Curr Dev Disord Rep*, vol. 5, no. 4, p. 226, Dec. 2018, doi: 10.1007/S40474-018-0153-2.
- [29] A. Aidinis, “PHONEMES, MORPHEMES AND LITERACY DEVELOPMENT: EVIDENCE FROM GREEK”.
- [30] S. Falk and C. D. Tsang, “6- to 9-Month old infants discriminate vowel durations in variable speech contexts,” *Infant Behav Dev*, vol. 61, 2020, doi: 10.1016/j.infbeh.2020.101475.
- [31] P. Ladefoged and I. Maddieson, “Vowels of the world’s languages,” *J Phon*, vol. 18, no. 2, pp. 93–122, 1990, doi: 10.1016/s0095-4470(19)30396-1.

- [32] H. Chen, D. T. Lee, Z. Luo, R. Y. Lai, H. Cheung, and T. Nazzi, “Variation in phonological bias: Bias for vowels, rather than consonants or tones in lexical processing by Cantonese-learning toddlers,” *Cognition*, vol. 213, 2021, doi: 10.1016/j.cognition.2020.104486.
- [33] A. A. Torres-García, “Análisis y clasificación de electroencefalogramas (EEG) registrados durante el habla imaginada,” *Inaoe*, 2016, [Online]. Available: <https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/53/1/TorresGaAA.pdf>
- [34] S. Ghosh, M. Laha, A. Konar, P. Rakshit, and A. K. Nagar, “Vowel Sound Imagery Decoding by a Capsule Network for the Design of an Automatic Mind-Driven Type-Writer,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206754.
- [35] A. B. Wohler, “Event-Related Brain Potentials Preceding Speech and Nonspeech Oral Movements of Varying Complexity,” *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 5, pp. 897–905, Oct. 1993, doi: 10.1044/jshr.3605.897.
- [36] Z. Halim, R. Baig, and S. Bashir, “Temporal patterns analysis in EEG data using sonification,” in *2007 International Conference on Information and Emerging Technologies, ICIET*, 2007. doi: 10.1109/ICIET.2007.4381303.
- [37] A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, “A survey on EEG-based imagined speech classification,” in *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications*, 2021. doi: 10.1016/B978-0-12-820125-1.00025-7.
- [38] G. Hickok, “The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model,” *J Commun Disord*, vol. 45, no. 6, pp. 393–402, 2012, doi: 10.1016/j.jcomdis.2012.06.004.
- [39] J. D. O’Connor and J. L. M. Trim, “Vowel, Consonant, and Syllable—A Phonological Definition,” *WORD*, vol. 9, no. 2, pp. 103–122, 1953, doi: 10.1080/00437956.1953.11659461.
- [40] K. Zhang and G. Peng, “The time course of normalizing speech variability in vowels,” *Brain Lang*, vol. 222, 2021, doi: 10.1016/j.bandl.2021.105028.
- [41] R. A. Sree *et al.*, “Classifying Syllables in Imagined Speech using EEG Data,” *Sci Rep*, vol. 6, no. 1, 2017.

- [42] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009, doi: 10.1016/j.neunet.2009.05.008.
- [43] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagines and articulated speech," *IEEE International Conference on Acoustics*, pp. 992–996, 2015.
- [44] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: A New Approach using Riemannian Manifold Features," *J Neural Eng*, pp. 1–16, 2017.
- [45] G. A. Pressel Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," *12th International Symposium on Medical Information Processing and Analysis*, vol. 10160, no. December, 2017, doi: 10.1117/12.2255697.
- [46] A. Antonio and T. García, "Clasificación de palabras no pronunciadas presentes en Electroencefalogramas (EEG)," 2011.
- [47] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989, doi: 10.1044/jshd.5404.482.
- [48] A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, and J. M. Ramírez-Cortés, "Análisis de señales electroencefalográficas para la clasificación de habla imaginada," *Revista Mexicana de Ingenieria Biomedica*, vol. 34, no. 1, pp. 23–39, 2013.
- [49] K. Brigham and B. V. K. V. Kumar, "Imagined Speech Classification with EEG Signals for Silent Communication: A Preliminary Investigation into Synthetic Telepathy," in *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, 2010, pp. 1–4. doi: 10.1109/ICBBE.2010.5515807.
- [50] B. Min, J. Kim, H. Park, and B. Lee, "Vowel Imagery Decoding toward Silent Speech BCI Using Extreme Learning Machine with Electroencephalogram," *Biomed Res Int*, vol. 2016, pp. 1–11, 2016, doi: 10.1155/2016/2618265.
- [51] A. Balaji *et al.*, "EEG-based classification of bilingual unspoken speech using ANN," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1022–1025. doi: 10.1109/EMBC.2017.8037000.

