Universidad Veracruzana

Instituto de Investigaciones en Inteligencia Artificial

Tesis de Maestría

Aprendizaje y Reconocimiento de Fonemas del Español Mexicano a partir del Movimiento de Labios en Imágenes Digitales

Que para obtener el grado de

Maestra en Inteligencia Artificial

Autor:

Lic. Nora Esmeralda Cancela García

Director:

Dr. Homero Vladimir Ríos Figueroa

Xalapa - Enríquez, Ver., agosto 2024

Contenido

Tablas

Figuras

Agradecimientos

A Dios.

A mi esposo e hijos.

A mis padres, hermana, hermano, sobrinas.

A mi director y maestro, Dr. Homero Ríos por su apoyo, guía y confianza.

A mis revisores, gracias por enriquecer este trabajo con sus correcciones y recomendaciones.

Amigos, Dr. Efrén Mezura y M. en I.A. Roberto Cruz gracias por motivarme y su apoyo.

1. Introducción

1.1 Antecedentes

El ser humano es un ser social, ha vivido siempre en comunidad, la comunicación ha sido un factor importante en su desarrollo. El medio principal para la comunicación en nuestra sociedad es el lenguaje, siendo la expresión oral la principal forma de comunicarnos.

El reconocimiento automático del habla (ASR - automatic speech recognition) es un área de interés desde la década de los 50's, ha evolucionado y se han obtenido mejores resultados a medida que ha evolucionado la tecnología. Estos avances han permitido la creación de aplicaciones mediante las cuales podemos comunicarnos con una computadora u otros dispositivos tal es el caso de Cortana, Alexa, Siri. Estas formas de comunicación se basan en la interpretación del sonido de la voz. Para tener buenos resultados estas aplicaciones generalmente funcionan en ambientes controlados donde el ruido es casi nulo. Existen investigaciones como la de McGurk y MacDonald [1] que demuestran que la percepción del habla implica tanto información auditiva como visual, ya que la información visual permite captar mejor el mensaje. McGurk demuestra que el cerebro integra señales de distintas modalidades en una única representación perceptiva, en sus experimentos percibe que cuando hay un desfase en las señales de video con las de sonido, las personas tienen dificultades para entender el mensaje, con esto demuestra que la información visual de los movimientos de los labios juega un papel importante en el reconocimiento del mensaje al comunicarnos.

En la actualidad hay investigaciones que muestran que los sistemas de reconocimiento automático del habla multimodal (MSR - multimodal speech recognition), buscan un equilibrio entre la información acústica y visual, permitiendo el desarrollo de sistemas más robustos que mejoran el rendimiento de los sistemas de reconocimiento del habla en espacios donde los resultados del procesamiento exclusivamente de audio no brindan buenos resultados. De ellos el área más abordada ha sido la que incorporan audio e información visual, llamada reconocimiento automático del habla audiovisuales (AV-ASR Audio-Visual Automatic Speech Recognition)

En ambientes donde existen ruido o en aquellos casos donde la dicción del hablante no es muy clara integrar la información visual mediante la lectura de labios ha dado grandes resultados.

Existe en la actualidad un gran interés en el desarrollo de aplicaciones que permitan la Lectura Automática de Labios (ALR -Automatic Lip Reading). Dentro de la

inteligencia artificial el área de reconocimiento de patrones explora la aplicación de varios algoritmos de aprendizaje automático buscando mejorar los resultados en la lectura labial.

Cuando nos comunicamos de forma oral expresamos frases (oraciones), que están formadas por palabras y estas a su vez por caracteres, para el caso del procesamiento del habla la unidad mínima no es el carácter sino el fonema. El fonema es definido como el sonido mínimo distinguible que puede cambiar el significado de una palabra [2]. Para analizar la información visual se utiliza el visema, entendida como la unidad mínima distinguible del habla en el análisis de video y es la descripción visual de un fonema, estos definen rasgos de la cara y boca cuando una persona se encuentra hablando [3]

Actualmente existen varios conjuntos de datos (bases de datos audiovisuales) que se han utilizado en trabajos de reconocimiento del habla, como las descritas por Adriana Fernández y Federico Sukno [4], sin embargo todas las mencionadas aquí son para el lenguaje inglés y otros idiomas. Ronquillo [5] usa conjunto de datos de RTVE para desarrollar un sistema reconocimiento automático del habla en español, usando lectura labial. El Conjunto de datos RTVE fue creado por la Universidad de Zaragoza en colaboración con la Corporación Radio Televisión Española, a partir de una colección de programas emitidos por la televisión pública española (RTVE) entre 2018 y 2019 [6]. Aparte de RTVE, son escasos los conjuntos de datos audiovisuales publicados para el idioma español.

Existe un gran número de aplicaciones que se pueden dar a los sistemas MSR, entre las que podemos mencionar ayudar a personas sordas o con problemas de audición a comunicarse sin tener la necesidad de saber leer los labios. Personas con problemas de salud y dificultades para comunicarse en los hospitales, les permitiría interactuar con personal y familiares. Y aquellas personas que requieren terapias del lenguaje para poder comunicarse.

1.2. Planteamiento del Problema

Según datos del INEGI en México en el censo 2020 un 16.5% de la población posee algún tipo de discapacidad, de ellos el 15.7% tiene problemas para hablar y comunicarse y el 24.4% para oír [7]. El tema de inclusión de las personas con discapacidad es una tarea en todos los ámbitos a nivel mundial según lo estable la estrategia para la inclusión definida por organización de las naciones unidas y la organización mundial de la salud. Integrar a las personas con alguna discapacidad que le impide comunicarse es un gran reto de los gobiernos y la sociedad. Desde el punto de vista tecnológico y la computación se están haciendo esfuerzos en el

desarrollo de sistemas para el reconocimiento automático del lenguaje multimodal. El audio ha sido una de las áreas más trabajadas y con grandes avances de reconocimiento de frases, ha logrado disminuir mucho los márgenes de error. Sin embargo, la lectura labial es un problema de mayor complejidad ya que las investigaciones han demostrado que hay características no visibles o difíciles de identificar que participan en el proceso de producción de fonemas (sonidos) que permiten construir las palabras con que nos comunicamos.

La lectura automática de labios no solo apoya la comunicación de las personas con discapacidad definitiva del lenguaje, sino a aquellas que por algún problema de salud presentan de manera temporal dificultades de comunicación oral, además de robustecer el reconocimiento de frases en ambientes ruidosos usando canales de reconocimiento de audio y visuales.

El presente trabajo propone la aplicación de un algoritmo de clasificación para la lectura labial y el reconocimiento de fonemas en español utilizados por el método Adryna [8]. Este método es empleado por algunos terapeutas del lenguaje para enseñar a personas con discapacidad del habla, por ejemplo, aquellos con algún nivel de espectro autista. El alfabeto utilizado consta de 21 fonemas.

1.3. Hipótesis

¿ Es posible realizar la clasificación de fonemas del idioma español a partir de análisis automático de video digital?

1.4. Objetivo General

Desarrollar un algoritmo de clasificación para el reconocimiento automático de fonemas del español utilizando información visual.

1.5. Objetivos Específicos

Para lograr el objetivo general propuesto en este trabajo se plantean los siguientes objetivos específicos:

- Construir un conjunto de datos audiovisual con información de audio, video RGB y profundidad de fonemas del método Adryna.
- Identificar la región de interés y realizar la extracción de características.
- Implementar un algoritmo de clasificación.

- Evaluar el modelo de aprendizaje y realizar el análisis de los resultados.

1.6. Justificación

En la actualidad, la lectura labial ha sido ampliamente abordada. El reto de detectar lo que dice una persona a partir de fotogramas de vídeo es un trabajo que aún no logra los resultados como los alcanzados con en el procesamiento de audio.

Hay desarrollos de sistemas para la lectura automática de labios para varios idiomas como inglés, alemán, chino, coreano, turco, y pocos son los trabajos que se encuentran para el español. Además, se sabe que incluso en México la forma de pronunciar es diferente, los estudiosos del lenguaje lo clasifican en tres regiones, español del norte, centro y sur. En la búsqueda realizada no encontramos un dataset ni trabajos de investigación de reconocimiento de fonemas del español mexicano.

La mayoría de las investigaciones actuales de lectura labial usan redes neuronales profundas para la identificación y reconocimiento de automática del habla, sin embargo, estos modelos requieren una gran cantidad de datos. Este trabajo propone la construcción de un data set audiovisual de fonemas del método Adryna para el idioma español mexicano y el desarrollo de un algoritmo de clasificación de los fonemas.

En la actualidad se pueden distinguir dos enfoques para la lectura labial, un enfoque ve este problema como una tarea de clasificación basada en palabras o frases, intentando a partir de video predecir una palabra o frase. El otro enfoque y más reciente se basa en la predicción de secuencias de caracteres o una secuencia de visemas para la construcción de palabras o frases [9].Construir palabras a través del reconocimiento de fonemas es de menor complejidad y por consiguiente la cantidad de cómputo requerido para la tarea de clasificación es menor. Este último enfoque es el considerado en el presente trabajo.

1.7 Preguntas de Investigación

El presente trabajo busca responder a las siguientes preguntas:

- ¿Es posible realizar identificación de fonemas del español usando información visual obtenida de una secuencia de frames de video?
- ¿La distancia de similitud resultado de aplicar el algoritmo DTW (Dynamic time warping) es una medida útil para hacer reconocimiento de fonemas usando el algoritmo KNN?

- ¿El algoritmo KNN es útil en la clasificación de fonemas con un conjunto de datos con centenas de secuencias (504 videos) ?

1.8 Variables de la Investigación

1.8.1 Variables Dependientes

Las variables dependientes de esta investigación son los grupos de fonemas con los que se van a realizar las pruebas del algoritmo de clasificación.

1.8.2 Variables Independientes

Las variables independientes en esta investigación son la secuencia de imágenes que se obtienen de cada persona para determinar que fonema pronunció.

1.9 Contribución

Aunado a la revisión del estado del arte de la lectura automática de labios, la presente investigación contribuye con:

- La creación de una Conjunto de Datos Audiovisual para el idioma español mexicano, basando en los fonemas del método Adryna.
- El reconocimiento de fonemas usando el algoritmo KNN y la distancia de similitud obtenida de dos secuencias de video con el algoritmo DTW, para un conjunto de datos pequeño, obtenido resultados que compitan con descritos en la literatura.

2. Trabajos Relacionados

En los últimos años se ha incrementado el interés por la lectura automática de labios. Las estadísticas de bases de datos como scopus y web of science muestran un incremento en los documentos relacionados con la investigación en lectura labial. La tabla 1 muestra algunos de los trabajos relacionados con el tema. Podemos observar que la mayoría de los conjuntos de datos utilizados son para el idioma inglés, y por lo general usan como técnica de clasificación alguna arquitectura basada en redes neuronales.

Por el tamaño del conjunto de datos usado para entrenar el modelo y la tarea de reconocimiento en el presente trabajo compararemos los resultados obtenidos con el trabajo de Adriana Fernández y Federico Sukno [10] quienes proponen un modelo basado en redes neuronales convolucionales; el conjunto de datos que usan es VLRF integrada por 24 hablantes que pronunciaron 25 oraciones cada uno, el conjunto de datos proporciona etiquetas de 31 fonemas.

Otro de los trabajos seleccionados es el de Stavros Petridis et al. [11] quienes proponen el uso de redes de memoria larga y corta (LSTM) para conjuntos de datos pequeños, dentro de los experimentos que realizan hacen el reconocimiento de caracteres usando los conjuntos de datos AVLetters y AVLetters2. AVLetters2 [12] es un conjunto de datos del alfabeto del idioma inglés, donde participan 5 personas quienes hacen 7 repeticiones de cada letra. Y AVLetters [41] es un conjunto de datos de 1998, primera versión del anterior donde participaron 10 personas haciendo 3 repeticiones de cada letra.

Finalmente consideraremos el trabajo de Randa El-Bialy et al. [9] quienes hacen el reconocimiento de palabras a través del reconocimiento de fonemas individuales utilizando una red neuronal convolucional y el conjunto de datos BBC Lip Reading Sentences 2 (LRS2). LRS2 es un conjunto de datos con más de 46,000 vídeos, extraídos de la BBC de 6 programas de televisión diferentes, con una gama de posiciones faciales que van desde la frontal hasta el perfil.

Titulo	Año	Idioma	Conjunto de Datos	Técnica de Clasificación	Tarea de Reconocimiento
Lip Reading Sentences Using Deep Learning with Only Visual Cues [13]	2020	Inglés	BBC LRS2	Red Neuronal Convolucional	Palabras y Frases
Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition [14]	2020	Inglés	LRW LRS3-TED	DCNN	Palabras y Frases
Lipreading with DenseNet and resBi-LSTM [15]	2020	Chino Mandarín	Conjunto de datos construido	DCNN	Palabras
End-to-end visual speech recognition for small-scale datasets [11]	*2020*	*Inglés*	*OuluVS2 , AVLetters, CUAVE y AVLetters2*	*Redes de memoria larga y corta (LSTM)*	*Palabras y caracteres*
Speaker-Independent Speech Recognition using Visual Features [16]	2020	Inglés	MIRACLVC1	CNN	Palabras
Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition [17]	2021	Inglés	VISWa (dígitos en Ingles)	Red Neuronal Artificial (RNA) Máquina de vectores de soporte (SVM) Naive Bayes (NB)	Dígitos
Using the Hand Preceding Model for Multi-Modal Fusion in Automatic Continuous Cued Speech Recognition [18]	2021	Frances Inglés	Conjunto de datos construido	CNN Modelo Oculto de Márkov de múltiples flujos	Vocales y Consonantes
Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions [19]	2021	Griego e Inglés	LRGW10 - Griego LRW500 - Ingles	CNN	Palabras
End-to-End Lip-Reading Without Large-Scale Data [10]	*2022*	*Español*	*VLRF*	*CNN*	*Fonemas*
Lip Reading Multiclass Classification by Using Dilated CNN with Turkish Dataset [20]	2022	Turco	Conjunto de datos Construido	DCNN/CNN	Frases y palabras
Turkish lip-reading using Bi-LSTM and deep learning models [21]	2022	Turco	Conjunto de datos Construido	CNN / Bi-LSTM	Palabras y Oraciones
Deep Learning-Based Approach for Arabic Visual Speech Recognition [22]	2022	Árabe	Conjunto de datos Construido	DCNN	Palabras y Dígitos
Developing phoneme-based lip-reading sentences system for silent speech recognition [9]	*2023*	*Inglés*	*BBC LRS2*	*CNN*	*Fonemas*
Application of deep learning in Mandarin Chinese lip-reading recognition [23]	2023	Chino Mandarín	Conjunto de datos construido	CNN	Palabras
Efficient DNN Model for Word Lip-Reading [24]	2023	Inglés / Japones	LRW, OuluVS, CUAVE y SSSD	CNN	Palabras
Visual Lip-Reading for Quranic Arabic Alphabets and Words Using Deep Learning [25]	2023	Árabe	AQAND Construido	CNN	Caracteres y Palabras

Tabla 1 Algunos trabajos relacionados con lectura labial en los últimos años. El análisis mostrado en la tabla es producto de la revisión de literatura.

3. Marco teórico

3.1. Reconocimiento Automático del Habla (ASR)

El reconocimiento automático del habla (ASR) ha sido una actividad que se encuentra en investigación desde hace 5 décadas, consiste en convertir el lenguaje hablado en texto legible por una computadora [26]. El área de ASR está estrechamente relacionada con la lingüística computacional, debido a su estrecha relación con el lenguaje natural y la fonética por la gran variedad de sonidos que puede producir el ser humano.

Un sistema ASR puede describirse como: dada una entrada de muestras de audio *X* de una señal de voz grabada, se aplica una función *f* para asignarla a una secuencia de palabras *w* que representan la transcripción de lo que se dijo [26].

$$w = f(x)$$

La arquitectura típica de un sistema ASR se ilustra en la figura 1. Podemos observar que un sistema ASR tiene cuatro componentes principales: procesamiento de señales y extracción de características, modelo acústico, modelo lingüístico y búsqueda de hipótesis [27] .

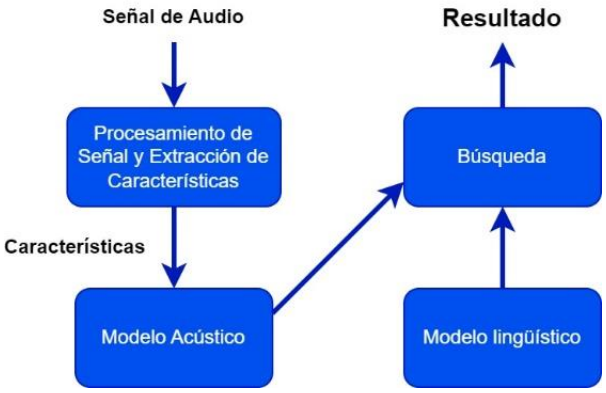


Fig. 1 Arquitectura Típica de un Sistema ASR. La imagen fue diseñada para mostrar los elementos de una arquitectura típica de un sistema ASR

La señal de audio es recibida por el módulo dedicado al procesamiento de la señal y extracción de características, éste mejora la señal eliminando ruidos y distorsiones, convierte la señal del dominio temporal al dominio frecuencial y extrae vectores de características. El modelo acústico integra conocimientos sobre acústica y fonética, toma como entrada las características generadas y produce una señal de salida para la secuencia de características de longitud variable. El modelo lingüístico estima la probabilidad de una secuencia de palabras hipotética. El módulo de búsqueda combina las salidas del modelo acústico y modelo lingüístico dada la secuencia del vector de características y la secuencia de palabras hipotética y emite

la secuencia de palabras con la puntuación más alta como resultado del reconocimiento [27].

3.2. Reconocimiento Automático del Habla Audio-Visual (AV-ASR)

El reconocimiento automático del habla audio-visual (AV-ASR), también conocido como reconocimiento del habla basado en la modalidad dual, busca mejorar la precisión de clasificación aprovechando eficazmente las ventajas de ambas modalidades de señales [28]. De la señal de video se obtiene la información visual como movimientos de los labios y rasgos faciales.

Un sistema AV-ASR consta de un canal de flujo visual y un canal de flujo de audio, un módulo que fusiona las características extraídas de ambas señales. De la señal de video se identifica la región de interés (ROI), se transforma generalmente el video en imágenes y de ellas se extraen las características visuales [29]. En la figura 2 se muestran los bloques que forman un sistema AV-ASR.

