Lectura de labios en videos sin audio mediante Dynamic Time Warping y Machine Learning

Sanny Jheremmy Matienzo Reyes

Resumen— Este trabajo presenta el desarrollo de un sistema de lectura de labios basado en Visión por Computador y Machine Learning, con el objetivo de transcribir palabras pronunciadas en videos sin audio. Este sistema busca facilitar la comunicación para personas con dificultades auditivas y explorar nuevas aplicaciones en contextos donde el sonido no es una opción viable. Para su desarrollo, se ha utilizado la librería Mediapipe para la detección de los labios, descriptores de Fourier para representar su movimiento, Dynamic Time Warping (DTW) para medir similitudes entre secuencias temporales y K-Nearest Neighbors (KNN) para la clasificación de palabras. Los resultados muestran que el sistema logra una alta precisión en entornos controlados, pero presenta dificultades al generalizar a diferentes hablantes y cambios en la orientación de la cabeza. Este estudio confirma el potencial de la lectura de labios automatizada y destaca la necesidad de mejorar la robustez del modelo para su aplicación en entornos más diversos.

Palabras clave— Inteligencia Artificial, Visión por Computador, Lectura de Labios, Dynamic Time Warping (DTW), Descriptores de Fourier, K-Nearest Neighbors (KNN), Mediapipe.

Abstract— This work presents the development of a lip-reading system based on Computer Vision and Machine Learning, aiming to transcribe spoken words from videos without audio. This system seeks to facilitate communication for individuals with hearing impairments and explore new applications in scenarios where audio is not a viable option. The development process involved Mediapipe for lip detection, Fourier descriptors for movement representation, Dynamic Time Warping (DTW) to measure temporal sequence similarities, and K-Nearest Neighbors (KNN) for word classification. The results indicate that the system achieves high accuracy in controlled environments but struggles to generalize across different speakers and variations in head orientation. This study highlights the potential of automated lip reading and the need to enhance model robustness for broader real-world applications.

Index Terms— Artificial Intelligence, Computer Vision, Lip Reading, Dynamic Time Warping (DTW), Fourier Descriptors, K-Nearest Neighbors (KNN), Mediapipe.

1 Introducción

La comunicación es una herramienta fundamental en nuestras vidas. Muchas personas ni siquiera pensamos en cómo sería nuestra vida si perdiéramos la capacidad de comunicarnos. Sin embargo, existen personas que luchan día a día contra este problema, como, por ejemplo, las personas sordomudas. Para las personas sordomudas, ésta es una realidad con la que viven todos los días. No pueden escuchar ni hablar de la misma forma que el resto, así que tienen que buscar maneras diferentes de expresarse.

En la mayoría de los casos, las personas sordomudas no pueden desarrollar el habla porque nacieron sin la capacidad de oír, lo que afecta el aprendizaje del habla. Sin embargo, también existen enfermedades que pueden limitar la capacidad del habla. Por ejemplo, si nos vamos a un caso mediático, podemos hablar de Bruce Willis, a quien se le diagnosticó afasia, un trastorno del lenguaje causado por daños en las áreas del cerebro responsables de la comunicación. Otro caso es el de Val Kilmer, a quien, debido a un cáncer de garganta, tuvieron que realizarle una traqueotomía, lo cual limitó su capacidad de hablar.

Figura 1: Retrato de Bruce Willis

A causa de este tipo de situaciones, se crearon técnicas como la lectura de labios. Esta técnica permite entender lo que diga otra persona solo mirando los movimientos de los labios, sin necesidad de sonido. La lectura de labios no solo beneficia a personas con problemas auditivos o de habla, sino que tiene una gran cantidad de usos que pueden ayudar a cualquier persona en su vida cotidiana. Uno de estos usos podría ser, por ejemplo, poder mantener una

- *E-mail de contacto: jheremmymatienzo@gmail.com*
- *Mención realizada: Computación*
- *Trabajo tutorizado por: Coen Antens*
- *Curso 2024/25*

conversación en un lugar muy ruidoso o, al contrario, en lugares donde no se puede hacer ruido, como, por ejemplo, una biblioteca. En general, la lectura de labios puede ser útil en situaciones donde el sonido no es una opción.

Hoy en día, con el avance de la Inteligencia Artificial y la Visión por Computador, es posible hablar de sistemas automáticos capaces de interpretar el movimiento de los labios y transformarlo en palabras, como por ejemplo "*LipType*", un software que permite a las personas enviar mensajes privados mientras se encuentran en un espacio público o en una reunión, con solo "decir" las palabras sin ningún sonido [1]. Gracias a la evolución de este tipo de tecnologías, sería posible utilizar el movimiento de los labios para que, por ejemplo, las personas con un familiar que ha perdido la capacidad de hablar puedan comunicarse de una forma más fácil con ellos.

Este tipo de avances tecnológicos inspiran a seguir ayudando a las personas a no sentirse apartadas en la sociedad e impulsa a seguir investigando sobre el tema. Así que, precisamente sobre este tema, irá el proyecto.

Este proyecto se centrará en el desarrollo de un software capaz de transcribir lo que una persona diga en un video sin audio, pasado como input. Con ayuda de técnicas de Visión por Computador, el software será capaz de detectar el movimiento de los labios, con el que, gracias a la ayuda de la Inteligencia Artificial, podrá transcribir lo que diga la persona en el video.

2 Objetivos

El objetivo principal del proyecto es facilitar la comunicación de personas sordomudas mediante el desarrollo de un sistema basado en Inteligencia Artificial y Visión por Computador, capaz de reconocer palabras pronunciadas a partir del movimiento de los labios en un video.

Para lograr este objetivo general de una forma progresiva, se definieron los siguientes objetivos específicos:

- **Preprocesamiento de datos para el reconocimiento de palabras**:
  Diseñar y desarrollar un pipeline de preprocesamiento que permita identificar los movimientos de la boca en cada *frame*, recortando la región de interés, estandarizando las dimensiones de las imágenes y extrayendo características relevantes, como las coordenadas de diferentes puntos de los labios, para su uso en el análisis posterior.

- **Método para medir similitudes entre movimientos labiales:**
  Investigar y desarrollar un método que permita poner en correspondencia el movimiento de los labios realizado a la hora de pronunciar una palabra en dos videos distintos, y de esta forma, poder medir similitudes entre movimientos.

- **Clasificación de palabras mediante Machine Learning:**
  Diseñar e implementar un sistema que, utilizando las similitudes obtenidas de los movimientos labiales, sea capaz de clasificar palabras pronunciadas en videos sin audio.

3 Estado del arte

3.1 Trabajos relacionados

A lo largo del tiempo, ha habido personas que ya comenzaron a investigar sobre la lectura de labios. Algunos trabajos destacables son:

- **Lipread Net** [2]: mediante el uso de 3DCNN (*3D Convolutional Neural Networks*) y LSTM (*Long Short-Term Memory*), con un 93% de *accuracy*, ha sido capaz de transcribir el habla a partir de únicamente el movimiento de los labios.

- **LipNet** [3]: mediante el uso de STCNNs (*SpatioTemporal Neural Network*), una RNN (*Recurrent Neuronal Network*) y una CTC (*Connectionist Temporal Classification Loss*), ha conseguido desarrollar un sistema *End-to-End* que transcribe oraciones, con un 92% de *accuracy*.

- **LipType** [4]: es una mejora de *LipNet*, pero añadiéndole un modelo de reparación independiente que mejora el reconocimiento en condiciones adversas. *LipType* consiguió su objetivo con un 92.5% de *accuracy* en condiciones controladas y un 85.3% en condiciones adversas.