- [52] A. Rezazadeh Sereshkeh, R. Trott, A. Bricout, and T. Chau, “EEG Classification of Covert Speech Using Regularized Neural Networks,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 25, no. 12, pp. 2292–2300, 2017, doi: 10.1109/TASLP.2017.2758164.
- [53] C. Cooney, R. Folli, and D. Coyle, “Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG,” in *2018 29th Irish Signals and Systems Conference (ISSC)*, 2018, pp. 1–7. doi: 10.1109/ISSC.2018.8585291.
- [54] A. Kapur, S. Kapur, and P. Maes, “AlterEgo: A Personalized Wearable Silent Speech Interface,” *23rd International Conference on Intelligent User Interfaces*, 2018.
- [55] P. Saha, M. Abdul-Mageed, and S. Fels, “SPEAK YOUR MIND! Towards Imagined Speech Recognition with Hierarchical Deep Learning,” Apr. 2019, pp. 141–145. doi: 10.21437/Interspeech.2019-3041.
- [56] G. Krishna, C. Tran, M. Carnahan, and A. Tewfik, “Advancing Speech Recognition With No Speech Or With Noisy Speech,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5. doi: 10.23919/EUSIPCO.2019.8902943.
- [57] K. Anandan, S. Chengaiyan, and A. S. Rajesh, “Identification of vowels in consonant-vowel-consonant words from speech imagery based EEG signals,” *Cogn Neurodyn*, vol. 14, Apr. 2020, doi: 10.1007/s11571-019-09558-5.
- [58] J. T. Panachakel and R. A. G., “Classification of Phonological Categories in Imagined Speech using Phase Synchronization Measure,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 2226–2229. doi: 10.1109/EMBC46164.2021.9630699.
- [59] J. R. Anaraki, J. Moon, and T. Chau, “Revisiting the Application of Feature Selection Methods to Speech Imagery BCI Datasets,” pp. 1–5, 2020, [Online]. Available: <http://arxiv.org/abs/2008.07660>
- [60] M. Angrick *et al.*, “Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity,” *bioRxiv*, no. 2, 2020, doi: 10.1101/2020.12.11.421149.
- [61] T. Proix *et al.*, “Imagined speech can be decoded from low-and cross-frequency features in perceptual space,” *bioRxiv*, no. January, p. 2021.01.26.428315, 2021, doi: 10.1101/2021.01.26.428315.

- [62] N. Kaongoen, J. Choi, and S. Jo, “Speech-imagery-based brain – computer interface system using,” *J Neural Eng*, vol. 18, no. 1, 2021.
- [63] M. Jiménez-Guarneros and P. Gómez-Gil, “Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition,” *Pattern Recognit Lett*, vol. 141, pp. 54–60, 2021, doi: 10.1016/j.patrec.2020.11.013.
- [64] J. Manuel Macias-Macias, J. Alberto Ramirez-Quintana, G. Ramirez-Alonso, and M. Ignacio Chacon-Murguia, “Deep Learning Networks for Vowel Speech Imagery,” in *2020 17th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 2020, pp. 1–6. doi: 10.1109/cce50788.2020.9299143.
- [65] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, “Sonification and textification: Proposing methods for classifying unspoken words from EEG signals,” *Biomed Signal Process Control*, vol. 37, pp. 82–91, 2017, doi: 10.1016/j.bspc.2016.10.012.
- [66] M. O. Tamm, Y. Muhammad, and N. Muhammad, “Classification of vowels from imagined speech with convolutional neural networks,” *Computers*, vol. 9, no. 2, 2020, doi: 10.3390/computers9020046.
- [67] G. E. Peterson and J. E. Shoup, “A physiological theory of phonetics.,” *J Speech Hear Res*, vol. 9, no. 1, 1966, doi: 10.1044/jshr.0901.05.
- [68] J. Monte-Ordoño and J. M. Toro, “Different ERP profiles for learning rules over consonants and vowels,” *Neuropsychologia*, vol. 97, pp. 104–111, 2017, doi: <https://doi.org/10.1016/j.neuropsychologia.2017.02.014>.
- [69] L. Deecke, M. Engel, W. Lang, and H. H. Kornhuber, “Bereitschaftspotential preceding speech after holding breath*,” 1986.
- [70] A. B. Wohlert, “Event-Related Brain Potentials Preceding Speech and Nonspeech Oral Movements of Varying Complexity,” *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 5, pp. 897–905, Oct. 1993, doi: 10.1044/jshr.3605.897.
- [71] J. Bucci, C. Vilain, N. Nguyen, J. L. Schwartz, and S. Dufour, “Mapping vowel sounds onto phonemic categories in two regional varieties of French: An ERP study,” *J Neurolinguistics*, vol. 54, 2020, doi: 10.1016/j.jneuroling.2020.100891.

- [72] Y. H. Yu and V. L. Shafer, “Neural Representation of the English Vowel Feature [High]: Evidence From /ε/ vs. /ɪ/,” *Front Hum Neurosci*, vol. 15, 2021, doi: 10.3389/fnhum.2021.629517.
- [73] M. Carreiras, M. Gillon-Dowens, M. Vergara, and M. Perea, “Are Vowels and Consonants Processed Differently? Event-related Potential Evidence with a Delayed Letter Paradigm,” *J Cogn Neurosci*, pp. 21 (2):275-288, 2009, doi: <https://doi.org/10.1162/jocn.2008.21023>.
- [74] A. Rezazadeh Sereshkeh, R. Yousefi, A. T. Wong, and T. Chau, “Online classification of imagined speech using functional near-infrared spectroscopy signals,” *J Neural Eng*, vol. 16, no. 1, pp. 1–13, 2019, doi: 10.1088/1741-2552/aae4b9.
- [75] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, “Speech Recognition with No Speech or with Noisy Speech,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 1090–1094, 2019, doi: 10.1109/ICASSP.2019.8683453.
- [76] J. T. Panachakel, A. G. Ramakrishnan, and A. G. Ramakrishnan, “Decoding Imagined Speech using Wavelet Features and Deep Neural Networks,” 2020, doi: 10.1109/INDICON47234.2019.9028925.
- [77] R. A. Sharon, S. Narayanan, M. Sur, and H. A. Murthy, “An Empirical Study of Speech Processing in the Brain by Analyzing the Temporal Syllable Structure in Speech-input Induced EEG,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 4090–4094, 2019, doi: 10.1109/ICASSP.2019.8683572.
- [78] G. Maegherman, H. E. Nuttall, J. T. Devlin, and P. Adank, “Motor Imagery of Speech: The Involvement of Primary Motor Cortex in Manual and Articulatory Motor Imagery,” *Front Hum Neurosci*, vol. 13, no. June, pp. 1–10, 2019, doi: 10.3389/fnhum.2019.00195.
- [79] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, “Synchronization between overt speech envelope and EEG oscillations during imagined speech,” *Neurosci Res*, 2019, doi: 10.1016/j.neures.2019.04.004.
- [80] S. Iqbal, Muhammed Shanir P.P., Y. U. Khan, and O. Farooq, “EEG Analysis of Imagined Speech,” *International Journal of Rough Sets and Data Analysis*, vol. 3, no. 2, pp. 32–44, 2016, doi: 10.4018/ijrsda.2016040103.
- [81] C. E. Herff, “Speech Processes for Brain-Computer Interfaces,” vol. 3, 2016.