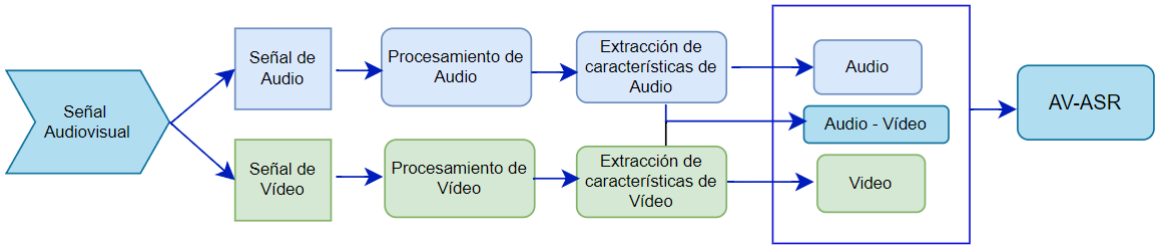


Fig. 2 Diagrama de un Sistema de Reconocimiento Automático del Habla Audio-Visual.

El interés por el desarrollo de sistemas AV-ASR han tenido un gran auge. A partir de 2019 se ha incrementado considerablemente las citas en esta área. A pesar de ello, existen grandes retos como [29] [30]:

- La construcción de bases de datos audiovisuales de dominio público para los diferentes idiomas.
- La selección y extracción de características visuales.
- La forma de combinar o integrar las características de Audio y Video para lograr mejorar el rendimiento.

Los sistemas AV-ASR en la actualidad son más precisos y generalizan mejor, sin embargo, siguen siendo computacionalmente costosos. Lo anterior aunado a los retos antes mencionados y su estrecha relación el procesamiento de imágenes y la visión por computadora ha provocado un gran interés entre los investigadores de estas áreas de la Inteligencia artificial.

3.3. Lectura Automática de Labios (ALR)

La lectura automática de labios (ALR) es una técnica que extrae información del movimiento de los labios de una persona mientras esta se encuentra hablando y determina las letras, palabras o frases que se están pronunciando [4], [13], [31], [32].

La estructura general de un sistema ALR consta principalmente de 3 módulos [4], como se muestra en la figura 3, el primer módulo realiza la detección de rostro y labios, posteriormente se encuentra el módulo que realiza la extracción de características visuales y finalmente se encuentra el módulo que realiza la clasificación.

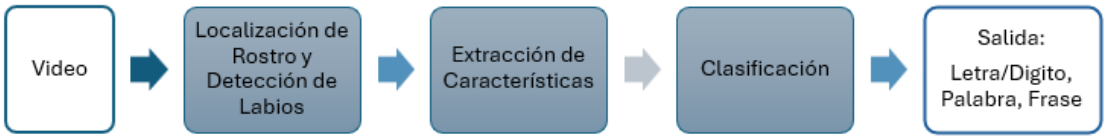


Fig. 3 Estructura general de un sistema ALR

Los sistemas de ALR tienen grandes retos en la actualidad, uno de los principales es la ambigüedad visual causado por los homofonemas, estos son caracteres que se confunden porque los movimientos labiales son iguales o muy similares, por ejemplo /p/,/b/ y /m/ [4].

Para los sistemas ASR el elemento mínimo distinguible es el fonema. Algunos investigadores utilizan el concepto de visema [33], concebida como la unidad mínima distinguible del habla en un video. Debido a que varios fonemas producen movimientos labiales que no se pueden distinguir implica que no existe una correspondencia uno a uno entre fonemas y visemas. Tal es el caso de los fonemas /b/ y /p/ ya que la vocalización se produce en un área no visible en un video. Otro de los problemas se presenta en aquellos fonemas donde la posición de la lengua cambian el aspecto visual, tal es el caso de la consonantes velares por ejemplo /g/ o /k/ [4], [10]. Fernández y Sukno [4] definen como uno de los principales retos que tiene los sistemas de ALR es diseñar sistemas robustos frente a las ambigüedades visuales.

Aunado al gran reto anterior, tenemos que los movimientos de los labios varían entre personas, idioma, región, contexto. Otros retos son los que se presentan en el área de visión por computadora como son la iluminación, que el objeto a reconocer no se vea por completo en el video, entre otros.

3.4. Conjuntos de Datos

Para obtener buenos resultados en cualquier problema donde se aplica alguna técnica de inteligencia artificial, los datos juegan un papel muy importante y son en gran medida la clave del éxito.

Como se mencionó anteriormente uno de los retos en la construcción de sistemas de reconocimiento audiovisual son el escaso número de bases de datos audiovisuales públicas existentes. Esto se debe en gran medida a la complejidad del lenguaje, a la variedad de idiomas y lenguas y a los intereses de cada uno de los investigadores, sumado a esto, no existe un criterio único aceptado para evaluar los conjuntos de datos audiovisuales [4] [30].

La cantidad de vocabulario grabado, la calidad del vídeo, la iluminación de la zona de pruebas, la orientación de la cabeza, los temas de las grabaciones y la frecuencia con la que los participantes reproducen el sonido, el vocabulario, la resolución, etc., son aún temas de gran interés entre los investigadores[30].

Algunas de las bases de datos audio visuales más citadas para el reconocimiento de dígitos y alfabetos se muestra en la tabla 2:

Nombre	Idioma	Año	Tarea	Participantes	Repeticiones por participante
AVDigits [34]	Ingles	2018	Dígitos	53	9
AVLetters2 [12]	Inglés	2008	Alfabeto	5	7
AVOZES [35]	Inglés	2004	Dígitos	20	-
AVICAR [36]	Inglés	2004	Alfabeto /Dígitos	86	5
AV@CAR [37]	Español	2004	Alfabeto /Dígitos	20	-
CUAVE [38]	Inglés	2004	Dígitos	36	-
BANCA [39]	Múltiple	2003	Dígitos	208	-
XM2VTS [40]	Inglés	1999	Dígitos	295	4
AVLetters [41]	Inglés	1998	Alfabeto	10	3

Tabla 2 Bases de Datos Audiovisuales para reconocimiento de Dígitos y Alfabetos. El análisis mostrado en la tabla es producto de la revisión de literatura.

Podemos observar que para el español solo se ha reportado una base datos audiovisual para el reconocimiento de alfabeto y dígitos es Av@car. Existen otras bases de datos para el reconocimiento de palabras o frases, particularmente para el idioma español se encontró VLRF con 24 participantes publicada en 2017 [42], RTVE construida a partir de videos de varios programas de televisión [43]. Debido a que en la presente propuesta se construyó una base de datos audiovisual para un grupo de fonemas no se profundizará en las bases de datos creadas para el reconocimiento de palabras o frases. Independientemente de tratarse de una base de datos audiovisual para el reconocimiento de alfabeto, dígitos, palabras o frases, en la búsqueda se encontró que en su mayoría están enfocadas a un solo idioma y el predominante es el inglés.

3.5. Algoritmos de Clasificación

El área de reconocimiento de patrones se encarga de descubrir de forma automática regularidades en los datos mediante el uso de algoritmos computacionales y la utilización de estas regularidades para clasificar los datos en diferentes categorías. Bishop [44] propone abordar el tema de clasificación como un enfoque de aprendizaje automático en el que un conjunto, denominado conjunto de entrenamiento, se utiliza para ajustar los parámetros de un modelo (algoritmo de aprendizaje), el algoritmo se puede expresar como una función y(x) que se determina durante la fase de entrenamiento a partir de los datos de entrenamiento. Una vez entrenado el modelo se puede pasar un nuevo conjunto de datos, denominado de prueba con los que se puede evaluar su capacidad de generalizar, es decir su capacidad de categorizar correctamente los datos del conjunto de prueba.

Los problemas de clasificación tienen el objetivo de asignar cada vector de entrada a un número finito de categorías discretas [44]. De una manera más formal el objetivo de los algoritmos de clasificación es encontrar una correspondencia entre las entradas x a las salidas y, donde $y \in \{1, \ldots, C\}$, siendo C el número de clases. Si C = 2, se conoce como clasificación binaria; si C > 2, se denomina clasificación multiclase [45].

Cuando hablamos de algoritmos de aprendizaje automático, encontramos que estos se clasifican en [44], [45]:

> *Aprendizaje supervisado*: para cada elemento de entrada del conjunto de entrenamiento se tiene la salida correspondiente. Cuando la salida es categórica el problema es de clasificación o reconocimiento de patrones, cuando la salida es real el problema es de regresión.

Aprendizaje no supervisado: los datos del conjunto de entrenamiento no tienen valor correspondiente (o etiqueta) de salida. Este tipo de aprendizaje busca encontrar similitudes en los datos y agruparlos en categorías de forma automática.

Aprendizaje por Refuerzo: el problema consiste en encontrar las acciones adecuadas en una situación dada para maximizar una recompensa. Descubre los resultados óptimos a través de prueba y error.

3.5.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN – K vecinos más cercanos) es un algoritmo de aprendizaje automático supervisado ampliamente utilizado en problemas de clasificación y regresión. La idea general del algoritmo es buscar los K puntos más cercanos a un punto específico e infiere su valor (figura 4). Este algoritmo asume que puntos de datos cercanos tienden a tener la misma etiqueta, es decir, comparte similitud en sus características [45]. Para poder conocer los K puntos más cercanos se usan métricas para el cálculo de la distancia, la más comúnmente utilizada es la distancia euclidiana.

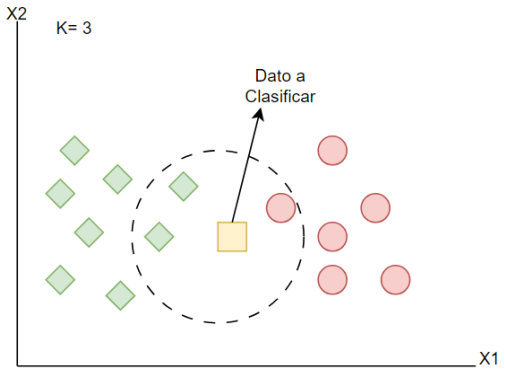


Fig. 4 Ilustración del algoritmo KNN para K=3, el dato a clasificar se asignaría al grupo de datos de la etiqueta verde. Imagen creada para ejemplificar el algoritmo KNN

Las etapas del algoritmo KNN son:

- Elegir el valor de K, es decir, el número de vecinos que se consideraran para clasificar el nuevo dato.
- Calcular la distancia que existe entre el nuevo dato con respecto a todos los datos que forman el conjunto de entrenamiento.

- Ordenar las distancias, y seleccionar los K datos con distancias más pequeñas. A este grupo de datos se le denomina K vecinos más cercanos al nuevo dato.
- El nuevo dato se clasifica según la etiqueta más frecuente de los K vecinos.

El algoritmo es sencillo y funciona correctamente siempre que se cuente con una buena métrica para el cálculo de la distancia y se seleccione un valor de K adecuado. El algoritmo es computacionalmente costoso cuando el conjunto de datos es muy grande.

El valor de K tiene un gran impacto en el comportamiento del modelo. Si K = 1, el método no comete errores en el conjunto de entrenamiento, el método no funcionara correctamente al momento de predecir los datos nuevos. A medida que K aumenta, las predicciones mejoran hasta que, en el límite de K = N, se predice la etiqueta mayoritaria de todo el conjunto de datos. A medida que K aumenta, la tasa de error en el conjunto de entrenamiento también. Se puede obtener un error mínimo en el conjunto de entrenamiento utilizando K = 1, pero el modelo sólo estaría memorizando los datos, se dice que el modelo se sobre ajusta y para valores de K grandes el modelo se sub ajusta, ambos casos afectan la capacidad de generalización del modelo. Para resolver el problema de la elección de K se utiliza la validación cruzada [45].

La idea general de la validación cruzada es dividir los datos de entrenamiento en K grupos (folds), entonces para cada grupo K $\in$ {1, . . . K}, entrenamos con todos los grupos menos el K-ésimo, el cual es usado para probar el modelo. En la figura 5 se ilustra la validación cruzada [44], [45].

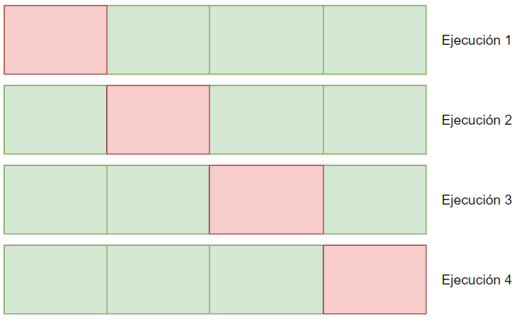


Fig. 5 Ilustración de validación cruzada, para k = 4, en cada ejecución se utilizará 3 grupos de datos para entrenamiento (color verde) y 1 para prueba (color rojo). Imagen creada para ejemplificar la validación cruzada.

4. Descripción de la Propuesta

4.1. Fonemas del Alfabeto

La lingüística es un campo científico que estudia el lenguaje y los principios de comprensión y producción del lenguaje. Sus campos de estudio son: la fonética, morfología, fonología, sintaxis, semántica, lexicografía, pragmática, sociolingüística, dialectología, lingüística histórica, psicolingüística y neurolingüística [46].

La fonética es el estudio de los sonidos de las lenguas, su articulación, características físicas y acústicas y la Fonología es el estudio de cómo los sonidos se relacionan como partes de un sistema. Los lingüistas usan el concepto de "dialectos" para distinguir entre dos o más variantes de un idioma. El idiolecto es el conjunto de características que conforman la particularidad lingüística de una persona, que incluye su vocabulario, gramática y pronunciación. Por ejemplo, el español de las Islas Canarias posee características particulares que no se encuentran en el español de Castilla-La Mancha, o el español mexicano. La variación del lenguaje regional, contextual, temporal y social son los cuatro tipos estudiados en la lingüística en general [47]. Los fonólogos son lingüistas que estudian la fonética y la fonología.

El Alfabeto Fonético Internacional (AFI) es un sistema que proporciona representaciones de los sonidos presentes en cualquier expresión oral. En este alfabeto a los fonos individuales se llaman fonemas y se escriben entre barras, por ejemplo /k/. Los fonemas se forman en la boca y participan diferentes órganos algunos de ellos se mueven como los labios, la lengua (el ápice, la lámina, y el dorso), el velo, la úvula, y las cuerdas vocales, y otros no se mueven como la nariz, la cavidad nasal, el alvéolo, el paladar, la cavidad bucal, y la cavidad laríngea [11]. La figura 6 muestra gráficamente los principales elementos que integran el aparto del habla y participan en la generación de sonidos.

Cada palabra en español está compuesta por letras que representan sonidos del habla. El español tiene 27 sonidos (cinco vocales más veintidós consonantes), con ellos se pueden articular los diferentes fonemas que componen su sistema fonológico [48]. El número de fonemas en español varía según la región, en [12] se describen 24 fonemas sin embargo se menciona que la mayoría de las regiones usa de 22 a 23 de ellos.

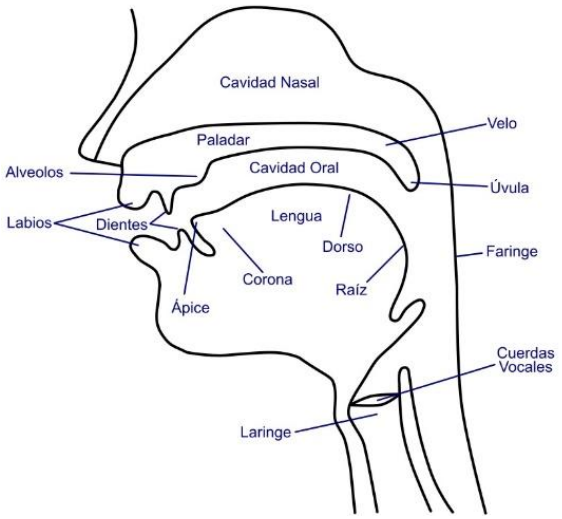


Fig. 6 Imagen creada para ejemplificar los elementos que integran el aparato del habla.

En AFI las vocales son clasificadas en: alargadas, conocidas también como débiles o cerradas porque no es necesario abrir tanto los labios. Y las redondeadas, fuertes o abiertas, porque los labios se abren y redondean para pronunciarlas. En la tabla 3 se muestra la clasificación de las vocales.

Clasificación de Vocales	
Alargadas	Redondas
/i/	/o/
/e/	/u/
/a/	

Tabla 3 Clasificación de Vocales de acuerdo con la formación de la boca

La AFI clasifica de acuerdo con el punto articulación, esto es el lugar donde se forman obstáculos para producir el sonido en:

- Bilabiales, los labios deben estar en contacto (completo o parcial).
- Labiodental se forma cuando los incisivos superiores e inferiores se tocan.
- Interdental se articula con el ápice de la lengua entre los dientes /θ/ de caza es la única consonante interdental.
- Dentales se caracterizan por el contacto del ápice de la lengua con los dientes superiores.
- Alveolares a su vez se clasifican en apicoalveolares y láminoalveolares. Las apicoalveolares se articulan por la aproximación o contacto de la punta de lengua con los alveolos. La láminoalveolar se produce por el contacto de la

lámina de la lengua con los alvéolos. Tomaremos estas en un solo grupo, alveolares.

- Palatales se dividen en alveopalatales y láminopalatales, estas últimas son cuando la lengua toca el paladar. Las consonantes que pronunciamos con la lengua detrás de la cresta alveolar y la lámina de la lengua hacia el paladar se conocen como alveopalatales.
- Velares se articulan con el dorso de la lengua en el Avelo.
- Laríngeas se articulan con una fricción en la laringe.

Clasificación de Consonantes							
Bilabiales	/p/	/b/	/m/				
labiodental	/f/						
interdental	/θ/						
Dentales	/t/	/d/					
Alveolares	/l/	/n/	/r/	/s/			
Palatales	/tʃ/	/dʒ/	/ʒ/	/ɲ/	/ʃ/	/ʝ/	/ʎ/
Velares	/k/	/g/	/x/				
Laríngeas	/h/						

Tabla 4 Clasificación de fonemas de acuerdo con el punto de Articulación según AFI

El presente trabajo propone la identificación de fonemas a partir de los movimientos de los labios, es decir, la lectura labial. Los fonemas a clasificar son los definidos por el método Adryna [8] usado por algunos terapeutas del lenguaje para mejorar la comunicación con personas con necesidades especiales de expresión oral. El método consiste en 21 fonemas, de los cuales 5 son corresponden a las vocales y 16 son consonantes. El método usa una imagen representativa para cada fonema. La Tabla 5 muestra los fonemas y la imagen correspondiente que usan los terapeutas.