3.2 Datasets

Para poder desarrollar la gran cantidad de proyectos existentes, se han tenido que utilizar *datasets* para poder entrenar y validar los modelos utilizados. Algunos *datasets* destacables son:

- **AVLetters2** [5]: contiene videos y audios de 5 personas, donde cada uno pronuncia las 26 letras del abecedario inglés 7 veces.

- **LRW** [6]: contiene videos y audios de más de 1000 personas pronunciando 500 palabras diferentes.

- **LRS2** [7]: contiene miles de videos de programas de la BBC, donde hay personas hablando en diversas situaciones.

- **LRS3** [8]: contiene más de 1000 videos de TED y TEDx de personas hablando, con subtítulos y límites de alineación de palabras.

3.3 Modelos

Actualmente, para desarrollar sistemas de lectura de labios, se utilizan modelos capaces de procesar secuencias de video y extraer patrones de movimiento de los labios. Algunos modelos destacados son:

- **Convolutional Neural Network (CNN)**: está diseñada para extraer características visuales de imágenes. Es útil a la hora de identificar patrones del movimiento de los labios, así como la forma de los labios y la posición de la boca.

- **Recurrent Neural Network (RNN):** está diseñada para procesar secuencias de datos, ya que utiliza información de pasos anteriores para comprender el contexto. Es útil para utilizar el contexto temporal en el movimiento de los labios. Las RNNs son menos efectivas para secuencias largas debido al problema del *vanishing gradient*. Este problema ocurre cuando, al hacer *backpropagation* para ajustar los pesos del modelo, los gradientes (que indican qué tanto deben actualizarse los parámetros) se vuelven extremadamente pequeños a medida que se propagan hacia atrás a través de muchas capas.

- **Gated Recurrent Unit (GRU):** es una variante de las RNNs. Está diseñada para procesar secuencias mediante una *update gate* y una *reset gate*, las cuales permiten utilizar la información relevante de pasos anteriores. De la misma forma que las RNNs, tienen limitaciones con secuencias largas debido al problema del *vanishing gradient*.

- **Long Short-Term Memory (LSTM):** es otra variante de las RNNs, pero ésta está diseñada para superar el problema del *vanishing gradient*. Gracias a que consta de un *input gate*, una *forget gate* y un *output gate*, es capaz de conservar la información relevante en secuencias largas, lo que hace que sea útil para la lectura de labios en videos de larga duración.

- **Transformer**: este modelo utiliza un mecanismo de atención que permite analizar toda la secuencia de movimientos labiales en paralelo y enfocarse en las partes más relevantes. De esta forma, los Transformers son útiles para procesar secuencias largas y complejas de manera eficiente.

3.4 Análisis del movimiento labial

Para realizar un análisis efectivo del movimiento de los labios, existen técnicas complementarias que cumplen objetivos distintos y que pueden integrarse para mejorar el procesamiento de datos en sistemas de lectura de labios.

Dynamic Time Warping (DTW):
El *Dynamic Time Warping* (DTW) [9] es una técnica que se puede usar para complementar los sistemas de lectura de labios. Básicamente, sirve para comparar dos secuencias temporales, aunque estas tengan diferentes duraciones o velocidades. En el contexto de la lectura de labios, se puede utilizar para alinear secuencias de movimiento labial extraídas de videos.

Para conseguir el alineamiento, primero se calcula la distancia euclidiana entre cada par de elementos de las secuencias $A = \{a_1, a_2, \dots, a_n\}$ y B $= \{b_1, b_2, \dots, b_m\}$, definida como:

$$d(a_i, b_j) = \sqrt{\sum_{k=1}^{K} (a_{ik} - b_{jk})^2}$$

Donde K representa la cantidad de características en cada punto. Luego, se construye una matriz de costos acumulados $C(i,j)$, donde cada celda representa el coste mínimo para alinear las secuencias hasta la posición (i,j), siguiendo la ecuación de recurrencia:

$$C(i,j) = d(a_i, b_j) + \min\{C(i-1,j), C(i,j-1), C(i-1,j-1)\}$$

Donde $C(i-1,j)$ representa el coste de alinear el elemento actual con el anterior en la secuencia A, $C(i,j-1)$ representa el coste de alinear el elemento actual con el anterior en la secuencia B y $C(i-1,j-1)$ representa el coste de alinear ambos elementos simultáneamente.

Una vez construida la matriz de costos acumulados, se obtiene el camino de alineación óptimo retrocediendo desde la celda final $C(N,M)$, hasta la celda inicial $C(0,0)$, siguiendo el camino de menor coste en cada paso, es decir, eligiendo la celda con menor valor entre $C(i-1,j)$, $C(i,j-1)$ y $C(i-1,j-1)$.

El resultado es un alineamiento óptimo que permite comparar variaciones en la velocidad de pronunciación y la duración del movimiento labial de manera robusta. Esto es especialmente útil para la lectura de labios, ya que cada persona puede pronunciar las palabras con diferentes tiempos y estilos.

Descriptores de Fourier:
Los descriptores de Fourier [10] son otra técnica complementaria que puede utilizarse para procesar y analizar movimientos de los labios.

Esta herramienta transforma datos espaciales o temporales al dominio de Fourier, lo que permite representarlos en términos de sus componentes frecuenciales en lugar de valores espaciales o temporales. Una de sus principales ventajas es que son invariantes a escala, rotación y traslación, lo que los hace ideales para comparar patrones que pueden variar en tamaño, orientación o posición.

En el contexto de la lectura de labios, los descriptores de Fourier se pueden usar para analizar los movimientos de los labios de una forma más simple y clara. Al trabajar con frecuencias, se eliminan detalles pequeños o variaciones que no son importantes, permitiendo centrarse en las

características principales que realmente indican cómo se forman las palabras. Además, esta representación facilita la comparación entre palabras o frases, ya que ignora diferencias que no afectan al significado de la palabra, como el tamaño o la orientación de la boca.

Los coeficientes de Fourier representan la forma del contorno de los labios en términos de frecuencias espaciales. Cada coeficiente captura una parte de la estructura del movimiento labial: los primeros coeficientes describen la forma global, mientras que los coeficientes de mayor índice contienen información sobre detalles más finos. Para calcular estos coeficientes, primero se representa el contorno de los labios como una serie de puntos complejos en el plano:

$$z_n = x_n + iy_n\,, \qquad n = 0,1,\dots,N-1$$

Donde x_n y y_n son las coordenadas del punto n del contorno i es la unidad imaginaria. Luego, se aplica la Transformada Discreta de Fourier (DFT) para obtener los coeficientes:

$$F_k = \frac{1}{N}\sum_{n=0}^{N-1} z_n e^{-i\frac{2\pi kn}{N}}, \quad k = 0,1,\dots,N-1$$

Para la clasificación de palabras, generalmente se seleccionan los coeficientes de menor frecuencia, ya que son los que aportan información relevante sin verse afectados por pequeñas variaciones en la forma de los labios. El número de coeficientes a utilizar depende del número de puntos del contorno, pero en muchos casos, entre 10 y 20 coeficientes son suficientes para capturar la información esencial sin incluir ruido innecesario.