- [82] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, “Transfer learning in imagined speech EEG-based BCIs,” *Biomed Signal Process Control*, vol. 50, pp. 151–157, 2019, doi: 10.1016/j.bspc.2019.01.006.
- [83] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, “Analysis and classification of speech imagery EEG for BCI,” *Biomed Signal Process Control*, vol. 8, no. 6, pp. 901–908, 2013, doi: 10.1016/j.bspc.2013.07.011.
- [84] S. Soman and B. K. Murthy, “Using Brain Computer Interface for synthesized speech communication for the physically disabled,” *Procedia Comput Sci*, vol. 46, no. Ictict 2014, pp. 292–298, 2015, doi: 10.1016/j.procs.2015.02.023.
- [85] L. Wang, W. Huang, Z. Yang, X. Hu, and C. Zhang, “A method from offline analysis to online training for the brain-computer interface based on motor imagery and speech imagery,” *Biomed Signal Process Control*, vol. 62, no. July, p. 102100, 2020, doi: 10.1016/j.bspc.2020.102100.
- [86] W. De Mulder, S. Bethard, and M. F. Moens, “A survey on the application of recurrent neural networks to statistical language modeling,” *Comput Speech Lang*, vol. 30, no. 1, pp. 61–98, 2015, doi: 10.1016/j.csl.2014.09.005.
- [87] G. Ciocca, S. Corchs, and F. Gasparini, “Complexity Perception of Texture Images,” in *Springer International Publishing Switzerland 2015*, 2015, pp. 119–126. doi: 10.1007/978-3-319-23222-5.
- [88] N. Hatami, Y. Gavet, and J. Debayle, “Classification of time-series images using deep convolutional neural networks,” *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 1241, no. c, 2018, doi: 10.1109/TNB.2018.2845103.
- [89] N. S. Kwak, K. R. Müller, and S. W. Lee, “A convolutional neural network for steady state visual evoked potential classification under ambulatory environment,” *PLoS One*, vol. 12, no. 2, pp. 1–20, 2017, doi: 10.1371/journal.pone.0172578.
- [90] M. Liu, W. Wu, Z. Gu, Z. Yu, F. Qi, and Y. Li, “Deep learning based on Batch Normalization for P300 signal detection,” *Neurocomputing*, p. to be published, 2017, doi: 10.1016/j.neucom.2017.08.039.
- [91] N. K. N. Aznan, S. Bonner, J. D. Connolly, N. Al Moubayed, and T. P. Breckon, “On the Classification of SSVEP-Based Dry-EEG Signals via Convolutional Neural Networks,” *IEEE Syst J*, 2018, [Online]. Available: <http://arxiv.org/abs/1805.04157>

- [92] E. Carabez, M. Sugi, I. Nambu, and Y. Wada, “Convolutional Neural Networks with 3D Input for P300 Identification in Auditory Brain-Computer Interfaces,” *Comput Intell Neurosci*, vol. 2017, 2017, doi: 10.1155/2017/8163949.
- [93] J. Ramírez-Quintana, J. Macias-Macias, A. Corral-Saenz, and M. Chacon-Murguia, “Novel SSVEP Processing Method Based on Correlation and Feedforward Neural Network for Embedded Brain Computer Interface,” in *Mexican Conference on Pattern Recognition*, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. A. Olvera-López, and J. Salas, Eds., Cham: Springer International Publishing, 2019, pp. 248–258.
- [94] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” *Adv Neural Inf Process Syst*, vol. 2017-Decem, no. Nips, pp. 3857–3867, 2017.
- [95] X. Jiang, Y. Wang, W. Liu, S. Li, and J. Liu, “CapsNet, CNN, FCN: Comparative performance evaluation for image classification,” *Int J Mach Learn Comput*, vol. 9, no. 6, 2019, doi: 10.18178/ijmlc.2019.9.6.881.
- [96] A. Fischer, “A special Newton-type optimization method,” *Optimization*, vol. 24, no. 3–4, pp. 269–284, 1992.
- [97] S. J. Wright, “Coordinate Descent Algorithms,” *Math Program*, vol. 151, no. 1, pp. 3–34, Feb. 2015, doi: 10.1007/s10107-015-0892-3.
- [98] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang, “A Unified Analysis of Stochastic Momentum Methods for Deep Learning,” *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2018-July, pp. 2955–2961, Aug. 2018, doi: 10.24963/ijcai.2018/410.
- [99] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming auto-encoders,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6791 LNCS, no. PART 1, pp. 44–51, 2011, doi: 10.1007/978-3-642-21735-7_6.
- [100] G. Hinton, S. Sabour, and N. Frosst, “Matrix Capsules with EM Routing,” in *ICLR 2018*, 2018, pp. 1–15. doi: 10.2514/1.562.
- [101] C. B. Do and S. Batzoglou, “What is the expectation maximization algorithm?,” *Nat Biotechnol*, vol. 26, no. 8, 2008, doi: 10.1038/nbt1406.
- [102] M. Hasani, A. N. Saravi, and H. Khotanlou, “An Efficient Approach for Using Expectation Maximization Algorithm in Capsule Networks,” 2019, [Online]. Available: <http://arxiv.org/abs/1912.05333>