Se sabe que no existe una relación clara y definida entre fonemas y visemas en el español. Para identificar los visemas en la lectura labial generalmente se analizan la forma de articulación del fonema. De la clasificación hecha por la AFI podemos observar que existen varios fonemas que comparten características al momento de articularse, estas son movimientos de los labios, posición de la lengua, el velo, entre otras. Uno de los problemas que ha incrementado la complejidad de la lectura labial automática radica en que un visema puede corresponder a varios fonemas, es decir que la representación visual es muy similar o idéntica, por ejemplo, las vocales /i/ y /e/, o las consonantes /b/, /p/ o /m/. Aunque el método Adryna utiliza imágenes (tabla 1) que intentan hacer distinción entre los fonemas la semejanza de la forma de los labios es mínima en algunos casos, influyendo esto en los resultados de la clasificación.

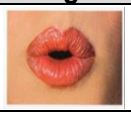
Fonema	Imagen	Fonema	Imagen	Fonema	Imagen
U		Ñ (Ña)		J (Ja)	
A		K (ka)		Y (Ya)	
I		G (Ga)		L (La)	
E		CH (Che)		F (Fa)	
O		R (Ra)		N (No)	
S (Si)		B (Be)		M (Ma)	
P (Pa)		T (Te)		D (Da)	

Tabla 5 Fonemas propuestos en el Método Adryna. Estas imágenes fueron tomadas de la referencia [8] sólo con propósitos educativos y de investigación.

La complejidad descrita en la identificación de visemas, nos motivó a reducir el problema de 21 clases de fonemas definidos en el método Adryna. Estos 21 fonemas fueron agrupados basándose en la clasificación por punto de articulación definida en AFI, ya que esta relaciona la forma o movimientos de los labios al articular el sonido. El problema de identificación de fonemas se redujo a 8 clases, considerando la división por puntos de articulación. La tabla 6 muestra la clasificación de los fonemas considerada para el presente proyecto.

Vocales	*Redondeadas*	U	O		
	Alargadas	A	I	E	
Consonantes	*Bilabial*	M - MA	B - BE	P - PA	
	Dental	T - TE	D - DA		
	Velar	K - KA	G -GA	J - JA	
	Labiodental	F - FA			
	Alveolar	S - SI	L – LA	N - NO	L - RA
	Palatal	Ñ - ÑA	CH - CHE	Y - YA	

Tabla 6 Clasificación de Fonemas de acuerdo con el punto de articulación para el proyecto

4.2. Modelo / Arquitectura

En el capítulo 3 revisamos la arquitectura general de los sistemas de Reconocimiento Automático del Habla (ASR) y los sistemas de Lectura Automática de Labios (ALR). Los elementos principales son captura de video, extracción de características y el algoritmo de clasificación. La arquitectura de la presente propuesta se muestra en la figura 7. Esta incluye la captura de videos y construcción de una base de datos audiovisual, posteriormente se realiza la detección de labios y la extracción de características, finalmente se aplica el algoritmo de clasificación para el reconocimiento de fonema. En las siguientes secciones se describe cada uno los elementos de la propuesta.

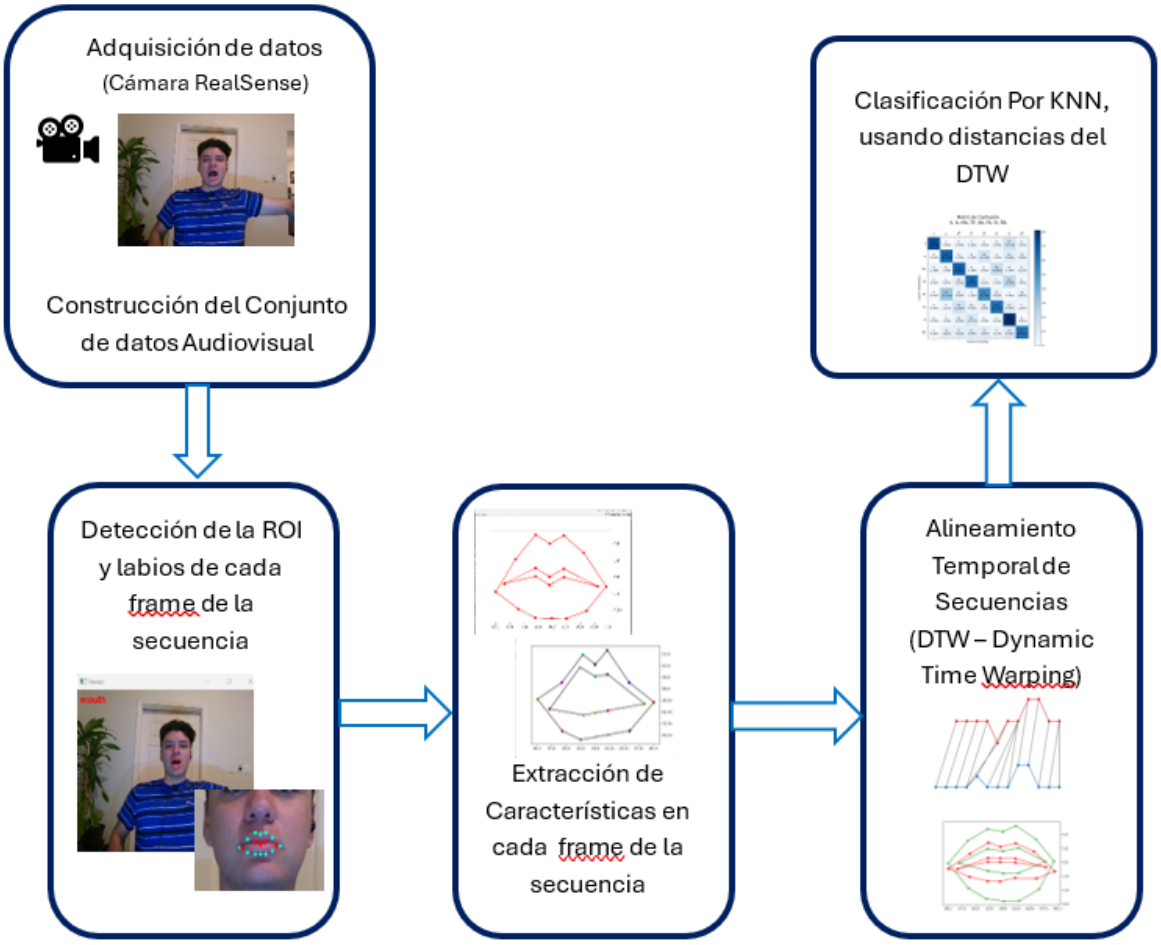


Fig. 7 Arquitectura propuesta para la clasificación de Fonemas

4.3. Materiales y Métodos

4.3.1 Materiales

Para la adquisición de datos se utilizó una cámara intel RealSense D455. La cámara cuenta con un sistema de visión estereoscópica, contiene un procesador de visión y un módulo de profundidad estereoscópica con conexión USB 2.0/USB 3.1. La figura 8 muestra el equipo utilizado.

Fig. 8 Equipo de adquisición de datos. Fotografía tomada del equipo utilizado

Los archivos generados por la cámara tienen la extensión bag, estos archivos contiene los datos de profundidad, la distancia a los objetos ubicados en el campo de visión de cámara, datos de color formados por los fotogramas en formato RGB y la información obtenida por el sensor de infrarrojo.

Aunque la cámara ofrece algunas aplicaciones para visualizar y grabar video, no fueron usadas para el proceso adquisición de datos, ya que en los inicios del proyecto se planteó utilizar la información de profundidad y audio. Para grabar la información de los diferentes canales se implementó un programa en Python para lograr la sincronización de la cámara con un micrófono para adquirir video y audio al mismo tiempo.

Para la implementación de este programa se usó la biblioteca Intel® RealSense™ SDK 2.0, ésta es una biblioteca multiplataforma para cámaras de profundidad de la serie 400. Específicamente se usó pyrealsense2, que es la versión para python de Intel RealSense SDK 2.0. Esta nos permite configurar la cámara, grabar los videos y extraer la información almacenada en el archivo bag.

La biblioteca permite crear un objeto para la configuración de la cámara y la definición de las especificaciones de cada stream de salida. Los frames por segundo (FPS) se definen a nivel de código, se puede modificar; en este proyecto se definió un valor de 30 FPS.

```
config = rs.config()
config.enable_stream(rs.stream.infrared, 1, 640, 480, rs.format.y8, FPS)
config.enable_stream(rs.stream.depth, 640, 480, rs.format.z16, FPS)
config.enable_stream(rs.stream.color, 640, 480, rs.format.bgr8, FPS)
```

Posteriormente se define y configura el objeto de flujo (pipeline), esto crea una secuencia de componentes de procesamiento de datos que permite capturar y manipular datos de los sensores de la cámara RealSense.

```
pipeline = rs.pipeline()
pipeline.start(config)
```

Después de la configuración se inicia la grabación (pipeline.wait_for_frames()) y se capturan los canales de información que se desean almacenar en el archivo.

```
color_frame = frames.get_color_frame()
ir_frame = frames.get_infrared_frame()
depth_frame = frames.get_depth_frame()
```

Para la grabación de audio se usó la biblioteca sounddevice de Python, que permite la reproducción y grabación de audio de alta calidad en tiempo real.

La resolución de las imágenes fue de 640 x 480 píxeles y una tasa de 30 FPS. La duración de cada video fue de 5 segundos.

Para la grabación de los videos las personas se sentaron en una silla dejando una distancia entre la cámara y el rostro de entre 65 a 70 cm. La cámara se montó sobre un tripie, la altura de este se ajustó tratando que el rostro de la persona estuviese centrado. Detrás de la cámara se tenía la computadora donde se estaba ejecutando el programa de grabación (figura 9). El programa emite un sonido “beep” para indicar a la persona que en ese momento se inicia la grabación del video y él inicie la producción del sonido del fonema. Los espacios físicos de grabación fueron en su mayoría salones de clase. Se buscó que se tuviera de fondo una pared. La iluminación fue la luz natural que entraba por las ventanas y los horarios de grabación fueron entre las 12:00 p.m. y las 4:00 pm.

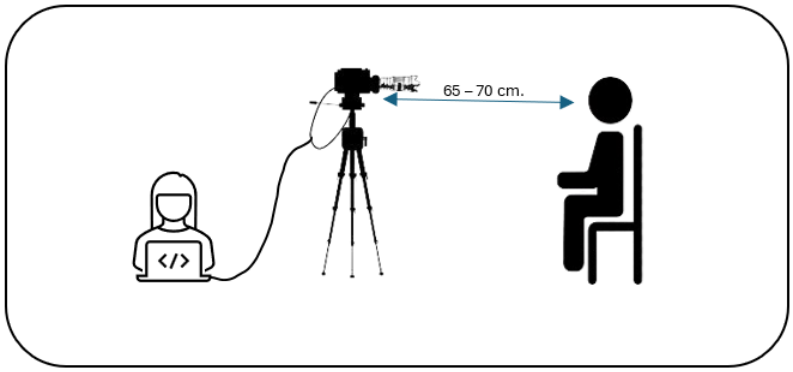


Fig. 9 Imagen creada para ilustrar el esquema de Grabación de vídeos

4.3.2 Participantes

En el proceso de adquisición de datos participaron 12 personas de los cuales 5 fueron mujeres y 7 hombres, la edad de los participantes fue entre 18 y 25 años (Tabla 7). Cada persona grabó los 21 fonemas dos veces cada uno. El total de videos obtenidos fue de 504 (12*21*2=504), con sus correspondientes archivos de audio.

Participantes	Hombres	Mujeres	Edades	Fonemas grabados	Repeticiones por Participante
12	7	5	18 a 25	21	2

Tabla 7 Datos de los participantes en la construcción del corpus

Para realizar la tarea de grabación se siguió la declaración de Helsinki [55], todos los participantes tuvieron consentimiento informado y fueron voluntarios. Desde un principio se le informó sobre el propósito de la investigación y del experimento. Tuvieron toda la libertad de abandonar el experimento en cualquier momento. Ningún participante sufrió daño por la experimentación. Para cada participante se recolectó un formato de consentimiento informado y fue leído y firmado de manera voluntaria.

4.3.3 Método

4.3.3.1 Preprocesamiento de los datos

Una vez concluida la grabación de los videos se procedió a procesarlos, y se extrajeron los frames de color, profundidad y el audio del archivo con extensión bag. Cada archivo de video se idéntico asignando una “P” seguida de un número que identifica a la persona, asi como si se trató de grabación 1 o 2, indicándolo con una V seguida del número 1 u 2. Finalmente, se indica el nombre del fonema al que corresponde el video. Por ejemplo “P1_V1_A.bag” Indica que se trata del video 1 del fonema /a/ correspondiente a la persona 1(Figura 10).

Al revisar los frames correspondientes a cada video, se observó que había muchos espacios sin sonido y por consiguiente movimiento de los labios que no abonan información que apoye el proceso de clasificación, y solo generarían ruido, se decidió eliminar esos frames de video. Esta actividad se hizo de forma manual, observado cada conjunto de frames del video.

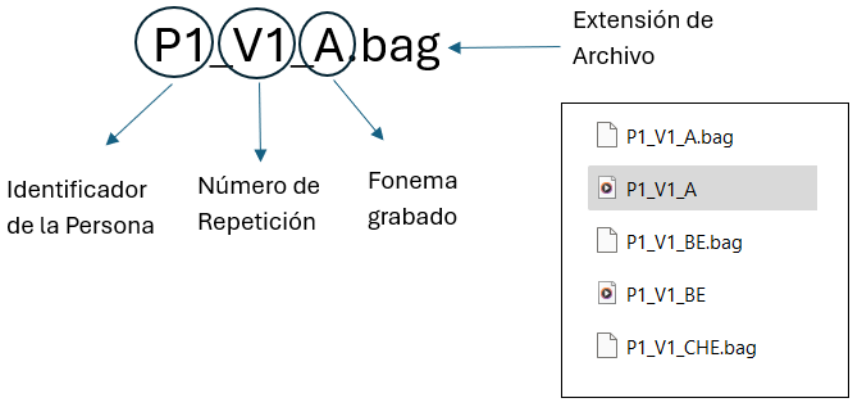


Fig. 10 Nomenclatura de los archivos de Audio y Video

De cada video correspondiente a cada fonema se obtuvieron finalmente entre 15 y 35 frames, el número cambia dependiendo de la persona, su velocidad de pronunciar y el fonema pronunciado.

Para cada persona se crea una carpeta, identificándola con la letra P más un número que lo identifica, por ejemplo, P1 indica que en esa carpeta corresponde a la información de la persona 1. Dentro de cada carpeta se encuentra una carpeta por cada fonema indicando el nombre del fonema y numero de video, por ejemplo, “P12V1_FA” indica que esa carpeta corresponde a la persona 12, grabación 1 (V1) del fonema “FA”, y dentro de ella se encuentran las carpetas donde se almacenan los frames de color y profundidad extraídos de cada video, así como una carpeta donde se encuentran las características obtenidas de cada frame, y las que ya se encuentran normalizadas (figura 11).

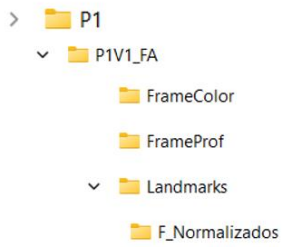


Fig. 11 Estructura del directorio con la información de color y profundidad extraídos de los videos y las características obtenidas de cada frame

4.3.3.2 Detección de la Región de Interés (ROI) y Extracción de Características

Posterior a la obtención de datos, la tarea es extraer la región de interés (ROI) y las características de la secuencia de video. Al realizar la revisión de la literatura se identificó que las características deben ser invariantes de los rasgos faciales de cada persona como barba, color de piel, bigote, entre otras. Además Gimeno Gómez [49] en su documento de investigación clasifica las características en:

- Geométricas: estas características consideraran la forma de la boca de las personas. Consiste en extraer una serie de puntos (landmarks) que identifican el contorno de los labios. Con estas características algunos autores calculan métricas como la altura, anchura y área bucal en cada frame del video. Esta característica se considera muy importante ya que al escuchar nos fijamos generalmente en los movimientos de los labios.
- Basadas en la Apariencia: este tipo características se extraen de la información visual contenida en los píxeles en el área de la boca.
- Basadas en el Movimiento: estas características buscan capturar los movimientos de dos frames consecutivos del video, en función de los cambios de intensidad producidos, basándose en la técnica de flujo óptico (Optical Flow).
- Basadas en Deep Learning: en este tipo se delega a la técnica de aprendizaje, en su mayoría Redes Neuronales Convolucionales realizan la extracción de características.
- Híbridas: combinación de varias de las técnicas.

En el presente trabajo se extraen características geométricas, el contorno interior y exterior que forman los labios.

Inicialmente se pensó en identificar todo el rostro para dentro de esta región, extraer las características que se requirieran para la clasificación, figura 5. Para la detección del rostro y los labios se utilizó una implementación en Python para la detección de objetos de la biblioteca dlib [56]. Dlib es un conjunto de herramientas de C++ que incluye algoritmos y herramientas de aprendizaje automático, procesamiento de imágenes entre otras, y que podemos usar en Python.

La función dlib.get_frontal_face_detector() es una función de la biblioteca dlib que se utiliza para crear un detector de rostros frontales. Este detector puede identificar rostros en imágenes utilizando un conjunto de características conocidas como histogramas de gradiente orientado (HOG) y un clasificador basado en máquina de vectores de soporte (SVM). Del sitio web de la biblioteca dlib se puede descargar el

archivo shape_predictor_68_face_landmarks.dat que contiene los datos del modelo pre entrenado proporcionado por la biblioteca.

El primer paso es identificar el rostro, para ello se pasa como parámetros al objeto creado para la predicción de rostros, la imagen en escala de grises y el número de rostro que desean identificar en la imagen: *detector(gray, 1).*

Para evaluar si se estaban identificando correctamente el rostro se graficó un rectángulo a partir de las coordenadas (x,y), de la esquina superior derecha y esquina inferior izquierda que retorna el método, y se recortó esta región para mostrarla en otra ventana, obteniendo resultados como el de la figura 12.