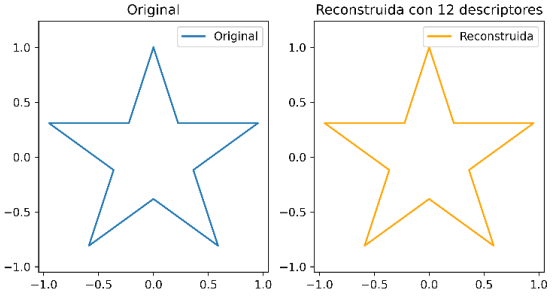

Figura 2: Comparación de la reconstrucción de una estrella con 12 descriptores de Fourier

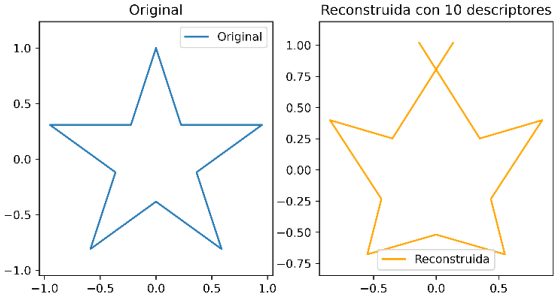

Figura 3: Comparación de la reconstrucción de una estrella con 10 descriptores de Fourier

En la Figura 2 se puede observar una comparación de una estrella reconstruida con 12 descriptores de Fourier, y en la Figura 3 una comparación con 10 descriptores de Fourier. Como se puede apreciar, con 12 descriptores es suficiente para reconstruir la figura a la perfección, en cambio, con 10 descriptores no. Sin embargo, se puede apreciar la estructura global de la estrella.

4 METODOLOGÍA

Para el desarrollo de este proyecto, se decidió utilizar la metodología *Scrumban* [11]. Esta metodología combina las mejores características de *Scrum* y *Kanban*.

Scrumban utiliza la estructura de *Scrum*, con el uso de *Sprints*, los cuales permiten dividir el trabajo en bloques de tiempo definidos. En cada *Sprint* se desarrollan una serie de tareas que son declaradas al principio, y al final de este, se hace una retrospección del trabajo realizado.

Por otro lado, se incorpora el flujo de trabajo visual y la flexibilidad de *Kanban*. *Kanban* aporta una visualización continua del estado el proyecto mediante un tablero. Gracias al uso de las columnas "*Backlog*", "*To do*", "*In progress*" y "*Done*", permite identificar de forma rápida el estado de cada tarea a realizar y detectar posibles cuellos de botella. *Kanban* también es conocido por su flexibilidad, ya que permite modificar el plan de tareas si es necesario, según la necesidad del equipo.

A continuación, se especifican las herramientas utilizadas para el desarrollo del proyecto:

- **Jira**: plataforma utilizada para la gestión del proyecto. Jira permite el uso eficiente de la metodología ágil *Scrumban*.
- **Git**: herramienta utilizada para el control de versiones. *Git* permite hacer un seguimiento de los cambios realizados a lo largo del desarrollo del proyecto.
- **GitHub**: herramienta complementaria a *Git*. *GitHub* permite almacenar el proyecto en un repositorio seguro en la nube, el cual es accesible desde cualquier dispositivo.
- **Visual Studio Code**: IDE utilizado para escribir y depurar el código realizado en *Python*, el cual es el lenguaje utilizado para el desarrollo del proyecto.
- **Windows Subsystem for Linux (WSL):** entorno de integración que permite utilizar recursos del sistema, como la GPU, de forma sencilla y eficiente, facilitando el entrenamiento de modelos
- **Microsoft Clipchamp**: herramienta de edición de video utilizada para el recorte manual de videos.

5 PLANIFICACIÓN

La planificación del proyecto se ha dividido en cinco fases principales, las cuales se han adaptado de manera progresiva para ajustarse a los cambios realizados en los objetivos durante su desarrollo.

En la primera fase, se llevaron a cabo las tareas esenciales para poder comenzar el proyecto, como, por ejemplo, una investigación y revisión preliminar del estado del arte sobre la lectura de labios, la definición inicial de los objetivos, la configuración del entorno de desarrollo y una serie de formaciones y pruebas prácticas iniciales.

La segunda fase se centró en el diseño y desarrollo de un pipeline de preprocesamiento de datos que fue aprovechado tanto en los objetivos iniciales del proyecto como en los finales. En esta fase se identificaron los *frames* útiles en los videos, se recortaron y estandarizaron los *frames* y se extrajeron características relevantes de los movimientos de los labios.

Tras el cambio de objetivos, la tercera fase se enfocó en desarrollar y evaluar métodos para encontrar correspondencias entre movimientos labiales. En esta fase se probaron diferentes enfoques para comparar movimientos labiales entre videos, y de esta forma, medir su similitud.

La cuarta fase se centró en diseñar e implementar un sistema basado en Machine Learning capaz de clasificar palabras en videos sin audio, utilizando los datos procesados y las similitudes obtenidas.

Finalmente, la quinta fase consistió en cerrar la fase de desarrollo del proyecto y comenzar a preparar la entrega final, incluyendo la redacción del informe y de la presentación final.

En el apartado A1 de apéndice, en la Figura 4, se pude observar el diagrama de Gantt del proyecto.

6 DESARROLLO

En este apartado se detalla el proceso de investigación e implementación del sistema de lectura de labios. El desarrollo comenzó con la experimentación preliminar de técnicas de transcripción de letras a partir de espectrogramas, lo que permitió familiarizarse con el procesamiento de datos y la clasificación basada en redes neuronales. A partir de esta base, se comenzó con la creación de un pipeline de preprocesamiento de videos para la extracción de características relevantes del movimiento labial. Seguidamente, se exploraron diferentes estrategias para medir la similitud entre movimientos labiales y, finalmente, se implementó un sistema de clasificación de palabras.

6.1 Letter Spectrogram Classifier

Para ir entrando en materia, lo primero que se desarrolló consistió en un clasificador binario de letras a partir de un espectrograma, usando, en este caso, las letras E y H.
Un espectrograma es una representación visual de la intensidad o amplitud de las frecuencias presentes en una señal (en este caso una señal de audio) a lo largo del tiempo. Se obtiene aplicando la Transformada de Fourier de Corto Tiempo (STFT) a fragmentos de la señal, lo que permite analizar la variación de las frecuencias en el tiempo y representarlas en una escala de colores.

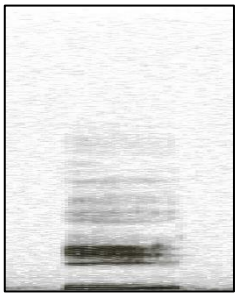

Figura 5: Espectrograma letra E

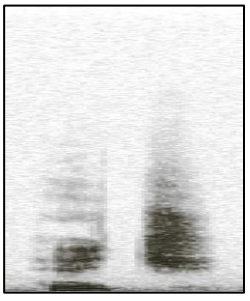

Figura 6: Espectrograma letra H

Para conseguir los espectrogramas se utilizó el *dataset* AVLetters2. Este *dataset* contiene videos de personas pronunciando letras, por lo que se transformó el audio de los videos en los que se pronuncian las letras E y H, en espectrogramas (Figura 5 y 6). Por cada letra se utilizaron 35 imágenes de 600x1000 pixeles, de las cuales se dividió un 80% para *train* y un 20% para *validation*. En este caso al tener pocas imágenes, se generaron audios propios pronunciando las letras para utilizarlo como test.

A partir de ahí, se configuró una arquitectura sencilla de una *Convolutional Neural Network* (CNN). Ésta contaba con dos capas convolucionales con función de activación *ReLU*, intercaladas con dos capas *Max Pooling*, permitiendo que la red aprenda a extraer patrones. Seguidamente una capa *Flatten* para convertir las características en un vector unidimensional, una capa *Dropout* con una tasa del 50% para evitar el *overfitting* y, por último, una capa *Dense* con activación *Softmax* para predecir la probabilidad que tiene la imagen de pertenecer a cada una de las 2 clases.