- [103] J. A. Ramirez-Quintana, L. Madrid-Herrera, M. I. Chacon-Murguia, and L. F. Corral-Martinez, "Brain-Computer Interface System Based on P300 Processing with Convolutional Neural Network, Novel Speller, and Low Number of Electrodes," *Cognit Comput*, 2020, doi: 10.1007/s12559-020-09744-2.
- [104] M. Liu, W. Wu, Z. Gu, Z. Yu, F. F. Qi, and Y. Li, "Deep learning based on Batch Normalization for P300 signal detection," *Neurocomputing*, vol. 275, pp. 288–297, 2018, doi: 10.1016/j.neucom.2017.08.039.
- [105] B. Xu *et al.*, "Wavelet Transform Time-Frequency Image and Convolutional Network-Based Motor Imagery EEG Classification," *IEEE Access*, vol. 7, no. MI, pp. 6084–6093, 2019, doi: 10.1109/ACCESS.2018.2889093.
- [106] J. Zhang, C. Yan, and X. Gong, "Deep convolutional neural network for decoding motor imagery based brain computer interface," *2017 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2017*, vol. 2017-Janua, pp. 1–5, 2017, doi: 10.1109/ICSPCC.2017.8242581.
- [107] K. W. Ha and J. W. Jeong, "Motor imagery EEG classification using capsule networks," *Sensors (Switzerland)*, vol. 19, no. 13, 2019, doi: 10.3390/s19132854.
- [108] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband eeg signals using capsnet," *Sensors (Switzerland)*, vol. 19, no. 9, 2019, doi: 10.3390/s19092212.
- [109] J. A. Ramirez-Quintana, J. M. Macias-Macias, G. Ramirez-Alonso, M. I. Chacon-Murguia, and L. F. Corral-Martinez, "A novel Deep Capsule Neural Network for Vowel Imagery patterns from EEG signals," *Biomed Signal Process Control*, vol. 81, p. 104500, 2023, doi: <https://doi.org/10.1016/j.bspc.2022.104500>.
- [110] R. Mukhometzianov and J. Carrillo, "CapsNet comparative performance evaluation for image classification," pp. 1–14, 2018, [Online]. Available: <http://arxiv.org/abs/1805.11195>
- [111] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," vol. 2, pp. 1701–1704, 2010, doi: 10.1109/IJCNN.1993.716981.
- [112] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: what does it estimate and how well does it do it?," *J Am Stat Assoc*, vol. 0, no. ja, pp. 1–22, 2023, doi: 10.1080/01621459.2023.2197686.
- [113] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," in

- Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, 2020. doi: 10.1109/ICCMC48092.2020.ICCMC-00016.
- [114] B. M. Idrees and O. Farooq, “Vowel classification using wavelet decomposition during speech imagery,” in *3rd International Conference on Signal Processing and Integrated Networks, SPIN 2016*, 2016, pp. 636–640. doi: 10.1109/SPIN.2016.7566774.
- [115] M. Kwabena Patrick, A. Felix Adekoya, A. Abra Mighty, and B. Y. Edward, “Capsule Networks – A survey,” *Journal of King Saud University - Computer and Information Sciences*, no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.09.014.
- [116] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” pp. 1–15, 2014, [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [117] L. Støhle and S. Wold, “Analysis of variance (ANOVA),” *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, 1989, doi: [https://doi.org/10.1016/0169-7439\(89\)80095-4](https://doi.org/10.1016/0169-7439(89)80095-4).
- [118] J. M. Macías-Macías, J. A. Ramírez-Quintana, M. I. Chacón-Murguía, A. A. Torres-García, and L. F. Corral-Martínez, “Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network,” *Comput Biol Med*, vol. 159, p. 106909, Jun. 2023, doi: 10.1016/J.COMPBIOMED.2023.106909.
- [119] S. Dodd and J. Mills, “Phonetics and Phonology,” 1996, pp. 13–33.
- [120] B. Hayes, R. Kirchner, and D. Steriade, *Phonetically based phonology*. 2004. doi: 10.1017/CBO9780511486401.
- [121] P. Saha, M. Abdul-Mageed, and S. Fels, “SPEAK YOUR MIND! Towards Imagined Speech Recognition With Hierarchical Deep Learning,” *Interspeech 2019*, pp. 1–5, 2019, [Online]. Available: <http://arxiv.org/abs/1904.05746>
- [122] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, “EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density,” *Biomed Signal Process Control*, vol. 59, p. 101899, 2020, doi: 10.1016/j.bspc.2020.101899.
- [123] M. M. Islam and M. M. H. Shuvo, “DenseNet based speech imagery EEG signal classification using Gramian angular field,” *2019 5th International Conference on Advances in Electrical Engineering, ICAEE 2019*, pp. 149–154, 2019, doi: 10.1109/ICAEE48663.2019.8975572.