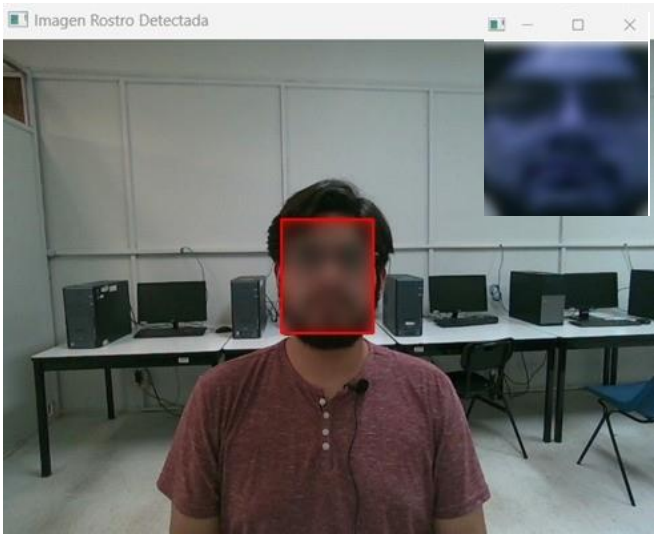


Fig. 12 Identificación del Rostro

El detector de puntos de referencia facial de dlib, permite identificar y etiquetar regiones faciales como: boca, cejas, ojos, nariz y mandíbula. Este detector de puntos de referencia facial es una implementación del trabajo de V. Kazemi y J. Sullivan [50]. El detector estima la ubicación de 68 coordenadas (x, y) que se asignan a las estructuras faciales de la cara, como las mostradas en la figura 13.

De la estructura obtenida por el detector se extrajeron los puntos 48 al 68 correspondientes al contorno de los labios. Estos puntos de cada frame fueron almacenados en un archivo. Se almacenaron en total 20 puntos correspondientes al contorno exterior e interior de los labios.

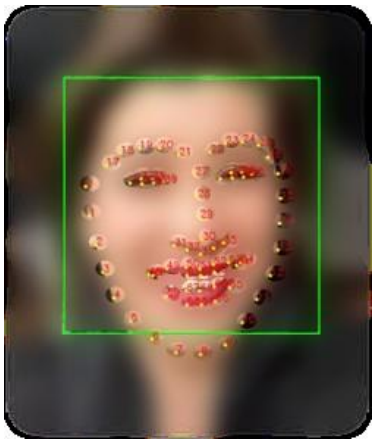


Fig. 13 Coordenadas de las estructuras faciales detectadas con dlib

Las características usadas en el proceso de clasificación además de ser invariantes a los rasgos individuales de cada persona se recomiendan sean invariantes a escala, rotación y traslación con el objetivo de minimizar los efectos que producen los movimientos de las personas al producir un fonema. Al momento de la grabación de los videos se instruyó a los participantes que dirigieran su mirada a la cámara, sin embargo, es difícil controlar que no inclinaran el rostro, que no acercaran su rostro un poco más adelante o lo giraran un poco. Una vez extraídos los datos de la región de interés se aplicaron normalizaciones: traslación, escala y rotación.

Para obtener la traslación de los n puntos que forman el contorno de la boca respecto al origen, se considera cada coordenada (x_i, y_i) correspondiente a cada punto que forman los labios, y el centroide $(x_c.y_c)$, a partir de estos datos se obtiene el nuevo valor de cada punto (xt_i, yt_i):

$$xt_i = x_i - x_c$$

$$yt_i = y_i - y_c$$

Por ejemplo, para las coordenadas de 4 puntos de los labios se obtuvo el centroide obteniendo (figura 14, gráfica1):

Datos Originales= [[256, 154] [268, 145] [282, 153] [269, 159]]

Centroide =[268,152]

Se aplicó la traslación al origen obteniendo el resultado mostrado en la figura 14, gráfica 2:

[[-12, 2], [0, -7], [14, 1], [1, 7]]

Centroide Normalizado Traslación = [0, 0]

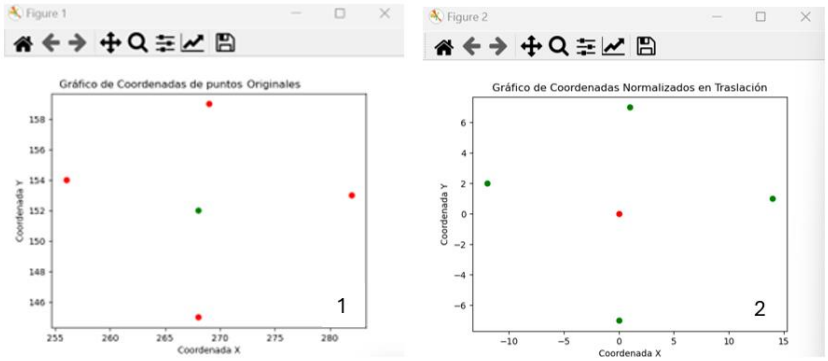


Fig. 14 Ejemplo de resultado de trasladar al origen un vector de puntos(Normalización en traslación).

Posteriormente a los puntos obtenidos como resultado de la traslación, fueron la normalización a escala para dejarlos en el rango de -1 y 1. Para esto primero obtuvimos la norma máxima:

$$NormaMax = \max\left[|(xt_i, yt_i)|\right]$$

Donde *max* es una función que retorna el valor máximo y

$$|(xt_i, yt_i)| = sqrt((xt_i{}^2, yt_i{}^2)$$

Posteriormente obtenemos el nuevo valor para cada punto

$$xe_i = xt_i / NormaMax$$

$$ye_i = yt_i / NormaMax$$

Para los datos resultado de traslación se aplica la normalización a escala y el resultado es mostrado en la figura 15. Los valores obtenidos son:

Normalización Escala =[[-0.854, 0.142], [0.0, -0.498], [0.997, 0.071], [0.071, 0.498]]

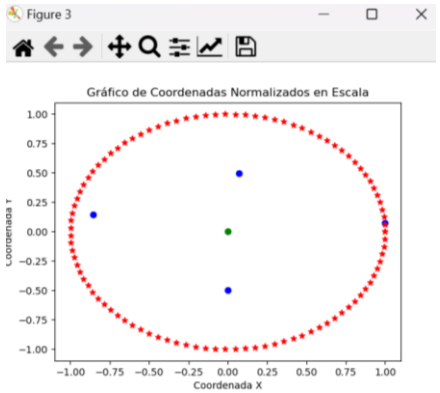


Fig. 15 Ejemplo de aplicar la normalización de escala a los puntos obtenidos en la normalización de traslación. Se graficó una circunferencia con radio 1, mostrados en rojo en la imagen, para percibir visualmente más fácil el resultado de la normalización a escala.

Finalmente se aplicó la normalización en rotación. Para los N puntos ya normalizados en traslación y escala (xe_i,ye_i). Se calcula el ángulo theta de la recta de mínima inercia que pasa por el centroide:

- Se calculan las variables auxiliares *Ixx, Iyy y Ixy*. Estos son momentos estadísticos de orden dos del conjunto de puntos

$$\text{Ixx} = \sum_{i=1}^{N} xe_i^2$$

$$\text{Ixy} = \sum_{i=1}^{N} xe_i \; ye_i$$

$$\text{Iyy} = \sum_{i=1}^{N} ye_i^2$$

- Con los valores obtenidos *Ixx, Iyy y Ixy*, se obtiene el ángulo theta con la siguiente formula:

$$\text{theta} = \frac{\arctan\left\{\frac{2\,Ixy}{Ixx - Iyy}\right\}}{2}$$

- Se obtiene el nuevo valor de (x_i,y_i) de cada punto producto de la rotación:

$$x_i = xe_i * \cos(theta) + ye_i * \sin(theta)$$
$$y_i = -xe_i * \sin(theta) + ye_i * \cos(theta)$$

La figura 17 muestra 4 imágenes, en la primera se muestran los puntos en las coordenadas originales extraídas de la imagen. La imagen 2 muestra el resultado de normalizarlos en traslación, respecto al origen (0,0). La imagen 4 muestra el resultado de aplicar normalización a escala en el rango de [1..-1]. Y finalmente la imagen 4 muestra el resultado de la aplicación de la normalización en rotación. Los datos obtenidos de todos los frames fueron almacenados en un archivo numpy, como matrices, su estructura es:

$$[\,[(x_1Frame_1; y_1Frame_1), (x_2Frame_1; y_2Frame_1), \ldots (x_{20}Frame_1; y_{20}Frame_1)],$$
$$[(x_1Frame_2; y_1Frame_2), (x_2Frame_2; y_2Frame_2), \ldots (x_{20}Frame_2; y_{20}Frame_2)],$$
$$\ldots$$
$$[(x_1Frame_n; y_1Frame_n), (x_2Frame_n; y_2Frame_n), \ldots (x_{20}Frame_n; y_{20}Frame_n)]\,]$$

Otras características que se obtuvieron fueron los ángulos que se forman entre los puntos de las comisuras de los labios y los puntos ubicados en el centro del labio inferior y superior, como se muestra en la figura 18. Estos puntos se muestran en color amarillo y los tres ángulos que se forman entre estos cuatro puntos de los labios. Con las características normalizadas se elimina gran parte del ruido que pudieran tener los datos para la clasificación.

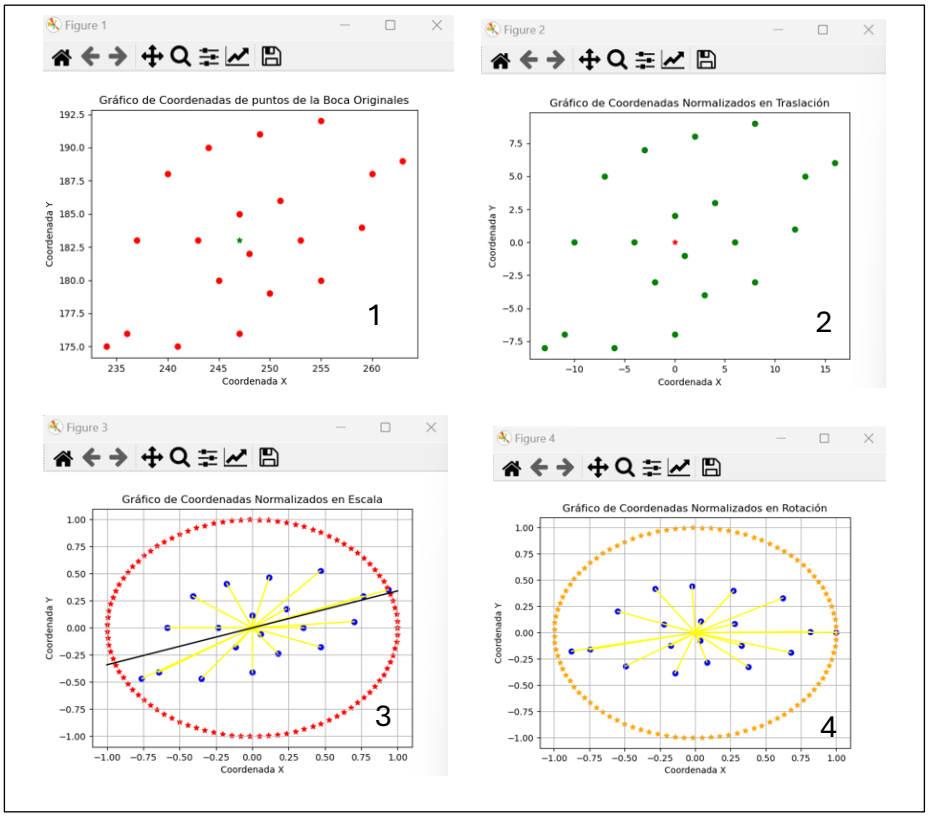


Fig. 16 Trasformaciones geométricas aplicadas a los datos de la región de interés.

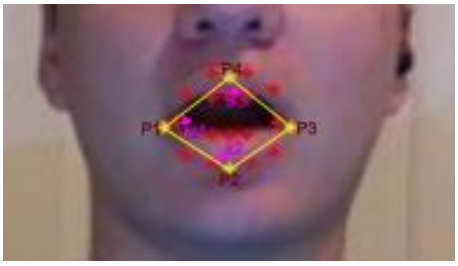


Fig. 17 Identificación de ángulos que forman 4 puntos de los labios

4.3.3.3 Alineación Temporal de Señales

La velocidad de cada persona para producir un fonema es diferente, incluso la misma persona puede modificar su velocidad y movimientos de los labios en repeticiones del mismo fonema. Para obtener una alineación de las series de tiempo formadas por las secuencias de video y obtener la distancia de similitud entre ellas, se usó el algoritmo DTW (Dynamic Time Warping). De esta forma con los datos obtenidos podemos comparar la similitud de secuencias.

Para la alineación temporal de señales se utilizó la librería de Python Tslearn [51], ésta cuenta con un gran número de algoritmos de aprendizaje automático para series temporales. Permite realizar tareas como clasificación, regresión y agrupamiento entre otras. Además, Tslearn proporciona una integración fácil con otras bibliotecas de aprendizaje automático de Python, como scikit-learn y TensorFlow .

En la extracción de características se obtuvieron las coordenadas (x,y) que forman los labios de las secuencias de video, estas se almacenaron en arrays numpy de dos dimensiones. Para aplicar el algoritmo DTW debían ser transformadas en series de tiempo. Esto se logró aplicando la función to_time_series de tslearn.utils.

Existen varias librerías que implementan el algoritmo DTW, para este proyecto se utilizó FastDTW implementación que proporciona alineaciones óptimas o casi óptimas con una complejidad de tiempo y memoria O(N) [52]. El método fastdtw recibe dos series de tiempo y retorna la medida de distancia y la alineación de cada elemento de la serie.

```
dtw_score, optimal_path  = fastdtw(formatted_time_series1, formatted_time_series2)
```

Anteriormente se mencionó que se agruparon los fonemas de acuerdo con la clasificación del punto de articulación, como se mostró en la tabla 4. Se obtuvieron 8 clases (Vocales Redondeadas, Vocales Alargadas, Bilabial, Dental, Velar, Labiodental, Alveolar, y Palatal). De estas se realizaron todas combinaciones tomando un elemento de cada clase resultando 1296 grupos, figura 19.

Grupos = vocales_redondas × vocales_alargadas × con_Bilabial × con_Dental × con_Velar × cons_Labiodental × cons_Aveolar × cons_Palantal

Número de grupos = 2 x 3 x 3 x 2 x 3 x 1 x 4 x 3 = 1296

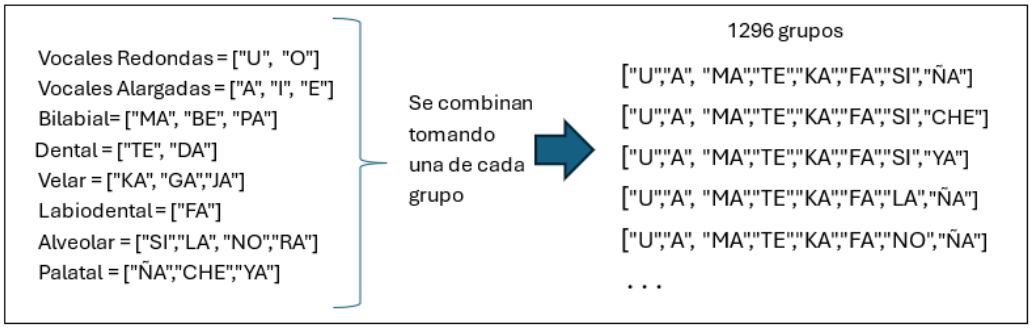


Fig. 18 Combinaciones de Fonemas para aplicar DTW

A cada grupo se aplica DTW, obtenido una matriz de distancias de similitud de 192 filas por 192 columnas, como la mostrada en la figura 17. El tamaño de la matriz de distancias se obtiene:

número de elementos por grupo * número de videos por fonema * número de personas
Tamaño de Matriz de Similitud = 8 * 2 * 12

En la matriz mostrada en la figura 20 se puede observar que en la celda de color amarillo se encuentra el valor 24.64, esta es la distancia de similitud que se obtiene de aplicar DTW al video 1 de la persona 1 del fonema /u/, con el video 2 de la misma persona y fonema. Cuando se compara el mismo video el resultado es 0, como se puede observar en la celda marcada en verde. En la celda azul se está mostrando el resultado donde se está comparando el video 2 de la persona 1 del fonema /a/ con el video 1 de la persona 1 del fonema /u/.

	P1V1_FU	P1V2_FU	P1V1_FA	P1V2_FA	P1V1_FMA	P1V2_FMA	P1V1_FTE	P1V2_FTE	P1V1_FKA	P1V2_FKA	P1V1_FFA	P1V2_FFA	P1V1_FSI	P1V2_FSI	P1V1_FÑA	P1V2_FÑA	P2V1_FU	P2V2_FU	P2V1_FA	P2V2_FA	. . .	P12V1_FFA
P1V1_FU	0.00	24.67	72.00	66.34	30.32	30.37	28.28	30.53	55.03	40.76	19.80	18.59	18.87	16.65	0.00	19.83	48.85	48.85	50.62	62.37		53.83
P1V2_FU	24.67	0.00	47.76	40.39	16.87	20.92	16.18	15.93	32.34	26.60	20.96	17.10	28.06	22.90	19.83	0.00	52.36	52.36	54.85	65.27		57.28
P1V1_FA	72.00	47.76	0.00	26.08	31.88	29.49	44.87	42.79	29.15	28.11	52.18	48.22	59.37	57.65	48.85	52.36	0.00	0.00	1.29	38.58		53.28
P1V2_FA	66.34	40.39	26.08	0.00	24.41	27.22	37.23	32.27	27.36	29.92	52.20	49.75	59.32	58.73	50.62	54.85	1.29	1.29	0.00	42.68		43.88
P1V1_FMA	30.32	16.87	31.88	24.41	0.00	11.72	11.50	12.53	24.17	20.29	68.91	61.16	73.21	68.46	62.37	65.27	38.58	38.58	42.68	0.00		51.29
P1V2_FMA	30.37	20.92	29.49	27.22	11.72	0.00	15.13	17.32	29.08	21.51	89.73	83.39	90.72	88.33	83.43	88.42	63.63	63.63	64.60	32.54		66.16
P1V1_FTE	28.28	16.18	44.87	37.23	11.50	15.13	0.00	13.27	31.85	24.81	32.30	27.82	32.85	29.42	23.20	29.93	28.70	28.70	29.79	31.71		45.27
P1V2_FTE	30.53	15.93	42.79	32.27	12.53	17.32	13.27	0.00	27.80	20.36	33.77	28.19	34.16	31.62	24.76	30.99	29.48	29.48	30.63	28.17		48.01
P1V1_FKA	55.03	32.34	29.15	27.36	24.17	29.08	31.85	27.80	0.00	25.43	48.12	40.32	52.27	46.80	39.52	44.03	31.17	31.17	33.23	34.61		28.07
P1V2_FKA	40.76	26.60	28.11	29.92	20.29	21.51	24.81	20.36	25.43	0.00	42.31	34.66	47.06	41.60	32.97	38.02	28.07	28.07	30.07	31.09		30.64
P1V1_FFA	28.27	17.96	30.77	24.66	8.39	15.56	13.72	14.59	25.32	21.01	64.10	55.07	67.43	61.79	48.04	58.07	30.64	30.64	33.02	23.90		28.45
P1V2_FFA	29.55	15.84	26.17	24.65	8.77	12.91	13.80	13.25	23.20	16.89	53.45	45.54	57.55	52.25	40.14	49.55	28.45	28.45	30.26	20.73		34.11
P1V1_FSI	34.32	24.32	43.39	33.63	14.02	20.71	17.25	14.25	30.17	25.21	49.24	43.86	53.27	48.08	37.03	44.63	34.11	34.11	36.45	45.22		34.16
P1V2_FSI	38.46	24.44	38.11	28.53	14.62	16.10	17.82	15.29	28.63	22.96	54.06	48.87	58.43	53.79	43.86	49.91	34.16	34.16	37.19	41.87		31.48
P1V1_FÑA	45.19	27.37	29.84	22.69	16.73	22.42	22.91	17.31	19.10	19.32	53.96	44.42	59.59	50.29	35.45	48.52	31.48	31.48	33.36	38.97		35.05
P1V2_FÑA	42.89	23.98	31.26	25.58	18.06	21.56	20.67	22.00	23.22	22.17	54.83	45.89	61.31	51.54	36.42	49.48	35.05	35.05	36.66	44.88		36.84
P2V1_FU	89.19	64.26	39.72	45.65	50.83	57.63	61.26	57.96	55.39	60.88	67.74	58.13	80.27	68.02	44.73	61.80	36.84	36.84	38.19	40.27		28.98
P2V2_FU	90.83	65.15	41.41	45.95	52.17	58.92	62.41	59.59	56.47	62.25	42.34	35.68	46.67	40.79	30.10	39.96	28.98	28.98	30.26	28.21		51.51
P2V1_FA	103.90	79.84	49.84	61.47	65.15	72.20	76.04	72.24	63.35	67.57	16.50	14.20	17.10	18.81	19.93	21.66	51.51	51.51	53.88	64.11		42.80
P2V2_FA	115.60	96.44	75.02	84.77	85.35	94.15	93.35	91.30	86.45	93.82	34.21	32.83	35.92	35.55	27.24	29.00	42.80	42.80	45.18	62.54		38.19
. . .																						
P12V1_FFA	86.12	61.48	33.26	36.22	44.89	53.83	57.28	53.28	43.88	51.29	66.16	45.27	48.01	55.13	49.54	64.60	52.80	39.26	48.24	54.79		0.00

Fig. 19 Ejemplo de matriz de distancias de similitud obtenida de aplicar DTW.