El modelo fue entrenado con *Categorical Crossentropy Loss Function*, optimizado con *Adam* y usando como métrica el *accuracy*. Se realizaron 50 épocas con un *batch size* de 8.

Una vez terminado el entrenamiento, se consiguió un *validation loss* del 0.00012697% y un *validation accuracy* del 100%.

Finalmente, al hacer *testing* del modelo con los audios generados, se pudo comprobar que el modelo fue capaz de clasificar bien todos los audios.

6.2 Preprocesamiento de datos

Para poder encontrar correspondencias entre movimientos labiales y posteriormente realizar una clasificación de manera correcta, primero se necesita que las imágenes utilizadas sean lo más estables posibles. Por este motivo, se necesita realizar un preprocesamiento de datos previo sobre los videos que se utilizarán.

Este preprocesamiento consiste en, primero, poder utilizar solo los *frames* útiles de los videos (*frames* donde la persona está diciendo la palabra). Y segundo, poder recortar los *frames* de un video por tal de solo mostrar la boca de la persona que esté hablando, además de que todos los *frames* recortados deben tener una medida estándar donde la boca no varie mucho su posición.

6.2.1 Búsqueda de frames útiles

Para poder utilizar el sistema de recorte de *frames* y, posteriormente, encontrar las similitudes entre videos, primero era necesario identificar los *frames* útiles, es decir, aquellos en los que la persona está pronunciando la palabra objetivo.

Inicialmente, se probaron dos sistemas automáticos para detectar estos *frames*:

- **Noise detector:**
  Se implementó un sistema basado en la detección de silencios mediante un umbral de decibelios. Aunque se normalizó la intensidad del sonido para mejorar la detección, este método resultó poco confiable, especialmente en casos donde las personas hablaban rápido o con variaciones en su pronunciación.

- **Whisper OpenAI:**
  Whisper se utilizó para transcribir el audio de los videos y extraer los tiempos de inicio y fin de las palabras detectadas. Aunque este modelo mostró una excelente precisión en la transcripción, los tiempos proporcionados no eran lo suficientemente exactos para delimitar con precisión los *frames* útiles. Se realizó un análisis donde comparaban los tiempos obtenidos por *Whisper* con tiempos anotados manualmente, y se pudo observar que *Whisper* tendía a cortar las palabras antes de que terminaran.

Finalmente, debido a las limitaciones de los dos sistemas automáticos probados, se optó por realizar el recorte manual de los videos. Gracias a la herramienta *Microsoft Clipchamp*, se pudo identificar correctamente los *frames* de inicio y fin de la pronunciación de la palabra de cada video.

6.2.2 Sistema de recorte de frames

La finalidad de recortar *frames*, de tal manera que solo se vea la boca de la persona que esté hablando, es simplificar el análisis en los movimientos de los labios, que es la parte más importante a la hora de encontrar similitudes entre *frames*. Esto es así ya que, de esta forma, se garantiza que el sistema solo procese la información necesaria, reduciendo ruido y datos irrelevantes.

Para conseguir esto, se ha desarrollado un sistema de recorte de *frames* automático. Este sistema funciona gracias a la librería *MediaPipe*, la cual es capaz de crear una máscara de puntos por toda la cara con información útil como la posición en pixeles de esos puntos. Por lo tanto, se ha utilizado los puntos ubicados en la zona interna de los labios para poder crear una *bounding box* la cual se utilizó para recortar la imagen.

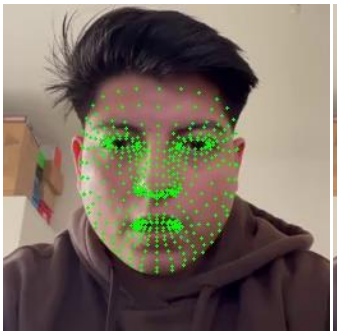

Figura 7: Máscara de puntos obtenida con MediaPipe

Figura 8: Máscara de puntos con solo landmarks internos de los labios

Debido a que el tamaño de la *bounding box* puede ser variable dependiendo lo cerca que esté la persona de la cámara, se realizó un proceso de normalización el cual consistió en aplicar un *resize* (a una dimensión estática definida) en las imágenes para que tengan todas las mismas dimensiones. Este proceso permite que las características extraídas sean compatibles entre diferentes videos.

Figura 9: Comparación frame sin recortar vs frame recortado y normalizado
Palabra: Blue

Observando los resultados de la Figura 9 se pudo ver que la deformación debido al *resize* a una dimensión estática es mínima, lo cual no afecta al desarrollo del sistema porque se utilizaron videos controlados. Pero en videos donde la persona puede que se esté moviendo o en diferentes videos con personas con diferentes cercanías a la cámara, podría generar una deformación más pronunciada. Por ello, en el futuro estaría bien implementar un sistema más elaborado donde el *resize* de la imagen se realice dependiendo de la distancia entre los ojos, de esta forma las dimensiones del *resize* pasarían a ser dinámicas.

6.3 Análisis y comparación de movimientos labiales

Para poder clasificar palabras basándonos en los movimientos labiales, es necesario encontrar una manera efectiva de medir la similitud entre diferentes videos. Esto requiere un método de comparación que sea capaz de capturar como varían los labios a lo largo de los *frames* del video, extrayendo información relevante que permita distinguir una palabra de otra, incluso cuando puedan compartir patrones de movimiento similares.

Uno de los principales problemas a la hora de comparar los movimientos labiales, es que las palabras pueden pronunciarse con duraciones distintas, dependiendo con la velocidad en la que se esté hablando, por lo que se necesita un método que permita comparar secuencias temporales de diferente longitud de manera eficiente.

6.3.1 Dynamic Time Warping (DTW)

Para poder solucionar este problema, se ha utilizado el *Dynamic Time Warping*, una técnica enfocada en la comparación de secuencias temporales. El DTW permite medir la similitud entre dos series de datos incluso si no tienen la misma longitud.

En este caso, el DTW es especialmente útil ya que es capaz de alinear las secuencias de características extraídas de los labios en diferentes videos, encontrando la correspondencia optima entre los *frames* y asegurando que las comparaciones se realicen sobre partes equivalentes del movimiento labial.

Para poder aplicar el DTW, primero se deben definir qué características de los labios se van a comparar entre los videos. Y para ello, se han probado diferentes enfoques para representar el movimiento de los labios.

6.3.1.1 Comparación basada en el área de los labios

En este enfoque, la característica utilizada para comparar los videos es el área formada por los puntos que rodean la parte interior de los labios en cada *frame*. La idea es que, como la apertura de la boca varía de manera significativa al pronunciar diferentes palabras, el área interna de los labios podría ser una buena característica que describa el movimiento.

Para cada *frame*, se calcula el área formada por los *landmarks* que rodean la parte interna de los labios. Luego, esta secuencia de valores es utilizada como entrada para el DTW, permitiendo alinear y comparar los patrones de cambio de área a lo largo de los *frames* del video.

6.3.1.2 Comparación basada en coordenadas 2D

Este enfoque consiste en utilizar directamente las coordenadas 2D de los puntos que delimitan la parte interior de los labios en cada *frame*. Con este enfoque, cada video se representa como una serie de puntos en el espacio, y la comparación entre videos se realiza comparando cómo cambian las posiciones de estos puntos a lo largo de los *frames* del video.