- [124] S. Martin *et al.*, “Word pair classification during imagined speech using direct brain recordings,” *Sci Rep*, vol. 6, 2016, doi: 10.1038/srep25803.
- [125] Y. Li, H. Luo, and X. Tian, “Mental operations in rhythm: Motor-to-sensory transformation mediates imagined singing,” *PLoS Biol*, vol. 18, no. 10, pp. 1–19, 2020, doi: 10.1371/journal.pbio.3000504.
- [126] Q. Rabbani, G. Milsap, and N. E. Crone, “The Potential for a Speech Brain–Computer Interface Using Chronic Electrocorticography,” *Neurotherapeutics*, vol. 16, no. 1, pp. 144–165, 2019, doi: 10.1007/s13311-018-00692-2.
- [127] J. Weaver, “How Brain Waves Help Us Make Sense of Speech,” *PLoS Biol*, vol. 11, no. 12, 2013, doi: 10.1371/journal.pbio.1001753.
- [128] O. B. Kolozsvári *et al.*, “Coherence Between Brain Activation and Speech Envelope at Word and Sentence Levels Showed Age-Related Differences in Low Frequency Bands,” *Neurobiology of Language*, vol. 2, no. 2, 2021, doi: 10.1162/nol_a_00033.
- [129] P. Sun and J. Qin, “Neural networks based EEG-Speech Models,” Dec. 2016, Accessed: May 08, 2023. [Online]. Available: <https://arxiv.org/abs/1612.05369v2>
- [130] R. A. Sharon and H. A. Murthy, “Correlation based Multi-phasal models for improved imagined speech EEG recognition,” 2020. doi: 10.21437/smm.2020-5.
- [131] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.
- [132] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.90.
- [133] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
- [134] A. Krizhevsky, G. E. Hinton, and I. Sutskenvet, “ImageNet Classification with Deep Convolutional Neural Networks,” pp. 1–9, 2012.
- [135] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, “Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features,” *J Neural Eng*, vol. 15, no. 1, 2018, doi: 10.1088/1741-2552/aa8235.

- [136] J. P. Montecinos, “Adquisición Y Desarrollo Del Lenguaje Y La Comunicación: Una Visión Pragmática Constructivista Centrada En Los Contextos,” *Límite*, vol. 7, pp. 54–66, 2000, [Online]. Available: <http://www.redalyc.org/pdf/836/83600704.pdf>
- [137] C. Díez, D. Pacheco, A. De Caso, J. García, and E. García, “El desarrollo de los componentes del lenguaje desde aspectos psicolingüísticos,” *Journal of Developmental and Educational Psychology*, vol. 2, no. 1, pp. 129–135, 2009.
- [138] T. Proix *et al.*, “Imagined speech can be decoded from low-and cross-frequency features in perceptual space,” *bioRxiv*, no. April, p. 2021.01.26.428315, 2021, doi: 10.1101/2021.01.26.428315.
- [139] R. D. Kent and F. D. Minifie, “Coarticulation in recent speech production models,” *J Phon*, vol. 5, no. 2, pp. 115–133, 1977, doi: 10.1016/s0095-4470(19)31123-4.
- [140] “OpenBCI - Open Source Biosensing Tools (EEG, EMG, EKG, and more),” 2016. <https://openbci.com/>
- [141] “RFDuino,” 2017. <http://docs.openbci.com/Hardware/02-Cyton>

ANEXO A. CREACIÓN DE BASE DE DATOS

En este trabajo se propone la aportación de una base de datos para habla imaginada, por lo que en este capítulo se detalla la composición de ésta y los protocolos con los que fue creada. El propósito de conformar una base de datos propia es analizar señales EEG bajo el mismo método de análisis con el que se analizaron las bases de datos de KaraOne y DaSalla, y con ello determinar una relación de características con las cápsulas de la red CapsVI. En la subsección A.1 se presenta la justificación de la creación de la base, en la subsección A.2 se detalla el protocolo de adquisición para obtener las señales, en la subsección A.3 se describe la creación de una interfaz cerebro computadora para la adquisición de señales EEG de habla imaginada, y, finalmente, en la subsección A.4 se detalla la composición de la base de datos.

A.1 Justificación

El habla imaginada se obtiene de la actividad cerebral en áreas que están relacionadas al procesamiento de lenguaje. Dichas áreas incluyen principalmente al área de Broca y Wernicke. Asimismo, recientemente, la literatura indica que también las áreas de la corteza premotora, el giro frontal inferior, el giro post-central, el giro temporal inferior, la corteza auditiva primaria y secundaria participan en la producción de lenguaje.

Sin embargo, actualmente se conoce de la complejidad que estudiar el fenómeno de habla imaginada conlleva. La información que proveen las áreas cerebrales involucradas en el proceso de lenguaje no se ha visto desde un punto de vista en el que para el área de procesamiento de señales se conoce como información a priori. La metodología que se ha seguido en el análisis de este fenómeno es similar a una caja negra llena de instrucciones, donde existe una adquisición de datos, una extracción de características y clasificación. No obstante, no se ha trabajado en caracterizar la información presente en señales EEG de SI.