Se graficó el warping path, para observar cómo llevó a cabo el algoritmo DTW la alineación de las secuencias, en la figura 21 muestra la alineación de fonema “da” de la persona 2 videos 1 y 2. Podemos observar que, aunque corresponde a la misma persona y mismo fonema, existen diferencias. Para este ejemplo la distancia obtenida fue de 19.06.

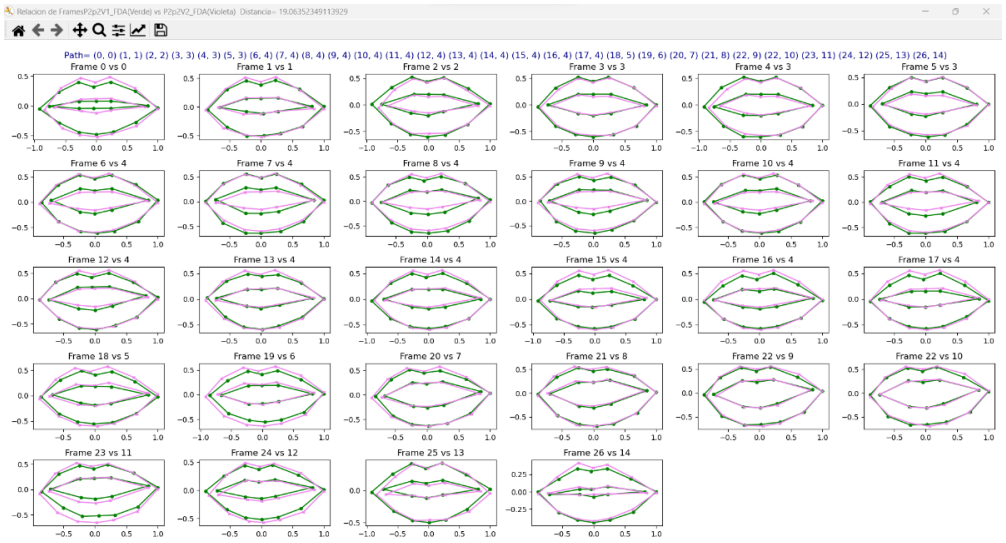


Fig. 20 Alineación de fonema "DA" de la persona 2, videos 1 y 2.El fonema se construye con 26 frames en el video 1 y 14 en el video 2. En verde se muestran las imágenes del video 1 y en purpura las del video 2.

Cuando se comparan fonemas diferentes la distancia se incrementa, y al graficar el warping path se puede observar esto. En la figura 22 muestra warping path del fonema /a/ de la persona 1, video 1 y el fonema /u/ de la misma persona. Una secuencia tiene una longitud de 21 frames y la otra de 24.La distancia es de 71.15.

Para determinar los datos que integrarían los conjuntos de prueba y de entrenamiento se utilizó el método de validación cruzada. Bishop [44] recomienda el método para evitar estimaciones con ruido del modelo de predicción cuando el número de datos no es muy grande, y por consiguiente el conjunto de datos de validación es pequeño. El método de validación cruzada consiste en dividir aleatoriamente el conjunto datos disponibles en S grupos, a continuación, se utiliza S - 1 de los grupos para entrenar un conjunto de modelos que se evalúan en el grupo restante. Este procedimiento se repite para todas las S opciones posibles para el grupo que no se ha utilizado.

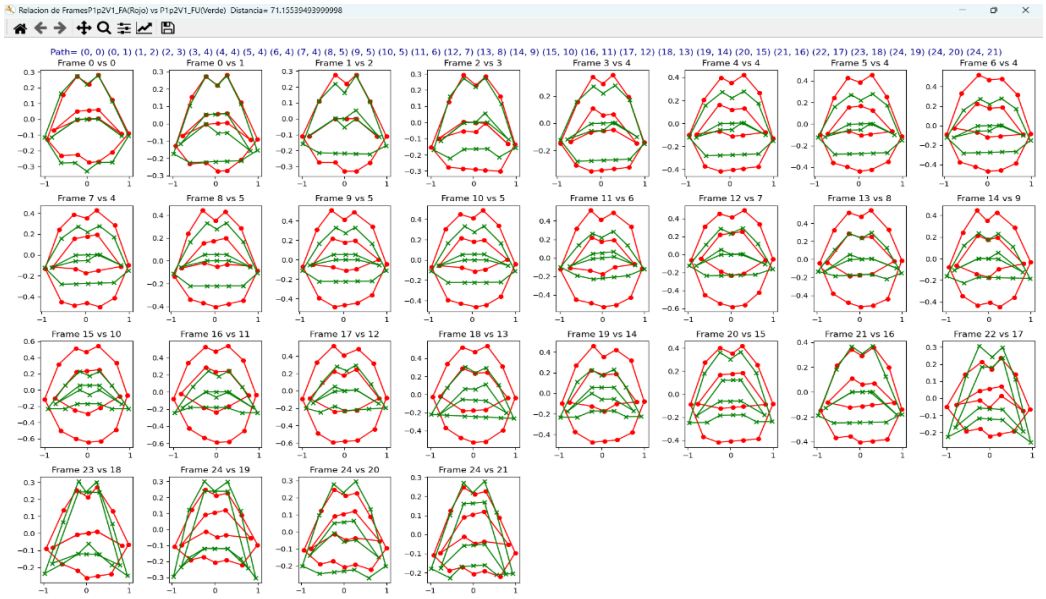


Fig. 21 Warping path fonema /u/ y /a/ persona 1. EL fonema /a/ tiene 24 frames y el /u/ 21. En color rojo se muestra las imágenes correspondientes al fonema /a/ y en verde el fonema /u/.

4.3.3.4 Algoritmo de Clasificación

Para el proceso de clasificación se usó el algoritmo K-NN ya que este algoritmo de aprendizaje permite obtener buenos resultados cuando el conjunto de datos no es muy grande, además nos permite utilizar la medida de distancia de similitud obtenida con DTW para determinar los k vecinos.

La entrada al algoritmo es una lista con las medidas de similitud que hay entre el fonema X al resto de los fonemas que pertenecen al grupo. Por ejemplo, el grupo [U,A,MA,TE,KA,FA,SI,ÑA] está integrado por un elemento de cada grupo de los definidos en la tabla 4. La longitud de la lista de distancias recibidas es de 192.

Algoritmo K-NN (Lista de distancias ,valor de K, foldPrueba)

De la lista de distancias se asigna valor infinito a los datos de entrenamiento

Ordena la lista de distancias DTW de menor a mayor

Si k = 1

Retorna la clase asociada a la primera distancia de la lista

De lo contrario

Calcula las frecuencias de cada clase de los K primero elementos de la lista.

Retorna la clase que tiene mayor frecuencia.

4.3.3.5 Métricas de Clasificación

El problema de clasificación planteado en el presente trabajo es un problema multiclase, para la evaluación del algoritmo de aprendizaje se construyó y utilizó la matriz de confusión, y las métricas accuracy, recall, precisión y F1-score.

La matriz de confusión que permite registrar el número de ocurrencias entre la clasificación real y la clasificación predicha. Las clases se enumeran en el mismo orden en las filas que en las columnas, por lo que los elementos correctamente clasificados se encuentran en la diagonal principal. En la figura 19 se muestra un ejemplo de una matriz de confusión.

Al observar la figura 23 podemos interpretar que la vocal /O/ fue predicha correctamente 103 veces, 19 veces fue confundida con un vocal /E/.Tambien observamos que la vocal /E/ fue 1 vez confundida con la vocal /O/, y 78 veces fue predicha correctamente.

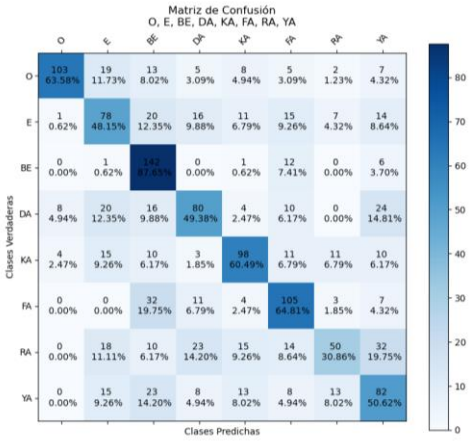


Fig. 22 Ejemplo gráfico de una Matriz de Confusión multiclase

De la matriz de confusión multiclase se obtienen los valores TP (True Positive), TN (True Negative), FP (False Positive) y FN (False Negative) para cada clase :

- TP (True Positive) : son los valores que el algoritmo clasifica como positivos y que realmente son positivos. Para el ejemplo de la matriz de confusión de la figura 18 tenemos:

  TP clase /O/ = 103 (Valor correspondiente a la clase de la diagonal principal de la matriz).

- TN (True Negative): son valores que el algoritmo clasifica como negativos y que verdaderamente son negativos. Para la matriz de ejemplo de la figura 18:

TN clase “/O/”= 1185 (suma de valores que de las filas y columnas excepto las correspondientes a la clase, en este caso todos los valores excepto de la fila 1 y columna 1)

- FP (False Positive): son valores que el algoritmo clasifica como positivo cuando realmente son negativos.

  FP clase “/O/”= 1 + 0 + 8+ 4 + 0 + 0 + 0 = 13 (suma de los valores de la columna correspondiente a la clase excepto el TP)

- FN (False Negative): son valores que el algoritmo clasifica como negativo cuando realmente son positivos.

  FN clase “/O/”= 19+13+5+8+5+2+7= 59 (Suma de los valores de la fila correspondiente a la clase excepto el TP)

Las métricas utilizadas para la medición del desempeño del algoritmo son:

- *Accuracy (precisión)* : es la probabilidad de que la predicción del modelo sea correcta, nos indica el porcentaje total de valores correctamente clasificados, tanto positivos como negativos[53].

$$Acurracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision*: es la fracción de instancias clasificadas correctamente de una clase específica respecto a todas las instancias que el modelo predijo que pertenecían a esa clase. La precisión mide la capacidad del modelo para identificar correctamente instancias de una clase particular [53].

$$Precision_{Clase\ X} = \frac{TP_{Clase\ X}}{TP_{Clase\ X} + FP_{Clase\ X}}$$

- *Recall (Sensibilidad*)**:** es la fracción de instancias en una clase que el modelo clasificó correctamente de todas las instancias en esa clase. La sensibilidad mide la capacidad del modelo para identificar todas las instancias de una clase particular, nos da información sobre el rendimiento con respecto a falsos negativos [53].

$$Recall_{Clase\ X} = \frac{TP_{Clase\ X}}{TP_{Clase\ X} + FN_{Clase\ X}}$$

- *F1-Score (Puntuación F1):* esta métrica considera la precisión y sensibilidad de un modelo de clasificación. Se obtiene a partir de las medidas *Precision* y *Recall.* Existen dos formas de obtener este valor conocidas como Macro Promedio F1 y Micro Promedio F1. En [53] se muestra que Micro Promedio F1 es equivalente a calcular *Accuracy* a partir de *Recall* y *Precision*. Aunado a lo anterior, se sabe que en el enfoque Macro, cada clase tiene el mismo peso, lo que implica que no hay distinción entre el tamaño de las clases, por lo que se decidió utilizar la fórmula de Macro Promedio F1.

  Para obtener Macro *F1-Score* se debe calcular antes Macro Promedio *Precision* y Macro Promedio *Recall*. La macro precisión y recuperación media se calculan como la media aritmética de las métricas para clases individuales:

$$Macro\ Promedio\ Precision = \frac{\sum_{C=1}^{C} Precision_C}{C}$$

$$Macro\ Promedio\ Recall = \frac{\sum_{C=1}^{C} Recall_C}{C}$$

  Donde *C* es el número de clases. A partir de estas medidas se calcula el Macro F1-Score con la siguiente formula:

$$Macro\ F1 = 2 * \left(\frac{Macro\ Promedio\ Precision * Macro\ Promedio\ Recall}{Macro\ Promedio\ Precision^{-1} + Macro\ Promedio\ Recall^{-1}}\right)$$

Otro elemento que se utilizó para el análisis de los resultados fue la curva de características operativas del receptor ROC (Receiver operating characteristics) y el área bajo la curva llamada AUC (Area under the ROC curve) . La curva ROC es una representación gráfica bidimensional del rendimiento de un clasificador. AUC es una porción del área del cuadrado unitario, su valor siempre estará entre 0 y 1, es equivalente a la probabilidad de que el clasificador clasifique una instancia positiva elegida al azar de forma correcta, suele ser utilizado como resumen cuantificable de la calidad del modelo [54].

La grafica ROC obtenida para un grupo de fonemas se muestra en la figura 23.

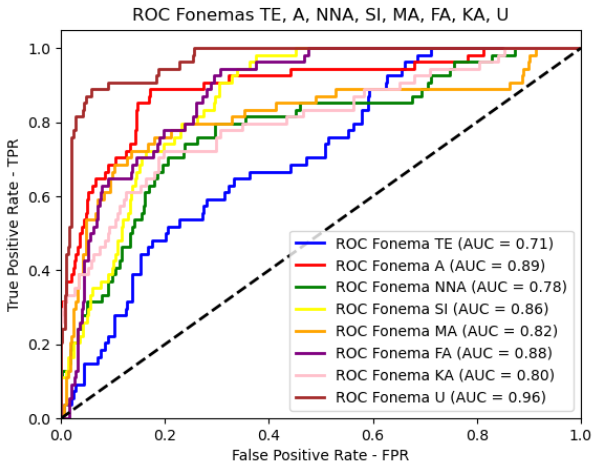


Fig. 23 Ejemplo de Grafica Roc y valores AUC para un grupo de fonemas

5. Experimentos, Resultados y Discusión

5.1 Descripción de los experimentos realizados

Se agruparon los experimentos realizados para probar el modelo en tres grupos:

- Selección de un fonema de cada grupo. Como se mencionó en la sección anterior los fonemas se agruparon de acuerdo con la clasificación por punto de articulación obteniendo ocho grupos. Se generaron todas las combinaciones de los elementos pertenecientes a los grupos de fonemas, obteniendo un total de 1296 combinaciones. Analizaremos los resultados de 2 grupos de fonemas en la siguiente sección.
- Uso de ángulos. Se probó el modelo con los ángulos obtenidos de los puntos de las comisuras de los labios, asi como el punto medio del labio inferior y superior, descritos en la sección 4.5 (figura 18) .
- Vocales. Otra prueba del modelo se realizó con las vocales exclusivamente, reduciendo el problema de clasificación a 5 clases.

5.2 Resultados

Experimento 1: un fonema por grupo (/U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/)

Con el primer grupo de fonemas formado por /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/ se obtuvo la matriz de confusión mostrada en la figura 25.

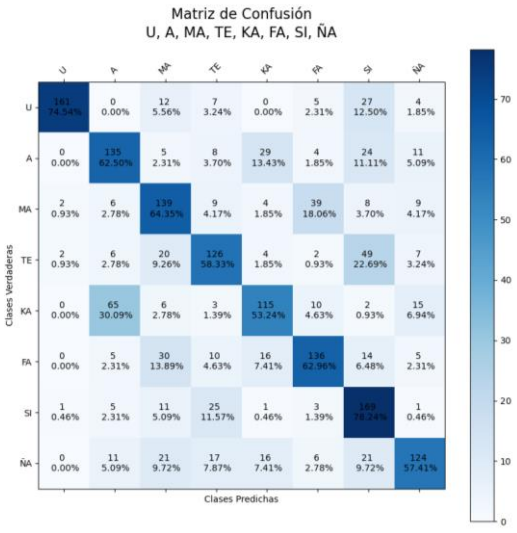


Fig. 24 Matriz de Confusión resultado de aplicar KNN a los fonemas /U/,/A/,/MA/, /TE/,/KA/,/FA/, /SI/, /ÑA/

Para ese mismo grupo de fonemas se obtuvieron los siguientes valores de las métricas:

Métricas Macro Promedio Fonemas /U/,/A/,/MA/, /TE/,/KA/,/FA/, /SI/, /ÑA/

	Precision	Recall	F1-score
Macro Promedio	0.657672901	0.63946759	0.64215333

Accuracy	0.639467593

Las métricas obtenidas para cada fonema se muestran en la tabla 8.