Al aplicar DTW sobre estas coordenadas, es posible medir la similitud entre las trayectorias de los puntos en diferentes videos. Esto permite capturar variaciones en la forma de los labios durante la pronunciación de una palabra, proporcionando una información más detallada que el área.

6.3.1.3 Comparación distancias al centro de masas

Este método proporciona una representación más estructurada del movimiento labial al describir la forma de los labios en función de distancias relativas en lugar de posiciones absolutas. Esto permite reducir la variabilidad causada por cambios en la forma de los labios, asegurando que la comparación entre secuencias se base en la estructura del movimiento y no en la posición exacta de los puntos.

Para ello, en cada *frame* se calcula el centro de masas del conjunto de puntos que delimitan la parte interior de los labios. Luego, en lugar de utilizar directamente las coordenadas 2D de los *landmarks*, se mide la distancia de cada punto a este centro. De esta manera, se obtiene una representación más estable que captura cómo varían estas distancias a lo largo de la pronunciación de la palabra.

6.3.1.4 Comparación basada en descriptores de Fourier

En este enfoque, la característica utilizada para comparar los videos es la transformación de los puntos que forman el contorno interior de los labios en el dominio de las frecuencias mediante descriptores de Fourier. Este método permite capturar la estructura global del movimiento sin verse afectado por traslaciones, rotaciones o cambios de escala.

Para cada *frame* del video, se extraen los *landmarks* del contorno interior de los labios y se representan como un conjunto de puntos con coordenadas 2D. Luego, estos puntos se convierten a números complejos, y se aplica la Transformada Rápida de Fourier (FFT) para obtener un conjunto de coeficientes que describen la forma de los labios en términos de frecuencias espaciales.

El uso de números complejos en la Transformada de Fourier permite descomponer una señal en términos de amplitud (magnitud) y fase (ángulo). La magnitud de los coeficientes de Fourier captura la contribución que realiza cada frecuencia a la hora de interpretar la estructura global de la zona interior de los labios, mientras que la fase captura información sobre la posición relativa de los puntos.

Sin embargo, para realizar la comparación mediante el DTW, al utilizar distancias euclidianas entre puntos, la librería utilizada no permite utilizar los descriptores de Fourier representados en números complejos. Por lo que, en este caso, se ha decidido utilizar únicamente la magnitud de los coeficientes de Fourier, eliminando la dependencia de la fase debido a que se han utilizado videos controlados y normalizados, por lo que la información relativa a la posición de los puntos es irrelevante. Además, se ha utilizado el máximo número de coeficientes, en este caso, 19 coeficientes, debido a la mala recreación con un numero inferior

de coeficientes. En la Figura 10 se pude observar una comparación entre la forma original de la parte interior de los labios y su recreación de con 19 descriptores de Fourier.

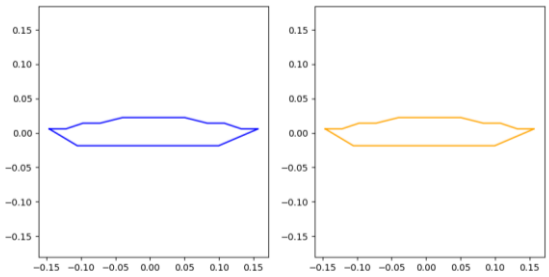

Figura 10: Comparación de la reconstrucción de la parte interior de la boca con 19 descriptores de Fourier

6.3.2 Comparación y decisión del mejor enfoque

Para determinar qué enfoque es más adecuado para medir la similitud entre movimientos labiales utilizando el DTW, se utilizaron dos videos en los que se pronuncia la palabra "HELLO", donde el primer video corresponde a una pronunciación normal y el segundo a una pronunciación más lenta, alargando la primera sílaba de esta forma: "HEEELLO". El objetivo era medir qué enfoque lograba capturar de manera más precisa la relación entre ambas secuencias, minimizando el coste de similitud y proporcionando una representación estable del movimiento labial.

En los apartados A2, A3, A4 y A5 del apéndice se pueden observar representaciones de las alineaciones temporales generadas por DTW para los distintos enfoques. Las filas intermedias muestran los *frames* originales de cada serie (la inferior corresponde a la primera serie y la superior a la segunda), mientras que las filas superior e inferior representan los *frames* alineados combinados, permitiendo visualizar la posición de los *landmarks* en cada *frame*. Las líneas azules indican la correspondencia entre *frames*: si un *frame* de la serie 1 se alinea con varios de la serie 2, significa que la segunda serie ha estirado ese segmento en el tiempo, y viceversa.

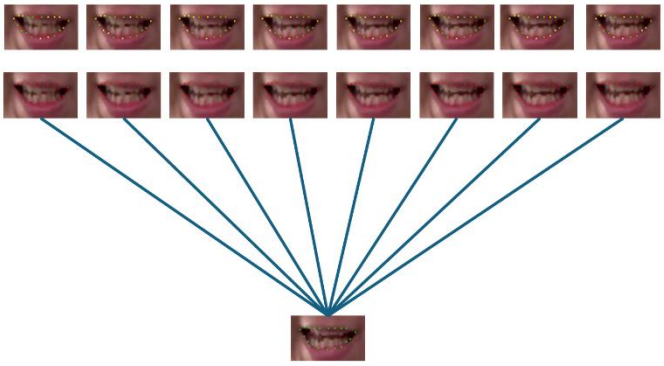

Figura 11: 1a parte del alineamiento del DTW con área

El primer enfoque evaluado fue el DTW con el área de los labios. El coste de similitud obtenido fue 11.83, lo que indica una gran deformación para alinear ambas secuencias. A pesar de este alto coste, como se observa en la Figura 11, el alineamiento refleja claramente que el inicio del segundo

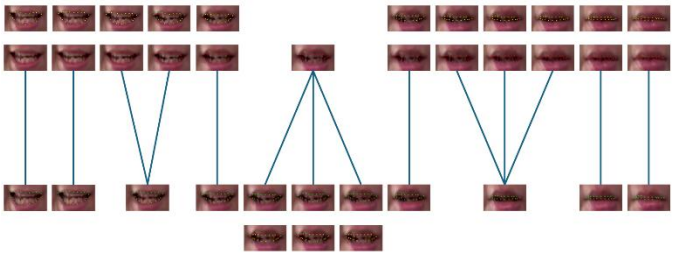

Figura 12: 1a parte del alineamiento del DTW con área

video está estirado respecto al primero, lo cual era esperable debido al alargamiento de la primera sílaba. Sin embargo, como se puede ver en la Figura 12, los *frames* posteriores muestran un alineamiento más estable. Este análisis permitió concluir que el área de los labios varía significativamente con la velocidad de pronunciación, lo que lo convierte en un método poco robusto ante cambios en la forma de hablar.

El segundo enfoque analizado fue el DTW con coordenadas 2D. Este método mejoró notablemente la alineación entre secuencias, reduciendo el coste de similitud a 2.55, lo que indica una mayor precisión en la correspondencia entre los movimientos labiales. A nivel de alineamiento, como se puede observar en el apartado A3 del apéndice, el inicio del segundo video está estirado respecto al primero, lo que, al igual que el anterior enfoque, concuerda con la pronunciación más alargada de la primera sílaba. En comparación con el enfoque anterior, la correspondencia entre *frames* es más estable a lo largo de toda la secuencia, lo que sugiere que utilizar directamente las coordenadas 2D permite capturar con mayor precisión la evolución del movimiento labial sin introducir tanta deformación. Sin embargo, este resultado se debe a que las imágenes estaban normalizadas, lo que reduce la variabilidad en la posición de la boca en términos de píxeles. Por lo que, en situaciones donde no se aplique normalización, donde la morfología de la boca varíe entre personas o incluso donde la librería que obtiene los *landmarks* produzca diferentes resultados, este método puede ser menos fiable debido a la variabilidad en las posiciones de los *landmarks*, lo que afectaría la precisión de la alineación.