Primero debemos de definir qué es lo que compone al lenguaje. Se conoce que el ser humano utiliza el habla articulada como la forma más natural y común de expresar y comunicar ideas o sentimientos. El habla está compuesta por palabras, las cuales se dividen en sílabas, que, a su vez, las cuales son unidades discretas de combinación de fonemas compuestos por una vocal y

consonantes. Las vocales son un segmento de sonido en el que no existe alguna obstrucción del tracto oral, mientras que las consonantes tienen un obstáculo en el flujo de aire.

La justificación de la creación de una base de datos de habla imaginada basada en vocales es que se toma como referencia el modelo de aprendizaje de habla de un humano. Un niño hablante de un idioma occidental primero aprende a pronunciar vocales, luego silabas por medio de la conjunción de vocales y consonantes para que, finalmente comience a pronunciar palabras [135]–[137]. Como se puede observar, el modelo de aprendizaje va de lo simple a lo complejo, siendo la articulación de vocales algo sencillo y la articulación de palabras algo complejo. Por esta razón, se propone la creación de la base de datos basada en vocales para generar el conocimiento suficiente de poder crear características a priori, teniendo en consideración la información que brinda la literatura, en la que se describe que al realizar una imaginación del habla se crean los SRP [42], los cuales son potenciales evocados relacionados a eventos que se generan en relación con el fonema que se imagina. De esta manera se pueden realizar clasificaciones con características a priori de las señales EEG de un sujeto.

En la Figura 2.1 se mostró un ejemplo de cómo se articulan las vocales en el idioma del inglés. Comúnmente, se representa de acuerdo con términos fisiológicos, específicamente la posición de la lengua en la boca y la forma de los labios, lo cual hace referencia a no generar obstrucciones en el tracto oral con diferentes posiciones de los músculos que son la lengua y los labios [138]. De esta forma, para cada una de las vocales la lengua se encuentra en una posición diferente, creando diferentes sonidos en el tracto oral en conjunto con la posición de los labios, los cuales se mostraron en la Figura 2.2, donde para las vocales /a/, /e/, e /i/ están abiertos, mientras que para las vocales /o/ y /u/ están redondeados.

De esta forma, en la literatura se ha creado un diagrama de articulación de sonidos con las diferentes combinaciones de posiciones que se pueden generar entre la lengua y los labios. Este diagrama se muestra en la Figura A.1, donde se puede observar que la lengua puede adoptar diferentes posiciones como lo es en frente, atrás, abajo arriba, redondeada y sin redondear. De esta manera se crean diferentes sonidos como son las vocales /a/, /e/, /i/, /o/, /u/, y los sonidos de /ɛ/, /α/ y /ɔ/.

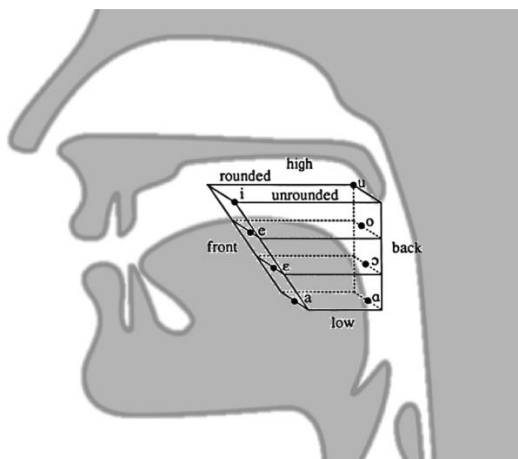


Figura A.1 Diagrama de articulación de sonidos en el tracto oral.

En el caso de la base de datos que se creará en el presente trabajo, se incluyen solamente las vocales /a/ y /u/ dadas las diferencias que existen en su articulación, puesto que se puede observar en el diagrama que ambas vocales están localizadas en extremos opuestos, y, según la lista de fonemas de Kent [139], son vocales con alta diferencia de articulación.

A.2 Protocolo de adquisición

Para la creación de la base de datos se elaboró un protocolo de adquisición basado en la base de datos de DaSalla, el cual se ejemplifica en la Figura A.2. Tal protocolo corresponde a la serie de 3 pasos donde su finalización se llamará sesión. Dicho esto, cada sesión de toma de muestras corresponde a las siguientes etapas:

1. En un monitor aparece una cruz por un tiempo de 2-3 segundos. El intervalo es aleatorio con el fin de evitar predisposición del usuario a la siguiente etapa. Además, este tiempo ayuda a evitar cualquier potencial relacionado a eventos que se pueda producir cuando la cruz apareció.
2. Aparece una de las tres imágenes a las que se les llamará tareas; labios redondos, labios abiertos o una cruz. Los labios redondos indican que el sujeto debe de comenzar la imaginación de la vocal /u/. Por otra parte, los labios cerrados indican la imaginación de la vocal /a/. la imagen de la cruz indica que el sujeto debe solamente de mantener la concentración en la pantalla del monitor sin realizar imaginación alguna. La duración de esta etapa es de 2.5 segundos.

3. Finalmente, la pantalla se vuelve completamente blanca durante 3 segundos, indicando al sujeto que puede tener un momento de relajación y puede moverse en caso de que lo desee. Al terminar esta etapa, concluye la sesión.

Para la creación de la base de datos, cada sujeto realiza 150 sesiones, 50 para cada una de las tareas. Aproximadamente cada sujeto le lleva un tiempo de 17.5 minutos llevar a cabo las 150 sesiones. Para la adquisición de señales se implementó este protocolo en una interfaz cerebro computadora basada en una Raspberry y el dispositivo de adquisición OpenBCI. Esta interfaz se explica en la siguiente subsección.