Clase	Precision	Recall	F1-score
A	0.579399142	0.625	0.6013363
FA	0.663414634	0.62962963	0.64608076
KA	0.621621622	0.53240741	0.57356608
MA	0.569672131	0.64351852	0.60434783
ÑA	0.704545455	0.57407407	0.63265306
SI	0.538216561	0.78240741	0.63773585
TE	0.614634146	0.58333333	0.59857482
U	0.969879518	0.74537037	0.84293194

Tabla 8 Métricas obtenidas por clase para el grupo 1

La grafica ROC correspondiente se muestra en la figura 23. La grafica A, fue construida utilizando un vector de probabilidades normalizadas, se inició calculando la probabilidad de cada clase:

$$P_{Ci} = \frac{1}{d_i}$$

Donde P_{Ci} *es la probabilidad* de la clase i y d_i es la distancia dtw de la clase i. Posteriormente se realiza la sumatoria de las probabilidades, para finalmente obtener el vector de probabilidades normalizadas:

$$Probabilidades\ Normalizadas = \left[\frac{P_{C1}}{\sum_{i=1}^{n} P_{Ci}}, \frac{P_{C2}}{\sum_{i=1}^{n} P_{Ci}}, \ldots \frac{P_{Cn}}{\sum_{i=1}^{n} P_{Ci}}\right]$$

Y la gráfica B se obtiene mediante el método One-Hot Encoding, este crea un vector del tamaño de números de clases y se coloca el valor 1 en la clase predicha y 0 en el resto.

El promedio de AUC para este grupo de fonemas es de 0.925 y se calcula las probabilidades normalizadas (gráfica A figura 26 y para el método one-hot encoding fue de 0.79375 (gráfica B figura 26).

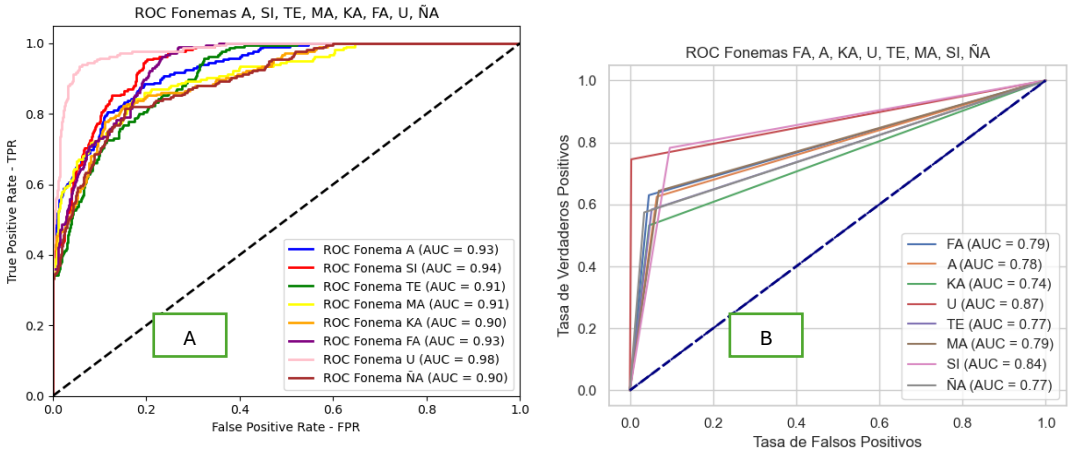


Fig. 25 Graficas ROC correspondientes a los fonemas /U/,/A/,/MA/, /TE/,/KA/,/FA/, /SI/, /ÑA/. Grafica A probabilidades Normalizadas y grafica B usando One-Hot Encoding

Experimento 2: un fonema por grupo (/E/, /MA/, /NO/, /JA/, /FA/, /CHE/, /U/, /DA/)

Para este experimento se seleccionó otro grupo de fonemas /E/, /MA/, /NO/, /JA/, /FA/, /CHE/, /U/, /DA/ la matriz de confusión se muestra en la figura 27.

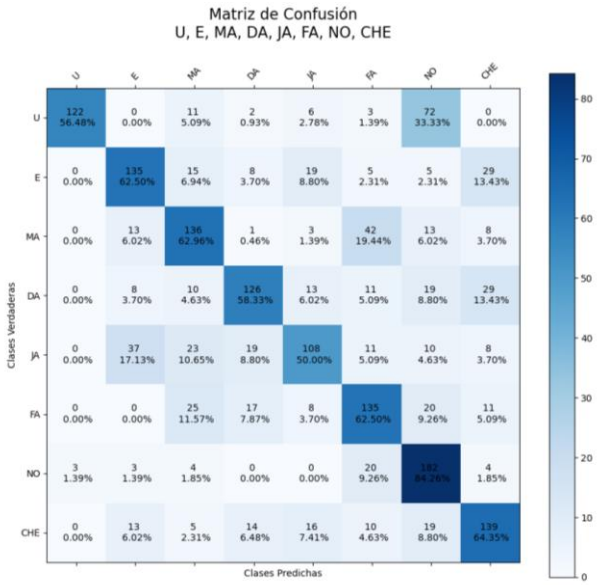


Fig. 26 Matriz de Confusión Grupo de fonemas /E/, /MA/, /NO/, /JA/, /FA/, /CHE/, /U/, /DA/

Los resultados obtenidos de las métricas en macro promedio y por cada fonema se muestran en la tabla 9.

Métricas Macro Promedio Fonemas /E/, /MA/, /NO/, /JA/, /FA/, /CHE/, /U/, /DA/

	Precision	Recall	F1-score
Macro Promedio	0.65355715	0.62673611	0.6274352

Accuracy	0.62673611

Clase	Precision	Recall	F1-score
CHE	0.60964912	0.64351852	0.62612613
DA	0.67379679	0.58333333	0.62531017
E	0.64593301	0.625	0.63529412
FA	0.56962025	0.625	0.59602649
JA	0.62427746	0.5	0.55526992
MA	0.59388646	0.62962963	0.61123596
NO	0.53529412	0.84259259	0.65467626
U	0.976	0.56481481	0.71554252

Tabla 9 Métricas Macro promedio y por clase para el grupo de Fonemas

La grafica ROC obtenida para estos fonemas se muestra en la figura 28 y se obtuvo un promedio de AUC de 0.932 para la gráfica A y 0.78 para la gráfica B.

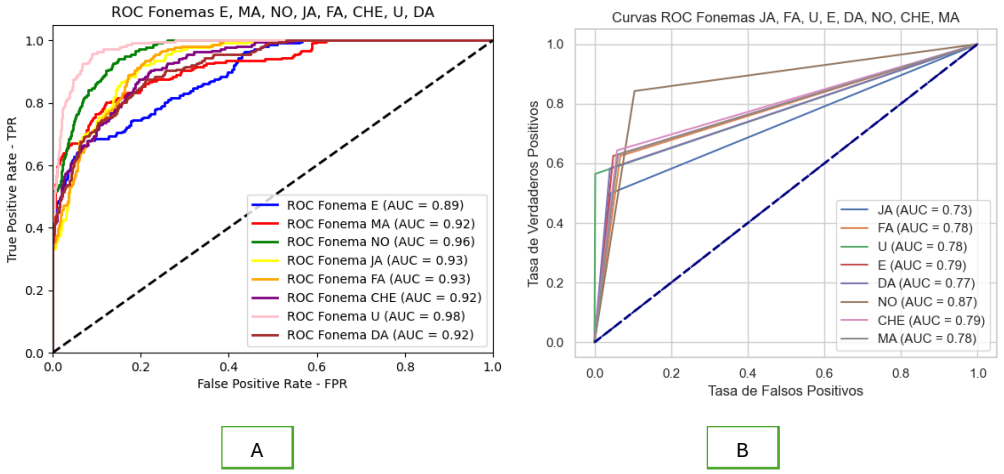


Fig. 27 Gráficas Roc fonemas /JÁ/, /FA7, /U/, /E/, /DA/, /NO/, /CHE/, /MA/. La gráfica A construida a partir de probabilidades normalizadas y B usando One-Hot Encoding.

Se obtuvieron los promedios de las métricas de la ejecución del algoritmo de clasificación aplicado a los 1296 grupos de datos, el accuracy promedio del modelo fue de 0.65272777 y las métricas por grupo se muestran en la siguiente tabla 10.

Métricas Promedio

Clase	Precision	Recall	F1-Score
Vocal Redonda	0.9647	0.7214	0.8228
Vocal Alargada	0.6331	0.6547	0.6415
Consonante Velar	0.6747	0.5580	0.6091
Consonante Labiodental	0.6765	0.6526	0.6624
Consonante Bilabial	0.5873	0.7932	0.6713
Consonante Dental	0.6049	0.6231	0.6117
Consonante Alveolar	0.6423	0.5949	0.5885
Consonante Palatal	0.6143	0.6240	0.6171

Tabla 10 Métricas por grupo de Fonemas

Experimento 3: Uso de ángulos fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

Otro experimento que se realizó fue la obtención de los ángulos que se forman con los 4 puntos que conforman la boca, con ellos se calculó la distancia de similitud DTW y posteriormente se utilizaron estas para la clasificación con el algoritmo K-NN. Para el grupo de fonemas formado por /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/ se obtuvo la matriz de confusión mostrada en la figura 29.

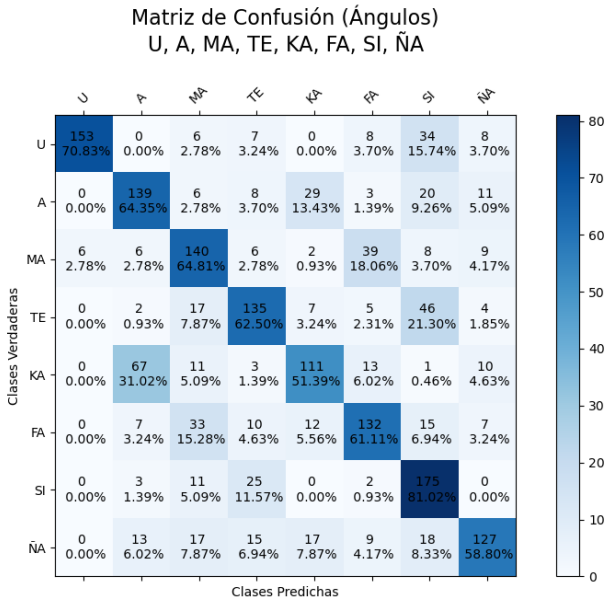


Fig. 28 Matriz de Confusión de los Fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/ con el uso de ángulos

Los valores de las métricas para grupo de fonemas se muestran en la tabla 11.

Métricas Macro Promedio Fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

	Precision	Recall	F1-score
Macro Promedio	0.66230	0.64352	0.64550

Accuracy	0.64352

Clase	Precision	Recall	F1-score
A	0.58649789	0.64351852	0.613686534
FA	0.625592417	0.61111111	0.618266979
KA	0.623595506	0.51388889	0.563451777
MA	0.580912863	0.64814815	0.612691466
NNA	0.721590909	0.58796296	0.647959184
SI	0.552050473	0.81018519	0.656660413
TE	0.645933014	0.625	0.635294118
U	0.962264151	0.70833333	0.816

Tabla 11 Métricas Macro promedio y por clase para el grupo de Fonemas usando Ángulos /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

La grafica ROC correspondiente se muestra en la figura 30, con un AUC promedio de 0.929 para la gráfica A y 0.79 para la gráfica B.

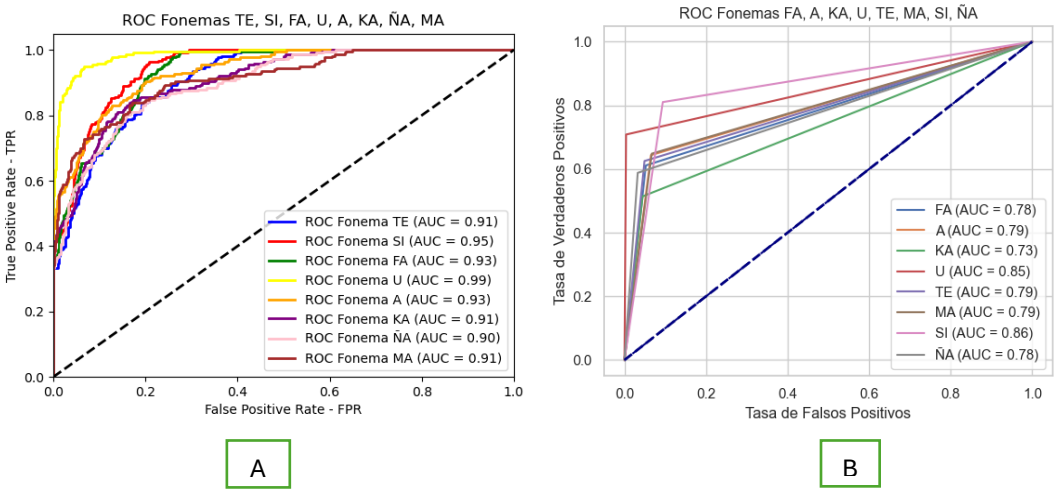


Fig. 29 Grafica ROC del grupo de fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/. La gráfica A construida a partir de probabilidades normalizadas y B usando One-Hot Encoding

Los resultados promedio obtenidos utilizando los ángulos de las métricas ejecutando el algoritmo de clasificación con los 1296 grupos de datos, el accuracy

promedio del modelo fue de 0.653 y las métricas por grupo se muestran en la siguiente tabla 12.

Métricas Promedio Ángulos

Clase	Precision	Recall	F1-score
Vocal Redonda	0.96439945	0.72125772	0.822560743
Vocal Alargada	0.63389144	0.65484254	0.641946592
Consonante Velar	0.67438408	0.5573131	0.608596351
Consonante Labiodetal	0.67561976	0.65215263	0.661773972
Consonante Bilabial	0.5864622	0.79249186	0.670577579
Consonante Dental	0.60450971	0.6233782	0.611678866
Consonante Aveolar	0.64260282	0.59552898	0.588895789
Consonante Palantal	0.61487415	0.62388189	0.617403133

Tabla 12 Métricas promedio por grupo de fonemas haciendo uso de los ángulos

Comparando resultados Puntos Vs Ángulos

Un comparativo de los resultados obtenidos para el mismo grupo de fonemas /U/, /A/, /MA/ ,/TE/,/KA/,/FA/, /SI/, /ÑA/ utilizando los puntos que forman en contorno de los labios y los ángulos podemos observar en la tabla 13 . El uso de ángulos mejora para algunos fonemas el desempeño del algoritmo de clasificación. Por ejemplo, en el caso del fonema /A/ aunque la precisión de incrementa muy poco (de 0.576 a 0.589), la métrica que nos indica la capacidad del modelo para evitar falsos positivos (recall) se incrementa de 0.625 con el uso de puntos a 0.644 con el uso de los ángulos. Así mismo el valor que nos indica el equilibrio entre falsos positivos y negativos (F1-Score) se incrementó con el uso de los ángulos a 0.614.

Comparativo de Métricas obtenidas utilizando Ángulos y Puntos para los fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

	Precision		Recall		F1-Score	
Clase	**Ángulos**	**Puntos**	**Ángulos**	**Puntos**	**Ángulos**	**Puntos**
A	0.586	0.579	0.644	0.625	0.614	0.601
U	0.962	0.970	0.708	0.745	0.816	0.843
MA	0.581	0.570	0.648	0.644	0.613	0.604
TE	0.646	0.615	0.625	0.583	0.635	0.599
KA	0.624	0.622	0.514	0.532	0.563	0.574
FA	0.626	0.663	0.611	0.630	0.618	0.646
SI	0.552	0.538	0.810	0.782	0.657	0.638
ÑA	0.722	0.705	0.588	0.574	0.648	0.633

Tabla 13 Comparativo de resultados obtenidos en las Métricas utilizando Ángulos y Puntos para los fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

En el caso del fonema /U/ las métricas no incrementaron con el uso de ángulos. En la tabla 13 podemos observar que al igual que fonema /a/ el fonema /TE/ incrementó sus valores de las métricas, no todos los fonemas mejoran con el uso de los ángulos, tal es el caso de los fonemas /FA/ y /KA/.

Los resultados macro promedio de las métricas utilizando ángulos y puntos se muestran en la tabla 14. Podemos observar que para el grupo de fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/ las métricas mejoran ligeramente usando la distancia de similitud obtenida con los ángulos que se forman en los 4 puntos de la boca.

Métricas Fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

Macro Promedio	Ángulos	Puntos
Precision	0.662	0.658
Recall	0.644	0.639
F1-Score	0.646	0.642
Accuracy	0.639	0.644

Tabla 14 Resultados obtenidos para las métricas Macro Promedio Fonemas /U/,/A/,/MA/,/TE/,/KA/,/FA/,/SI/,/ÑA/

En la tabla 14 se observa que con el uso de ángulos el modelo tiene una ligera mejor en las métricas precision, recall y F1-Score, sin embrago es solo un grupo de los 1296 grupos de fonemas. En la tabla 15 se muestran los resultados obtenidos para la clasificación para el total de grupos de fonemas y se observa que no hay una mejora significativa en el desempeño del modelo haciendo uso de ángulos o puntos. El accuracy del modelo usando puntos es de 0.6526 comparado con el obtenido con el uso de ángulos de 0.6527, que tampoco representan una diferencia significativa en el desempeño del modelo.

	Precision		Recall		F1-score	
Clase	**Ángulos**	**Puntos**	**Ángulos**	**Puntos**	**Ángulos**	**Puntos**
Vocal Redonda	0.9644	0.9647	0.7213	0.7214	0.8226	0.8228
Vocal Alargada	0.6339	0.6331	0.6548	0.6547	0.6419	0.6415
Cons Velar	0.6744	0.6747	0.5573	0.5580	0.6086	0.6091
Cons Labiodetal	0.6756	0.6765	0.6522	0.6526	0.6618	0.6624
Cons Bilabial	0.5865	0.5873	0.7925	0.7932	0.6706	0.6713
Cons Dental	0.6045	0.6049	0.6234	0.6231	0.6117	0.6117
Cons Aveolar	0.6426	0.6423	0.5955	0.5949	0.5889	0.5885
Cons Palantal	0.6149	0.6143	0.6239	0.6240	0.6174	0.6171

Tabla 15 Comparativo de los resultados obtenidos para las métricas por clase de fonemas utilizando ángulos y puntos

Experimento 3: Vocales

Se realizó la evaluación del modelo con los fonemas correspondientes a las vocales. Para este problema de clasificación de 5 clases, uno para cada vocal se obtuvieron los resultados mostrados en la tabla 16.