El tercer enfoque analizado fue DTW con distancias al centro de masas. Este método obtuvo un coste de similitud de 4.16, ligeramente superior al de las coordenadas 2D, pero con la ventaja de ser más robusto ante variaciones en la morfología de los labios entre diferentes personas. En cuanto al alineamiento, como se puede observar en el apartado A4 del apéndice, este enfoque mostró los peores resultados hasta el momento. Se puede observar que hay una mayor cantidad de *frames* mal emparejados. Específicamente, la primera parte del segundo video presenta alineaciones con varios *frames* de la primera serie, lo que indica una menor precisión al capturar el alargamiento de la primera sílaba. Además, a lo largo de la secuencia se generan alineaciones que no son estables. A pesar de esto, el enfoque sigue siendo una alternativa útil cuando se trabaja con múltiples personas, ya que evita depender de la posición

absoluta de los labios. Sin embargo, en este caso específico, donde la normalización ya reduce las variaciones en la ubicación de la boca, la pérdida de precisión en el alineamiento sugiere que este método no es el más adecuado para capturar diferencias en la velocidad de pronunciación.

El último enfoque analizado fue el DTW con descriptores de Fourier. Este método proporcionó el mejor resultado, con un coste de similitud de 0.085, lo que indica una alineación mucho más precisa y estable. En el apartado A5 del apéndice se puede observar que, en términos de alineamiento, aunque no es más preciso que el alineamiento obtenido usando las coordenadas 2D, su representación en coeficientes permite capturar mejor características del movimiento labial. Se puede observar que la alineación sigue el patrón general del alargamiento en la primera parte de la palabra, pero con una alineación posterior menos estable. Sin embargo, los descriptores de Fourier logran modelar de manera más robusta la forma de los labios a lo largo del tiempo, reduciendo la influencia de variaciones de posición en los *landmarks*. Además, su capacidad para representar la forma de los labios de manera invariante a traslaciones y escalas lo hace altamente efectivo para comparar diferentes personas sin perder información relevante del movimiento. Esto lo convierte en el enfoque más adecuado ya que se busca una solución generalizable para la lectura de labios en distintas personas.

Una vez analizado los resultados obtenidos con cada enfoque, se llegó a la conclusión que el DTW con descriptores de Fourier es el mejor enfoque para medir similitud entre movimientos labiales, ya que minimiza el coste de alineación y ofrece una representación invariante a cambios en la forma y posición de los labios. Además, su capacidad de eliminar detalles irrelevantes y capturar únicamente las características más importantes del movimiento labial lo hace perfecto para la clasificación de palabras.

Por estas razones, se ha seleccionado los descriptores de Fourier como el enfoque principal para encontrar correspondencias entre movimientos labiales.

6.4 Clasificación de palabras

Para poder clasificar palabras, se decidió utilizar un clasificador *K-Nearest Neighbors* (KNN) junto con los costes de similitudes obtenidos mediante *Dynamic Time Warping* (DTW) con descriptores de Fourier. Para evaluar el rendimiento del clasificador se realizaron tres *tests* con conjuntos de datos progresivamente más complejos.

6.4.1 Primer test: Dataset controlado con una única persona

En este primer *test*, se utilizó un conjunto de datos compuesto por 8 videos grabados en un entorno controlado. Cada video contenía la pronunciación de una única palabra, con un total de 4 palabras distintas: “AFTERNOON”, “ALWAYS”, “CAT”, “UNDERSTAND”.

Para cada palabra, existían dos videos, donde la diferencia era en que el primer video contenía una pronunciación estándar, mientras que el segundo video tenía una sílaba extendida. Estos videos fueron grabados bajo condiciones controladas para garantizar que la boca estuviera los más centrada posible en el video y orientada paralelamente a la cámara. Además, en todos los videos aparecía siempre la misma persona. Dado que solo había dos videos por palabra, se optó por un KNN con k=1, ya que cada palabra tenía solo una muestra de referencia en el conjunto de entrenamiento.

Se probó el KNN con cada palabra, y se consiguió un 100% de *accuracy*. Todas las palabras fueron clasificadas correctamente, lo que indica que el método es eficaz en un entorno controlado con un solo hablante.

6.4.2 Segundo test: Dataset controlado con múltiples personas

En este segundo *test*, el conjunto de datos se amplió con 24 nuevos videos, aumentando el total a 32. Se añadieron 6 videos por palabra, incluyendo videos de dos nuevas personas, además de la persona original. Con esta expansión, el *dataset* ahora incluía tres personas diferentes, lo que le añadía variabilidad.

Como ahora existían más muestras por palabra, se incrementó el número de vecinos en el KNN a k=5, buscando mejorar la generalización en la clasificación. Sin embargo, los resultados mostraron una disminución en el *accuracy* con respecto al primer *test*. Las palabras pronunciadas por la persona original se identificaron correctamente. Sin embargo, las palabras pronunciadas por las nuevas personas fueron clasificadas de manera errónea. En general, consiguió un *accuracy* del 50% en cada palabra, con los errores concentrados en los videos de personas distintas a la original.

Este comportamiento demuestra que el modelo no está aprendiendo representaciones generalizables del movimiento labial, sino que se basa en características específicas de cada persona.

6.4.3 Tercer test: Dataset con variaciones en la orientación de la cabeza

En este tercer *test*, se añadieron 6 videos nuevos por palabra, donde las personas pronunciaban las palabras con una ligera inclinación de la cabeza respecto a la cámara. Esto generó un total de 56 videos en el *dataset*.

A pesar de existir más muestras en el dataset, se mantuvo el KNN con k=5. Los resultados fueron similares a los de la segunda prueba. Se pudo observar que los videos con la cabeza girada solo podían clasificarse correctamente con otros videos de la misma persona en una postura similar. Cuando se intentaba clasificar estos videos con otros donde la persona tenía la cabeza en posición frontal, es decir con una perspectiva de los labios diferente, la clasificación fallaba.

Este resultado demuestra que el modelo actual no es capaz de generalizar correctamente frente a cambios en la orientación de la cabeza. La razón principal de este problema es que los descriptores de Fourier están capturando formas distintas de la parte interior de los labios, por lo que, el DTW no consigue un coste bajo de similitud entre videos.

7 Conclusión

Este trabajo se ha centrado en el desarrollo de un sistema de lectura de labios basado en Inteligencia Artificial y Visión por Computador, utilizando descriptores de Fourier para representar el movimiento labial, *Dynamic Time Warping* (DTW) para medir similitudes entre secuencias temporales y *K-Nearest Neighbors* (KNN) para la clasificación de palabras pronunciadas en videos sin audio.

Los resultados obtenidos demuestran que el método propuesto es efectivo en entornos controlados, alcanzando un 100% de precisión cuando el *dataset* está compuesto por videos de una sola persona bajo condiciones uniformes. Sin embargo, al introducir mayor variabilidad, como diferentes personas o cambios en la orientación de la cabeza y en la perspectiva de los labios, el rendimiento del sistema se redujo significativamente, mostrando dificultades para generalizar correctamente.