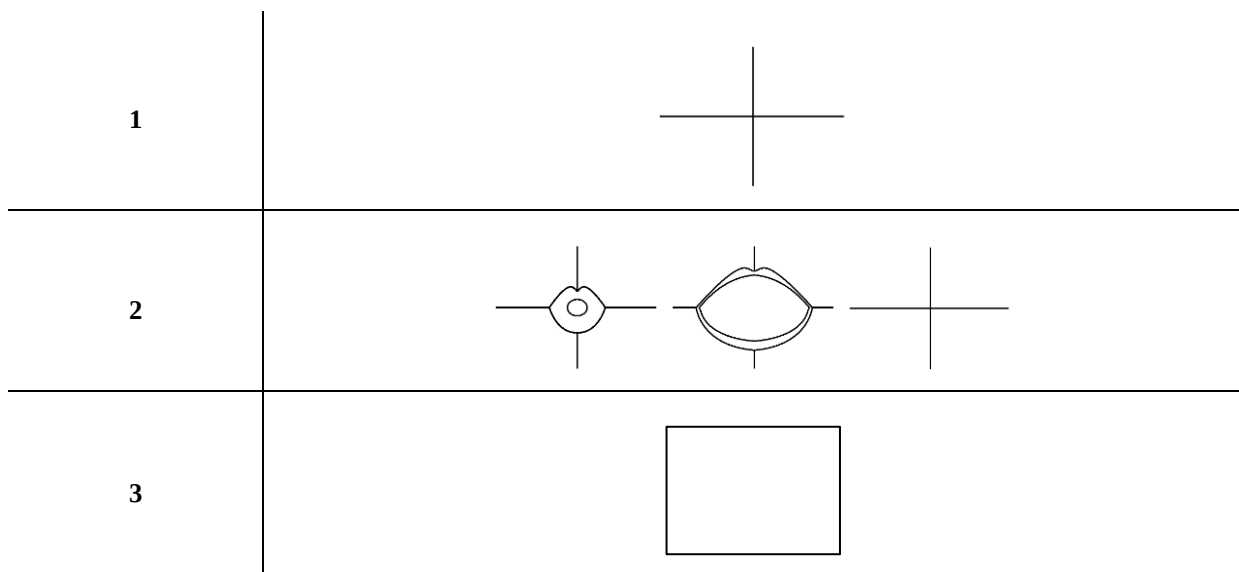


Figura A.2 Protocolo de adquisición de señales EEG.

A.3 BCI para la adquisición de señales EEG de habla imaginada.

Para la creación de la base de datos fue necesario crear una interfaz cerebro computadora e implementar el protocolo de adquisición descrito en la subsección A.2. Dicho sistema BCI se compone del dispositivo de adquisición Cython de OpenBCI y un sistema de cómputo RaspBerry Pi.

Cython es un dispositivo que se compone por dos tarjetas electrónicas: la tarjeta de adquisición y la tarjeta de comunicación, tal y como se observa en la Figura A.3. La tarjeta de adquisición (Figura A.3a) utiliza el microcontrolador PIC32MX250F128 con un ADC que realiza un muestreo de señales de 250 Hz [140] y se comunica a un sistema de computación vía inalámbrica mediante

la tarjeta de comunicación “Dongle” (Figura A.3b) que utiliza un módulo bluetooth 4.0 conocido como módulo de radio RFDuino [141].

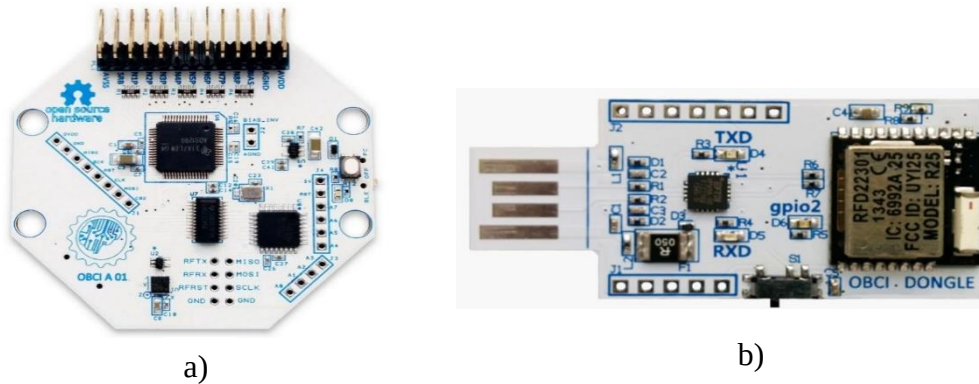


Figura A.3 Cython biosensing board [140].

En la tarjeta de adquisición es donde se conectan los electrodos. En este caso se utilizaron electrodos de copa, ya que son de los electrodos que presentan menor molestia a los usuarios, además de que existe evidencia en trabajos previos de su buen funcionamiento al adquirir señales. Asimismo, se utilizó gel Ten20[®] para acoplar impedancias entre el contacto de la piel y el mismo electrodo.

Se utilizaron 8 electrodos activos y 2 más, uno de referencia, y otro de tierra. La colocación de los electrodos se muestra a continuación:

Tabla A.1 Localización de electrodos en BCI para habla imaginada.