Métricas Macro Promedio
Vocales

Accuracy:	0.70
Precision:	0.73
Recall:	0.70
F1-score:	0.71

Métricas por Fonema

	Precision	Recall	F1-Score
A	0.78	0.72	0.75
E	0.61	0.63	0.62
I	0.54	0.71	0.62
O	0.78	0.73	0.75
U	0.87	0.7	0.78

Tabla 16 Métricas Macro promedio y por clase para los fonemas correspondientes a las vocales

Los valores de las métricas precision, recall, f1-score tienen valores del 0.70 y superiores, por lo que se considera un buen modelo para la clasificación de los fonemas asociados a las vocales. La Matriz de confusión correspondiente a las vocales se muestra en la figura 31 y la gráfica ROC en la figura 32.

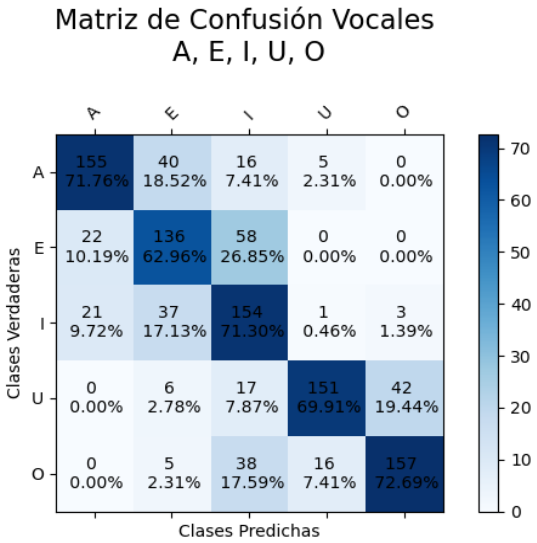


Fig. 30 Matriz de confusión correspondiente a las vocales

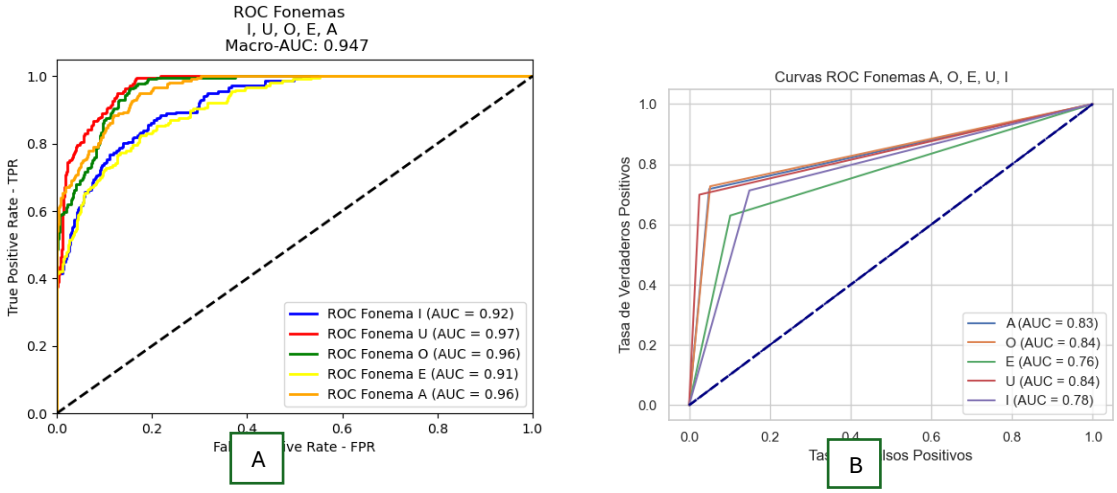


Fig. 31 Grafica ROC para los fonemas correspondientes a las vocales. La gráfica A tiene un Macro Auc de 0.94 y la gráfica B de 0.81

Uno de los principales retos de la lectura labial son los llamados homofonemas, es decir caracteres cuyos movimientos labiales son casi iguales y son difíciles de distinguir, por ejemplo /p/ y /b/, aunado a esto, en los grupos de fonemas que existen elementos que no son distinguibles visualmente y que juegan un rol importante en la producción del sonido, tal es el caso de las consonantes labiodentales, donde la posición de la lengua en los dientes es importante para la producción del sonido, otro ejemplo son el grupo de las consonantes alveolares donde la lengua toca o se encuentra muy cerca de alveolos para producir el sonido.

Analizando los resultados obtenidos de todos los grupos de fonemas podemos observar que las vocales redondas son en su mayoría clasificadas correctamente por el algoritmo propuesto, con una precisión del 96%, un valor de 72% para evitar los falsos positivos (recall), y un F1 de 82% lo que nos indica que mantiene un equilibrio en evitar falsos positivos y negativos. De igual forma podemos observar que los resultados más bajos se obtuvieron para la clase Alveolar, sobre todo para la métrica F1 donde su valor está en 58% lo que nos indica dificultad para poder identificar correctamente los falsos positivos y negativos de ese grupo de fonemas, como se menciona anteriormente la posición de la lengua, no visible por nuestra propuesta, juega un papel importante en la producción del sonido.

5.3 Discusión

Puesto que el desempeño del modelo no tuvo una diferencia significativa al utilizar para la clasificación ángulos, respecto a los puntos del contorno de los labios realizaremos el análisis de resultados con estos últimos.

En la figura 32 se muestra la gráfica con el accuracy obtenido por los grupos de fonemas. El accuracy promedio de los grupos fue de 0.6527, el valor máximo obtenido fue de 0.7188 correspondiente al grupo 193 que está integrado por los fonemas /GA/, /PA, /A/, /SI/,/ÑA/, /FA /, /U/, /DA/. El valor mínimo fue de 0.5938 correspondiente a grupo /E/,/LA/,/JA/, /FA/, /CHE/, /BE/, /TE/, /U/.

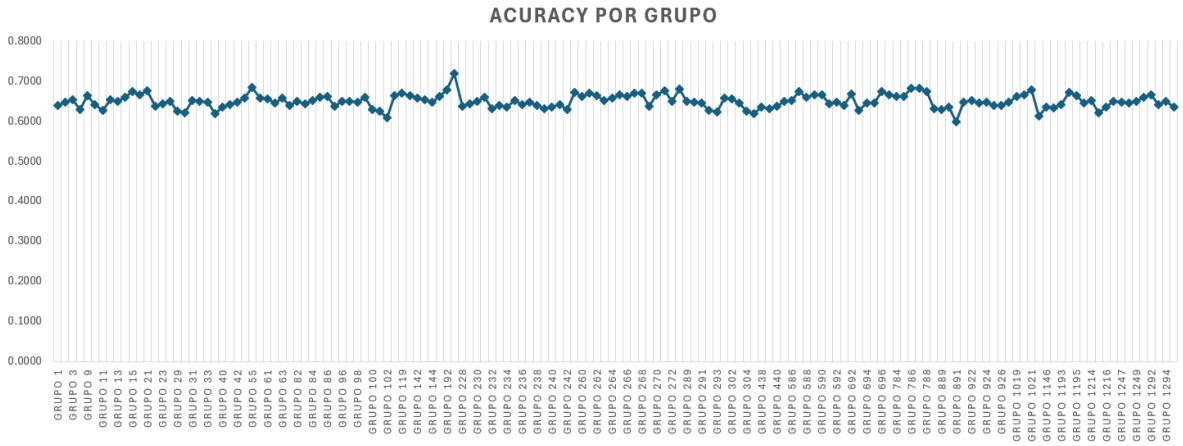


Fig. 32 Accuracy de grupos de fonemas seleccionados al azar.

Los valores promedio máximos obtenidos de las métricas accuracy, recall y F1-score corresponden al grupo 193 y está integrado por los fonemas /GA/,/PA7,/A/,/SI/,/ÑA/,/FA/,/U/,/DA/ como se muestra en la tabla 17. También, observamos que los valores mínimos de estas métricas corresponden al mismo grupo de fonemas (grupo 533) /E/,/LA/,/JA/,/FA/,/CHE/,/BE/,/TE/,/U/ . Para el caso de la precisión el valor máximo corresponde al mismo grupo de fonemas de las métricas anteriores, pero el valor mínimo corresponde al grupo de fonemas 466 integrado por /E/,/TE/,/MA/,/RA/,/JA/,/FA/,/U/,/ÑA/ (tabla 18).

	Accuracy	Recall	F1-score	Grupo	Fonemas
Valor Máximo	0.719	0.719	0.720	193	/GA/,/PA/,/A/,/SI/,/ÑA/,/FA/,/U/,/DA/
Valor Mínimo	0.594	0.594	0.590	533	/E/,/LA/,/JA/,/FA/,/CHE/,/BE/,/TE/,/U/

Tabla 17 Valores mínimos y máximos de las métricas accuracy, recall y f1-score de los 1296 grupos de prueba

	Precision	Grupo	Fonemas
Valor Máximo	0.737	193	/GA/,/PA/,/A/,/SI/,/ÑA/,/FA/,/U/,/DA/
Valor Mínimo	0.610	466	/E/,/TE/,/MA/,/RA/,/JA/,/FA/,/U/,/ÑA/

Tabla 18 Valor mínimo y máximo obtenido para la métrica precision y grupo de fonemas correspondientes

La figura 33 muestra la gráfica ROC y la matriz de confusión del grupo de fonemas con los valores promedio más altos en las métricas accuracy, precisión recall y f1-score (grupo 193). Podemos observar que ningún fonema tiene un porcentaje de verdaderos positivos menores a 60%. También podemos decir que el modelo tiene buena calidad para realizar predicciones, ya que de todos los casos que el modelo predijo como positivos el 73.7% fueron realmente positivos, además identificó el 71.9% de todos los casos positivos reales, y F1-score de 0.720 sugiere que el modelo se está desempeñando razonablemente bien, considerando tanto su capacidad para identificar verdaderos positivos (precisión) como para evitar falsos positivos (recall). EL AUC promedio es de 0.83 indica que el desempeño del modelo puede ser considerado como bueno.

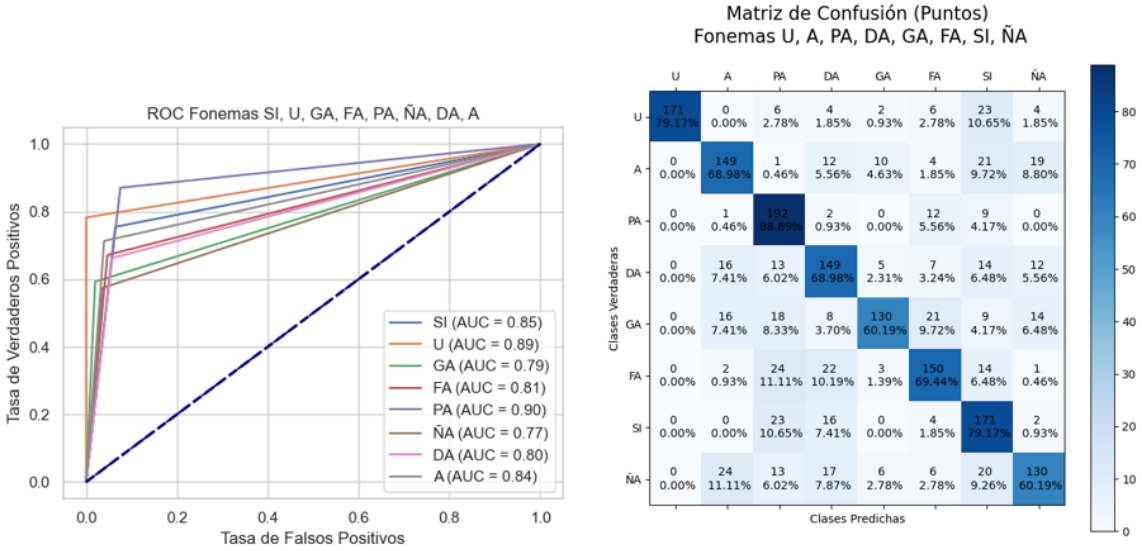


Fig. 33 Gráfica Roc y matriz de confusión del grupo de fonemas con valores más altos en sus métricas

Analizando las métricas de cada fonema del grupo (tabla 19) observamos que el fonema /GA/ y /ÑA/ son lo que tienen valor más bajo en la métrica de recall. Estos fonemas pertenecen a las clases velar y palatal. Para producir el sonido de estos grupos de fonemas la posición de la lengua es importante. Para los fonemas del grupo velares la lengua tocando el avelo y para las palatales la lengua tocando el paladar, la lengua no es un elemento que se identifiquen en este trabajo. En las

imágenes del método Adryna que distinguen estos fonemas podemos observar que existe gran relación entre de ellas y el resto de las imágenes. La imagen que más se distingue es el fonema /PA/ de la clase bilabial, para este fonema el modelo tiene mayor capacidad de identificar correctamente las instancias positivas de este fonema.

Analizando el detalle de los fonemas de grupo, observamos que la precision más baja la obtuvieron los fonemas /DA/ y /SI/ y el valor más bajo de F1-Score para los fonemas /DA/ y /ÑA/.

Fonema	Imagen	Clase	Precision	Recall	F1-Score
A		*Vocal Alargada*	0.716	0.690	0.703
DA		*Dental*	0.648	0.690	0.668
FA		*Labiodental*	0.714	0.694	0.704
GA		*Velar*	0.833	0.602	0.699
ÑA		*Palatal*	0.714	0.602	0.653
PA		*Bilabial*	0.662	0.889	0.759
SI		*Alveolar*	0.609	0.792	0.688
U		*Vocal Redonda*	1.000	0.792	0.884

Tabla 19 Métricas del grupo de fonemas 193, con valores de recall y F1-Score más altos de todos los grupos. Las imágenes mostradas fueron tomadas de la referencia [8] sólo con propósitos educativos y de investigación.

Con respecto a los valores obtenidos con la métrica accuracy sólo 8 grupos obtuvieron un valor menor a 0.6, sin embargo, este valor es más cercano al 0.6 que al 0.5 como se puede observar en la tabla 20.

Analizando las matrices de confusión de estos fonemas y las métricas se observó que los fonemas /LA/, /MA/, /JA/, /YA/ y /FA/ suele confundirlas el modelo. Por ejemplo, en la matriz de confusión de la Figura 35 podemos observar que cuando entra un fonema /LA/ el 21.76% de las veces lo predice como /JA/ (clase velar) y un 11.11% como /FA/ (clase labiodental). Podemos decir que estos fonemas presentan ambigüedad visual, por lo que sería importante utilizar alguna otra característica que permita distinguirlos.

	Accuracy	Fonemas
Grupo 533	0.594	/E/,/LA/,/JA/,/FA/,/CHE/,/BE/,/TE/,/U/
Grupo 893	0.594	/O/,/I/,/LA/,/YA/,/MA/,/JA/,/FA/,/TE/
Grupo 1107	0.598	/O/,/E/,/LA/,/MA/,/JA/,/FA/,/TE/,/ÑA/
Grupo 891	0.598	/O/,/I/,/LA/,/MA/,/JA/,/FA/,/TE/,/ÑA/
Grupo 892	0.598	/O/,/I/,/LA/,/MA/,/JA/,/FA/,/CHE/,/TE/
Grupo 965	0.598	/O/,/I/,/LA/,/YA/,/JA/,/FA/,/BE/,/TE/
Grupo 318	0.599	/I/,/LA/,/YA/,/JA/,/FA/,/BE/,/TE/,/U/
Grupo 928	0.599	/O/,/I/,/LA/,/MA/,/JA/,/FA/,/CHE/,/DA/

Tabla 20 Grupos con Accuracy por debajo del 0.6

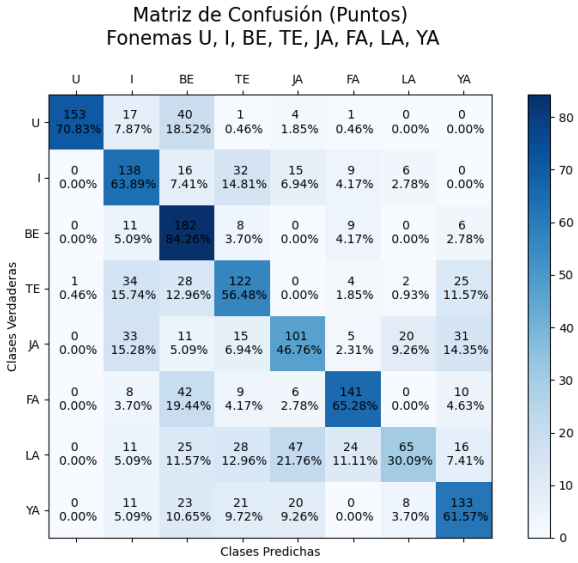


Fig. 34 Matriz de confusión del grupo de fonemas 318 correspondiente a los fonemas /I/,/LA/,/YA/,/JA/,/FA/,/BE/,/TE/,/U/

Con respecto a las métricas generales por clase, el modelo tiene una buena capacidad para distinguir los fonemas correspondientes a la clase Vocal Redonda, donde se encuentran los fonemas /U/ y /O/ ya que se obtuvo un 72.1% de certeza que se clasificaran correctamente. Además, observamos una capacidad del modelo de 82.2%, podemos decir que es un buen modelo para identificar tanto los verdaderos positivos (precision) como para evitar falsos positivos (recall).

Para las clases vocal alargada, velar, labiodental, bilabial, dental, y palantal tienen una capacidad mayor al 60% para identificar tanto los verdaderos positivos como los falsos positivos. Con valor de 68.8% en la métrica F1-score se encuentra la clase Aveolar, compuesta por los fonemas /SI/,/LA/, /NO/,/RA/.

En la tabla 21 observamos las métricas promedio del modelo para todos los grupos, podemos concluir que es un buen modelo para clasificar por grupos de fonemas ya que 67,47% de las predicciones positivas del modelo fueron correctas. Identificó

correctamente el 65,27% de todos los casos positivos reales. El resultado obtenido en la métrica F1-Score (0.6531) nos sugiere que el modelo se está desempeñando moderadamente bien, considerando tanto su capacidad para identificar verdaderos positivos (precisión) como para evitar falsos positivos (recall).

	Métricas Promedio del Modelo			
	Accuracy	**Precision**	**Recall**	**F1-score**
Promedio	0.653	0.675	0.653	0.653
Desviación Estándar	0.02136	0.02181	0.02136	0.02186

Tabla 21 Métricas promedio del modelo

En la tabla 21 se observa que la desviación estándar de las métricas es un valor muy pequeño, lo que permite afirmar que:

- existe una baja dispersión en los valores de las métricas, esto significa que sus valores se encuentran muy cercanos a la media.
- el modelo propuesto es consistente y con muy poca incertidumbre, lo que refleja que será capaz de generalizar nuevos datos sin que el rendimiento del modelo se vea afectado significativamente.