La combinación de descriptores de Fourier y DTW ha demostrado ser útil para medir similitudes entre secuencias de movimiento labial, pero los resultados indican que este enfoque es sensible a la variabilidad entre personas y cambios en la orientación de la cabeza. Esto sugiere que la representación de los movimientos labiales aún requiere mejoras para lograr un sistema más robusto y adaptable a diferentes condiciones.

Para mejorar la generalización del sistema y hacerlo más robusto a diferentes condiciones, se proponen las siguientes mejoras:

- Implementar un proceso de alineación automática basado en la distancia entre los ojos o la posición relativa de la boca en la cara. Esto ayudaría a compensar giros de la cabeza.

- Aplicar homografías para corregir la inclinación de la cabeza y normalizar la perspectiva de los labios en los videos. Esto permitiría transformar la región de interés a una vista más frontal, reduciendo el impacto de variaciones en la orientación de la cabeza sobre los descriptores utilizados en la clasificación.

- Ampliación del *dataset* para mejorar la generalización. La mala clasificación observada en los experimentos podría reducirse si usa un conjunto de datos más amplio y variado.

- El uso de Redes Neuronales Convolucionales (CNNs) en combinación con LSTMs o *Transformers* podría ayudar a capturar patrones de movimiento invariables a giros de la cabeza, logrando una mejor generalización.

Agradecimientos

Quiero agradecer a mi tutor, Coen Antens, por su guía y apoyo a lo largo de todo el proyecto. Su orientación fue clave para el desarrollo de este trabajo. También agradezco a Raul Ochoa y Oriol Cano por su colaboración en la grabación de los videos del *dataset*, permitiendo ampliar las pruebas del sistema. Por último, a mi familia, por su apoyo constante y por estar siempre ahí en cada etapa de este proceso.

Bibliografia

[1] "Lip-reading software helps users of all abilities to send secure messages," University of California, Oct. 17, 2023. [Online]. Available: https://www.universityofcalifornia.edu/news/lip-reading-software-helps-users-all-abilities-send-secure-messages.

[2] K. Vayadande, T. Adsare, N. Agrawal, T. Dharmik, A. Patil, and S. Zod, "LipreadNet: A deep learning approach to lip reading," in 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), 2023, pp. 1-6.

[3] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[4] Pandey, L., & Arif, A. S. (2021, May). Liptype: A silent speech recognizer augmented with an independent repair model. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-19).

[5] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

[6] Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 2024, 2016, Revised Selected Papers, Part II 13 (pp. 87-103). Springer International Publishing.

[7] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8717-8727.

[8] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.

[9] K. Bringmann, N. Fischer, I. van der Hoog, E. Kipouridis, T. Kociumaka, and E. Rotenberg, "Dynamic Dynamic Time Warping," arXiv preprint arXiv:2310.18128, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.18128

[10] C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," IEEE Transactions on Computers, vol. C-21, no. 3, pp. 269-281, Mar. 1972.

[11] D. Atlassian, "Scrumban: domina dos metodologías ágiles," Atlassian. [Online]. Available: https://www.atlassian.com/es/agile/project-management/scrumban.