Canal	Localización
FC3	Corteza premotora
FC1	Corteza premotora
F5	Giro frontal Inferior
C3	Giro postcentral
F7	Área de Broca
FT7	Giro temporal Inferior
T7	Giro temporal medio - Corteza auditiva secundaria
C5	Área de Wernicke - Corteza auditiva primaria

Tal y como se puede observar, la tarjeta Cython está limitada al uso de 8 canales como máximo. La metodología de selección de canales se deriva del análisis de la base de datos en KaraOne en la subsección 4.1. Dicho análisis en el que se ponderaron los canales con base en el resultado del

análisis ANOVA, se excluyen los canales C4, Cz y P3 ya que fueron los canales con menor significancia en el análisis. Lo anterior es lógico, ya que estos canales corresponden fisiológicamente al giro postcentral del cerebro y es de esperar que su medias intragrupos sean similares. En cuanto al canal P3, corresponde al Lóbulo superior parietal, y de todas las áreas que se incluyeron en la selección de canales de KaraOne, este fue el que menor información aportó de los 11, por esa razón se excluyó en este estudio. Los canales seleccionados se muestran en la Figura A.4 con círculos azules, mientras los canales que se excluyeron se muestran con círculos rojos.

Tal y como se observa en la Figura A.4, los canales seleccionados corresponden totalmente al hemisferio izquierdo. La razón de esto se basa en que, típicamente, el lenguaje o producción del habla se encuentra dominado por el hemisferio izquierdo. Lo anterior lo indica una serie de investigaciones que se han hecho a lo largo del tiempo (Broca y Wernicke), sin embargo, se encontró hace poco que probablemente en personas que son zurdas, la producción del habla se encuentre dominada por el hemisferio derecho.

Podemos concluir que los canales fueron seleccionados de forma óptima ya que, en la experimentación de la base de datos de KaraOne, se inició con una ponderación de canales bajo el análisis estadístico de ANOVA utilizando los 62 canales. Sin embargo, no todas las características de los canales están asociadas con la producción del habla, por lo que las clasificaciones rondaban alrededor del 57% seleccionando 20 de los 62. Luego de ello, utilizando la información que provee la literatura, se seleccionaron aquellos canales que bajo el sistema internacional 10-10 estén directamente asociados con las áreas que los modelos más recientes de producción del habla indican. Seleccionando solamente 8 canales que corresponden no solo fisiológicamente a las áreas encargadas del procesamiento de lenguaje, sino que también fueron los canales más importantes en el análisis de ANOVA y sus características, se demostró bajo en análisis ANOVA y redes tipo cápsula que estos canales aportaban información relacionada con el habla, además de que las ponderaciones de ANOVA indicaban un nivel más alto de relación entre características de estos canales.

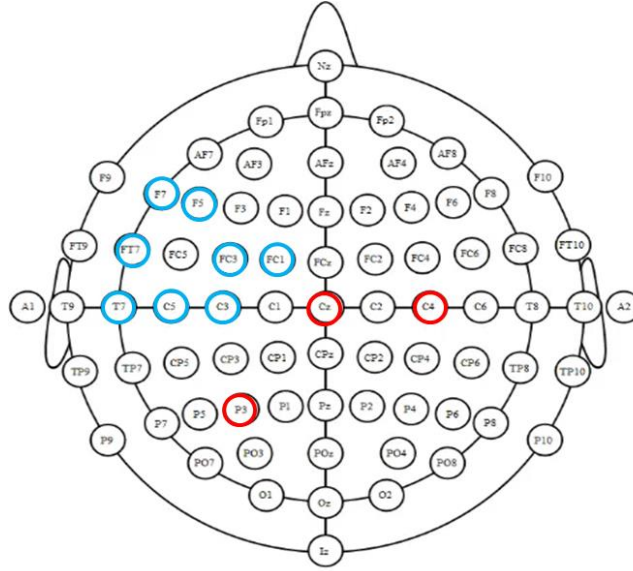


Figura A.4 Distribución de canales para la BCI de habla imaginada.

A.4 Composición de la base de datos

La base de datos se compone de 150 sesiones del protocolo diseñado para habla imaginada, 50 sesiones por cada tarea. Para la recolección de datos se situó a un sujeto en una silla cómoda a una distancia de 1 m hacia el monitor. Las condiciones ambientales de la habitación donde se encontraba son: temperatura a 20°, sin calefacción o dispositivos generadores de ruido, completamente en oscuridad para enfocar al sujeto solamente en el monitor, y puertas y ventanas cerradas con el fin de evitar ruidos externos que pudieran interferir en la imaginación.

El sujeto 1, corresponde a una persona de sexo masculino de 26 años, no cuenta con ingesta actual de algún medicamento, además de que no utiliza lentes o ha tenido cirugía de corrección de vista. Durante las 150 sesiones de toma de muestras, el sujeto no reportó algún cansancio, ni algún síntoma de adormilamiento. De esta forma, el progreso de la actual base de datos se compone de 150 señales EEG referentes a las 150 sesiones del protocolo de habla imaginada.

Donde cada señal $s_c(n)$ se compone de 625 muestras dado que la frecuencia de muestreo del dispositivo de adquisición es de 250 Hz, y la duración de la tarea de habla imaginada es de 2.5 segundos. De esta manera $s_c(n)$ contiene solamente las muestras en las que el sujeto realizaba la imaginación de una vocal o de la tarea de control. Cada una de las señales fue guardada en formato .csv y contiene un arreglo de 9x625, donde las primeras 8 filas corresponden a los 8 canales activos,

y la última fila hace referencia al número de tarea que apareció en el monitor durante la imaginación, lo cual corresponde a la etiqueta.