Revisando los trabajos relacionados encontramos algunos que hacen el proceso de reconocimiento de caracteres o fonemas basados exclusivamente con datos de video (tabla 22) con los que compararemos considerando la métrica de accuracy que fue la reportada por los autores y que calculamos para el modelo propuesto.

Titulo	Idioma	Tarea de Reconocimiento	Accuracy
End-to-end visual speech recognition for small-scale datasets [11]	*Inglés*	*Caracteres*	*AvLetters : 65.9 AvLetters-2: 36.8*
End-to-End Lip-Reading Without Large-Scale Data [10]	*Español*	*Fonemas*	*46.24*
Developing phoneme -based lip -reading sentences system for silent speech recognition [9]	*Inglés*	*Fonemas*	*70.0*
Nuestra Propuesta	*Español*	*Fonemas*	*65.2*

Tabla 22 Resultados obtenidos en los trabajos relacionados y nuestra propuesta.

En [10] realizan la clasificación de fonemas usando redes neuronales convolucionales obtenido un accuracy de 46.24% para la clasificación de fonemas.

En [11] encontramos que obtuvieron un accuracy promedio de 66.3% con las bases de datos AvLetters y 36.8% con la base de datos AvLetters, usando un solo canal de entrada, imágenes RGB extraídas del video como en nuestro caso.

Finalmente, en [9] se reporta una accuracy de 70% para el reconocimiento de fonemas utilizando redes neuronales para esta tarea.

El accuracy obtenido en la presente propuesta alcanza un promedio de 65.2 %, utilizando para la clasificación el algoritmo KNN basado en la distancia de similitud entre los frames obtenida con el algoritmo DTW. Considerando los modelos propuestos para la clasificación en los trabajos mencionados en la tabla 22, podemos considerar que con la presente propuesta se obtuvieron resultados competitivos.

En las investigaciones revisadas sobre de lectura de labios [4], [5], [9], [10], [13], [32] se argumentan problemas de reconocimiento en aquellos fonemas donde las imágenes visuales que representan el sonido son muy difíciles de distinguir, repercutiendo esto en los resultados de reconocimiento de los fonemas, mismos problemas que corroboramos en este trabajo.

Después de analizar los resultados y respondiendo las preguntas de investigación (sección 1.7), podemos concluir lo siguiente:

1. Es posible realizar identificación de fonemas del idioma español con la información obtenida del contorno de los labios de las secuencias del video, sin embargo, los resultados obtenidos en las métricas pudieran mejorarse utilizando otras características que ayuden a distinguir las ambigüedades visuales.

2. Por otro lado, la distancia de similitud obtenida con el algoritmo DTW aplicado a las secuencias de video, es una medida útil para hacer la clasificación de fonemas usando el algoritmo KNN.

3. Finalmente, el algoritmo KNN nos permite obtener resultados competitivos cuando trabajamos con conjuntos de datos formados por centenas de videos, considerando que en la literatura se utilizan en su mayoría modelos basados en redes neuronales y conjuntos de datos más grandes.

En cuanto a las limitaciones de la propuesta presentada en esta tesis, fue que a pesar de que se les explico a los participantes el método Adryna y la gesticulación sugerida, en las muestras que se les tomo, los participantes les costó seguir el método y pronunciaban los fonemas gesticulando de manera habitual. Otra limitación que pudo haber impedido mejores resultados fue el proceso de

depuración de videos, ya que éste se llevó a cabo de forma manual, analizar los 504 videos y eliminar los frames que no abonaba al fonema quedo a discreción del ojo humano, lo que puede no ser muy exacto o correcto.

6. Conclusiones y trabajos futuros

Los resultados arrojados en los experimentos realizados nos muestran que la lectura de labios basada en fonemas utilizando la distancia de similitud entre los diferentes frames de un fonema y la utilización del algoritmo de clasificación nos permite obtener buenos resultados de reconocimiento. Algunas de las razones por las que la propuesta es considerada viable para la lectura de labios son:

- El tiempo que le toma a cada persona pronunciar un fonema es distinto. Aquí la distancia de similitud nos ayuda a eliminar el ruido que esto pudiera producir al momento de comparar videos de un mismo fonema.
- KNN es un algoritmo que ofrece buenos resultados utilizando la distancia de similitud para conjuntos de pequeños, comparado con los necesarios para entrenar modelos de redes neuronales como las mencionadas en los trabajos relacionados, además de la capacidad de cómputo requerida por estos.
- Al comparar los resultados se observa que los resultados obtenidos son competitivos, considerado la cantidad de datos y la arquitectura utilizada.
- Los conjuntos de datos audiovisuales para el idioma español son escasos. Para la creación de un buen sistema de ASR para el idioma español mexicano es importante construir conjuntos de datos multimodales que integren video RGB y de profundidad, así como audio que permitan explorar técnicas que mejoren los resultados actuales. Este trabajo contribuye con la construcción de un conjunto de datos multimodal.

El tema de las ambigüedades visuales es un problema vigente e independiente de la arquitectura que se utilice para la construcción de los sistemas de lectura labial. Por otro lado, el uso de fonemas para la lectura de labios como el utilizado en este trabajo, ofrece mejores resultados que los visemas, ya que al hacer el mapeo de fonema a visema existe perdida de información causado posiblemente por que el número de visemas es menor al de fonemas[9]. Los visemas son un área de investigación vigente.

El reconocimiento de fonemas permite la construcción de palabras para el reconocimiento de frases a un costo computacional más bajo.

La lectura automática de labios permite la construcción de sistemas de reconocimiento del habla más robustos. Además, contribuye a la creación de interfaces que mejoren la interacción humano-máquina para personas que han visto afectada parcial o total su capacidad de habla.

Como trabajo futuro se considera la integración de la información que se puede obtener de los frames de profundidad para la clasificación, así como la identificación de otras características visuales que ayuden a la disminución de las ambigüedades y mejoren los resultados obtenidos.

Bibliografía

[1] H. Mcgurk y J. Macdonald, «Hearing lips and seeing voices», *Nature*, vol. 264, n.º 5588, Art. n.º 5588, dic. 1976, doi: 10.1038/264746a0.

[2] A. A. Hill, «William Freeman Twaddell», *Language*, vol. 59, n.º 2, pp. 347-354, 1983.

[3] yulin-li, «Obtención de la posición facial con visema - Azure Cognitive Services». Accedido: 4 de marzo de 2023. [En línea]. Disponible en: https://learn.microsoft.com/es-es/azure/cognitive-services/speech-service/how-to-speech-synthesis-viseme

[4] A. Fernandez-Lopez y F. M. Sukno, «Survey on automatic lip-reading in the era of deep learning», *Image Vis. Comput.*, vol. 78, pp. 53-72, oct. 2018, doi: 10.1016/j.imavis.2018.07.002.

[5] Ronquillo, Yadira, «Continuous lip reading in Spanish», 2022, Accedido: 4 de marzo de 2023. [En línea]. Disponible en: http://repositori.upf.edu/handle/10230/54585

[6] «RTVE2022DB.pdf». Accedido: 4 de marzo de 2023. [En línea]. Disponible en: http://catedrartve.unizar.es/reto2022/RTVE2022DB.pdf

[7] «Población. Discapacidad». Accedido: 23 de febrero de 2023. [En línea]. Disponible en: https://cuentame.inegi.org.mx/poblacion/discapacidad.aspx

[8] Trinidad Caparrós López, *Método Adryna: aprendizaje del habla con apoyo familiar*. Asociación Adryna. Asociación Integral a Discapacitados, 2018.

[9] R. El-Bialy *et al.*, «Developing phoneme-based lip-reading sentences system for silent speech recognition», *CAAI Trans. Intell. Technol.*, vol. 8, n.º 1, pp. 129-138, mar. 2023, doi: 10.1049/cit2.12131.

[10] A. Fernandez-Lopez y F. M. Sukno, «End-to-End Lip-Reading Without Large-Scale Data», *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2076-2090, 2022, doi: 10.1109/TASLP.2022.3182274.

[11] S. Petridis, Y. Wang, P. Ma, Z. Li, y M. Pantic, «End-to-end visual speech recognition for small-scale datasets», *Pattern Recognit. Lett.*, vol. 131, pp. 421-427, mar. 2020, doi: 10.1016/j.patrec.2020.01.022.

[12] S. Cox, R. Harvey, Y. Lan, J. Newman, y B.-J. Theobald, «The challenge of multispeaker lip-reading», presentado en Proc. International Conference on Auditory-Visual Speech Processing, 2008, pp. 179-184.

[13] S. Fenghour, D. Chen, K. Guo, y P. Xiao, «Lip Reading Sentences Using Deep Learning With Only Visual Cues», *IEEE Access*, vol. 8, pp. 215516-215530, 2020, doi: 10.1109/ACCESS.2020.3040906.

[14] B. Xu, J. Wang, C. Lu, y Y. Guo, «Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition», en *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA: IEEE, mar. 2020, pp. 1626-1635. doi: 10.1109/WACV45572.2020.9093314.

[15] X. Chen, J. Du, y H. Zhang, «Lipreading with DenseNet and resBi-LSTM», *Signal Image Video Process.*, vol. 14, n.º 5, pp. 981-989, jul. 2020, doi: 10.1007/s11760-019-01630-1.

[16] P. G, S. A, M. K, H. D, y K. Renuka, «Speaker-Independent Speech Recognition using Visual Features», *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, n.º 11, 2020, doi: 10.14569/IJACSA.2020.0111175.

[17] S. Debnath y P. Roy, «Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition», *Signal Image Video Process.*, vol. 15, n.º 1, pp. 25-32, feb. 2021, doi: 10.1007/s11760-020-01717-0.

[18] L. Liu, G. Feng, D. Beautemps, y X.-P. Zhang, «Re-Synchronization Using the Hand Preceding Model for Multi-Modal Fusion in Automatic Continuous Cued Speech Recognition», *IEEE Trans. Multimed.*, vol. 23, pp. 292-305, 2021, doi: 10.1109/TMM.2020.2976493.

[19] D. Tsourounis, D. Kastaniotis, y S. Fotopoulos, «Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions», *J. Imaging*, vol. 7, n.º 5, p. 91, may 2021, doi: 10.3390/jimaging7050091.

[20] N. P. Akman, T. T. Sivri, A. Berkol, y H. Erdem, «Lip Reading Multiclass Classification by Using Dilated CNN with Turkish Dataset», en *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Prague, Czech Republic: IEEE, jul. 2022, pp. 1-6. doi: 10.1109/ICECET55527.2022.9873011.

[21] Ü. Atila y F. Sabaz, «Turkish lip-reading using Bi-LSTM and deep learning models», *Eng. Sci. Technol. Int. J.*, vol. 35, p. 101206, nov. 2022, doi: 10.1016/j.jestch.2022.101206.

[22] I. Ullah, H. Zahid, F. Algarni, y M. Asghar Khan, «Deep Learning-Based Approach for Arabic Visual Speech Recognition», *Comput. Mater. Contin.*, vol. 71, n.º 1, pp. 85-108, 2022, doi: 10.32604/cmc.2022.019450.

[23] G. Xing, L. Han, Y. Zheng, y M. Zhao, «Application of deep learning in Mandarin Chinese lip-reading recognition», *EURASIP J. Wirel. Commun. Netw.*, vol. 2023, n.º 1, p. 90, sep. 2023, doi: 10.1186/s13638-023-02283-y.

[24] T. Arakane y T. Saitoh, «Efficient DNN Model for Word Lip-Reading», *Algorithms*, vol. 16, n.º 6, p. 269, may 2023, doi: 10.3390/a16060269.

[25] N. Faisal Aljohani y E. Sami Jaha, «Visual Lip-Reading for Quranic Arabic Alphabets and Words Using Deep Learning», *Comput. Syst. Sci. Eng.*, vol. 46, n.º 3, pp. 3037-3058, 2023, doi: 10.32604/csse.2023.037113.

[26] U. Kamath, J. Liu, y J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-14596-5.

[27] D. Yu y L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. en Signals and Communication Technology. London: Springer London, 2015. doi: 10.1007/978-1-4471-5779-3.

[28] S. Bhaskar y T. T. Madathil, «Multimodal Based Audio-Visual Speech Recognition for Hard-of-Hearing: State of the Art Techniques and Challenges», *Indones. J. Electr. Eng. Inform. IJEEI*, vol. 10, n.º 2, pp. 385-397, may 2022, doi: 10.52549/ijeei.v10i2.3683.

[29] L. Xia, G. Chen, X. Xu, J. Cui, y Y. Gao, «Audiovisual speech recognition: A review and forecast», *Int. J. Adv. Robot. Syst.*, vol. 17, n.º 6, p. 172988142097608, nov. 2020, doi: 10.1177/1729881420976082.

[30] A. Das, S. Patikar, y K. Medhi, «A survey on Audio-Visual Speech Recognition System», en *2023 4th International Conference on Computing and Communication Systems (I3CS)*, Shillong, India: IEEE, mar. 2023, pp. 1-5. doi: 10.1109/I3CS58314.2023.10127316.

[31] A. K. Jheel y Khadhim Mahdi Hashim, «Automatic lip reading classification using artificial neural network», *J. Manag. Inf. Decis. Sci.*, vol. Volume 24, 2021, doi: 10.13140/RG.2.2.19499.52007.

[32] Y. Lu, J. Yan, y K. Gu, «Review on Automatic Lip Reading Techniques», *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, n.º 07, p. 1856007, jul. 2018, doi: 10.1142/S0218001418560074.

[33] H. L. Bear, R. W. Harvey, B.-J. Theobald, y Y. Lan, «Which phoneme-to-viseme maps best improve visual-only computer lip-reading?», 3 de octubre de 2017. Accedido: 16 de marzo de 2024. [En línea]. Disponible en: http://arxiv.org/abs/1710.01093

[34] S. Petridis, J. Shen, D. Cetin, y M. Pantic, «Visual-Only Recognition of Normal, Whispered and Silent Speech», presentado en IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada: IEEE, 2018, pp. 6219-6223. doi: 10.1109/ICASSP.2018.8461596.

[35] R. Goecke y J.B. Millar, «The Audio-Video Australian English Speech Data Corpus AVOZES», presentado en Proceedings of the 8th International Conference on Spoken Language Processing INTERSPEECH 2004 – ICSLP, Jeju, Kopea, 2004, pp. 2525-2528.

[36] Bowon Lee *et al.*, «AVICAR: audio-visual speech corpus in a car environment», presentado en Proc. Interspeech 2004, 2004, pp. 2489-2492. doi: 10.21437/Interspeech.2004-424.

[37] A. Ortega *et al.*, «AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition», en *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa, Portugal: European Language Resources Association (ELRA), 2004. [En línea]. Disponible en: http://www.lrec-conf.org/proceedings/lrec2004/pdf/389.pdf

[38] E. K. Patterson, S. Gurbuz, Z. Tufekci, y J. N. Gowdy, «CUAVE: A new audio-visual database for multimodal human-computer interface research», presentado en International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA: IEEE, 2002, pp. 2027-2020. doi: 10.1109/ICASSP.2002.5745028.

[39] E. Bailly-Bailliére *et al.*, «The BANCA Database and Evaluation Protocol», en *Audio- and Video-Based Biometric Person Authentication*, vol. 2688, J. Kittler y M. S. Nixon, Eds., en Lecture Notes in Computer Science, vol. 2688. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 625-638. doi: 10.1007/3-540-44887-X_74.

[40] K. Messer, J. Matas, J. Kittler, J. Luettin, y G. Maitre, «XM2VTSDB: The Extended M2VTS Database», *Second Int. Conf. Audio Video-Based Biom. Pers. Authentication*, vol. 964, pp. 965-966.

[41] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, y R. Harvey, «Extraction of visual features for lipreading», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, n.º 2, pp. 198-213, feb. 2002, doi: 10.1109/34.982900.

[42] A. Fernandez-Lopez, O. Martinez, y F. M. Sukno, «Towards Estimating the Upper Bound of Visual-Speech Recognition: The Visual Lip-Reading Feasibility Database», en *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA: IEEE Computer Society, 2017, pp. 208-215. doi: 10.1109/FG.2017.34.

[43] D. Gimeno-Gómez y C.-D. Martínez-Hinarejos, «LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild», *IberSPEECH 2021*, pp. 2750-2758, 2022, doi: 10.21437/IberSPEECH.2021-47.

[44] C. M. Bishop, *Pattern recognition and machine learning*. en Information science and statistics. New York: Springer, 2006.

[45] K. P. Murphy, *Machine learning: a probabilistic perspective*. en Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012.

[46] J. I. Hualde y S. Colina, *Los sonidos del español*. Cambridge ; New York: Cambridge University Press, 2014.

[47] J. Muñoz-Basols, N. Moreno, T. Inma, y M. Lacorte, *Introducción a la lingüística hispánica actual: teoría y práctica*, 0 ed. Routledge, 2016. doi: 10.4324/9780203096758.

[48] G. Coloma, «IMPORTANCIA ECONÓMICA DE LAS CARACTERÍSTICAS FONÉTICAS DEL IDIOMA ESPAÑOL Y SUS VARIEDADES REGIONALES».

[49] D. Gimeno Gómez, «Lectura de labios mediante técnicas de machine learning», Universitat Politècnica de València, Departamento de Sistemas Informáticos y Computación, 2020. Accedido: 5 de marzo de 2024. [En línea]. Disponible en: http://polipapers.upv.es/index.php/IA/article/view/3293

[50] V. Kazemi y J. Sullivan, «One millisecond face alignment with an ensemble of regression trees», en *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH: IEEE, jun. 2014, pp. 1867-1874. doi: 10.1109/CVPR.2014.241.

[51] R. Tavenard *et al.*, «Tslearn, A Machine Learning Toolkit for Time Series Data», *J. Mach. Learn. Res.*, vol. 21, n.º 118, [En línea]. Disponible en: http://jmlr.org/papers/v21/20-091.html

[52] Kazuaki, Tanida, «FastDTW», fastdtw. [En línea]. Disponible en: https://pypi.org/project/fastdtw/

[53] M. Grandini, E. Bagli, y G. Visani, «Metrics for Multi-Class Classification: an Overview», 13 de agosto de 2020, *arXiv*: arXiv:2008.05756. Accedido: 24 de marzo de 2024. [En línea]. Disponible en: http://arxiv.org/abs/2008.05756

[54] T. Fawcett, «An introduction to ROC analysis», *Pattern Recognit. Lett.*, vol. 27, n.º 8, pp. 861-874, jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[55] World Medical Association, «WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. Accedido: 10 de abril de 2024. [En línea]. Disponible en: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

[56] "dlib C++ Library", «dlib C++ Library. Accedido el 20 de enero de 2024. [En línea]. Disponible: http://dlib.net/