APÉNDICE

A1. DIAGRAMA DE GANTT

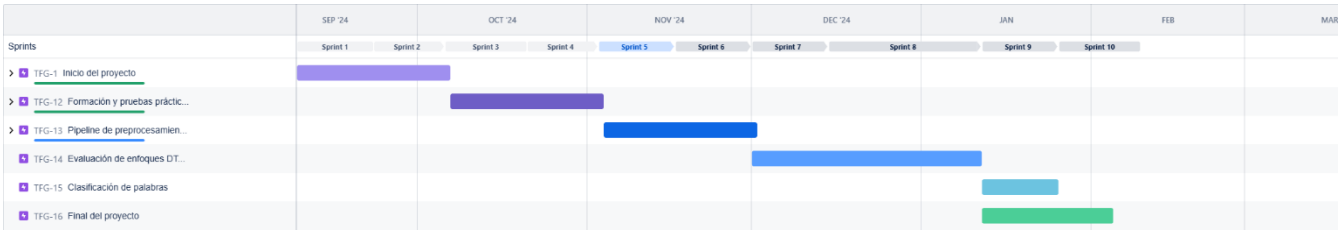

Figura 4: Diagrama de Gantt del proyecto

A2. Alineamiento DTW con área

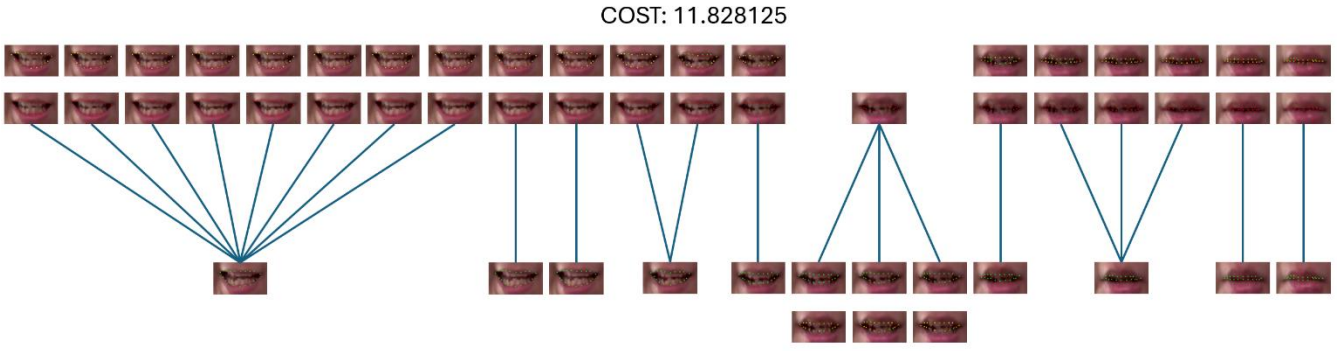

Figura 13: alineamiento completo del DTW con área

A3. Alineamiento DTW con coordenadas 2D

Figura 14: alineamiento completo del DTW con coordenadas 2D

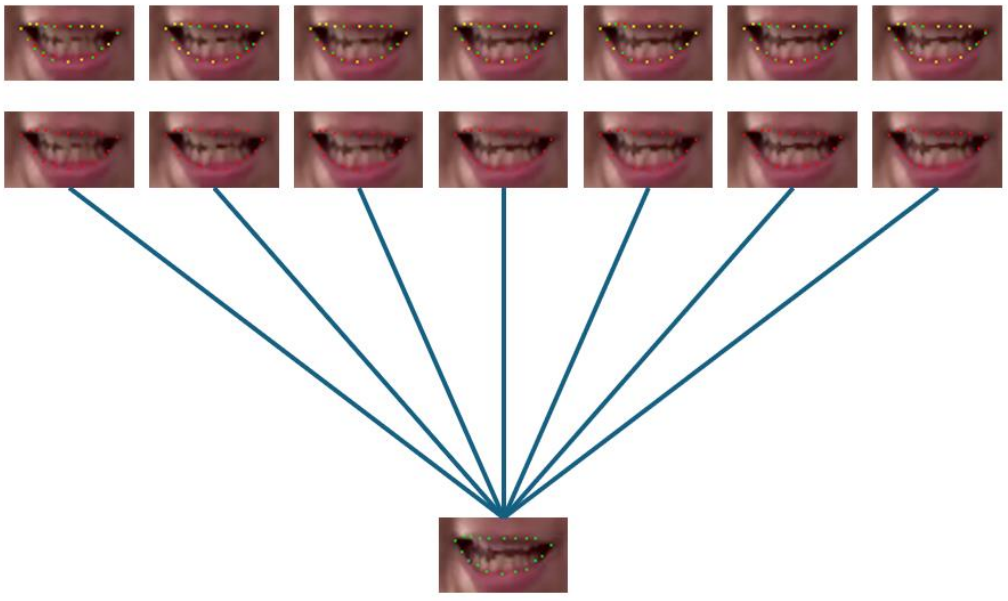

Figura 15: 1a parte del alineamiento del DTW con coordenadas 2D

Figura 16: 2a parte del alineamiento del DTW con coordenadas 2D

A4. ALINEAMIENTO DTW CON DISTANCIAS AL CENTRO DE MASAS

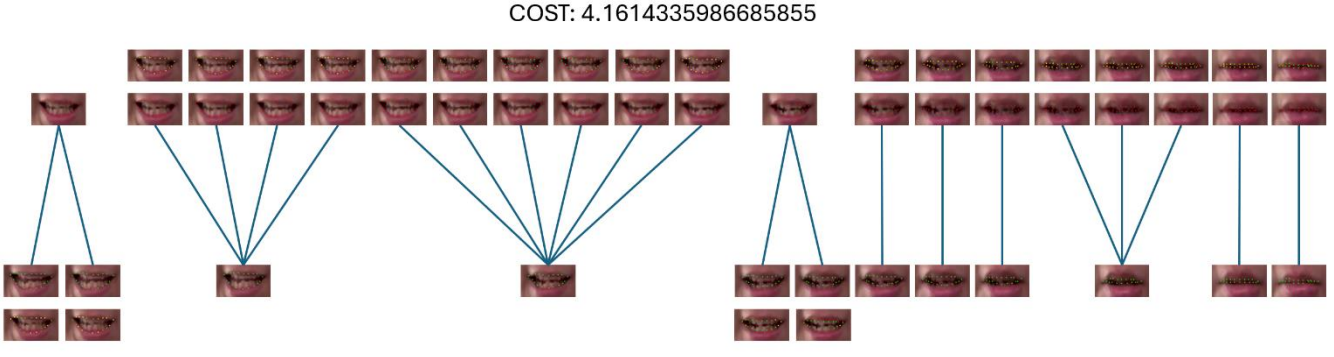

Figura 17: alineamiento completo del DTW con distancias al centro de masas

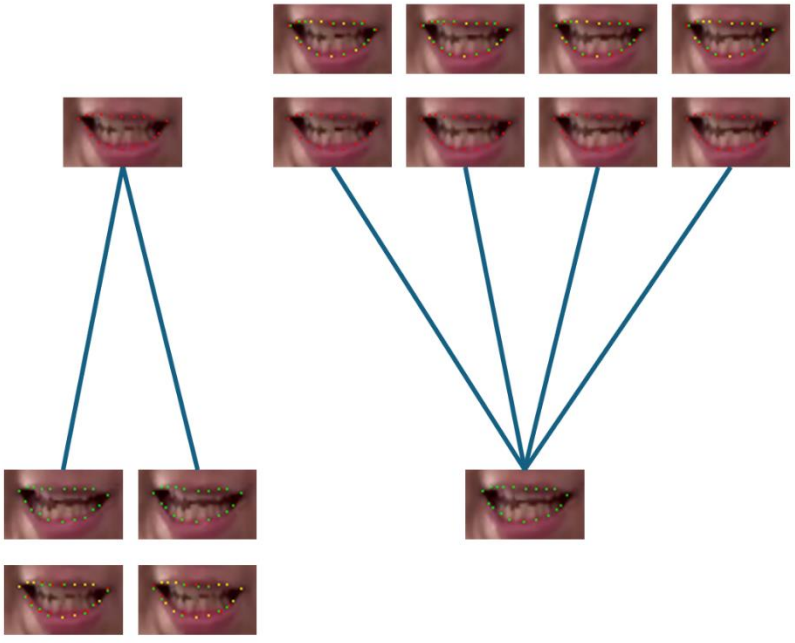

Figura 18: 1a parte del alineamiento del DTW con distancias al centro de masas

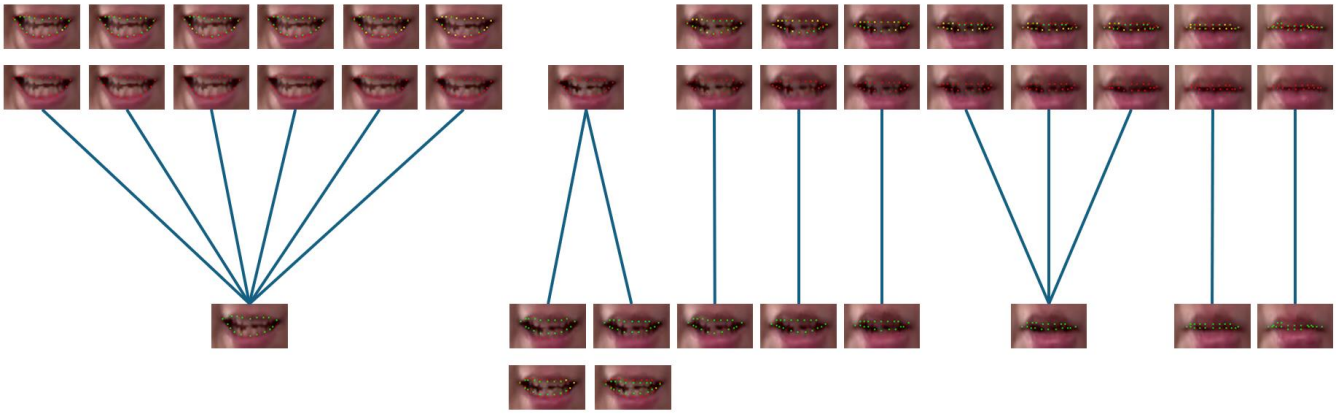

Figura 19: 2a parte del alineamiento del DTW con distancias al centro de masas

A5. Alineamiento DTW con descriptores de Fourier

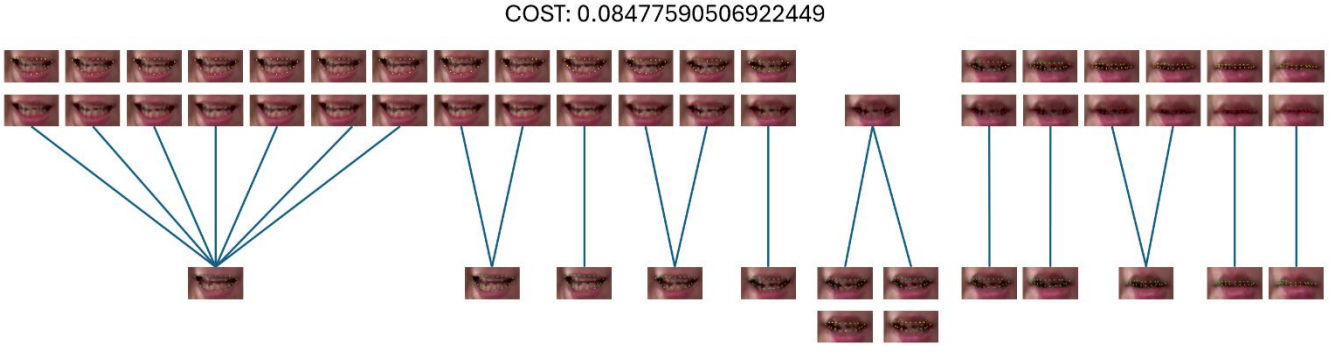

Figura 20: alineamiento completo del DTW con descriptores de Fourier

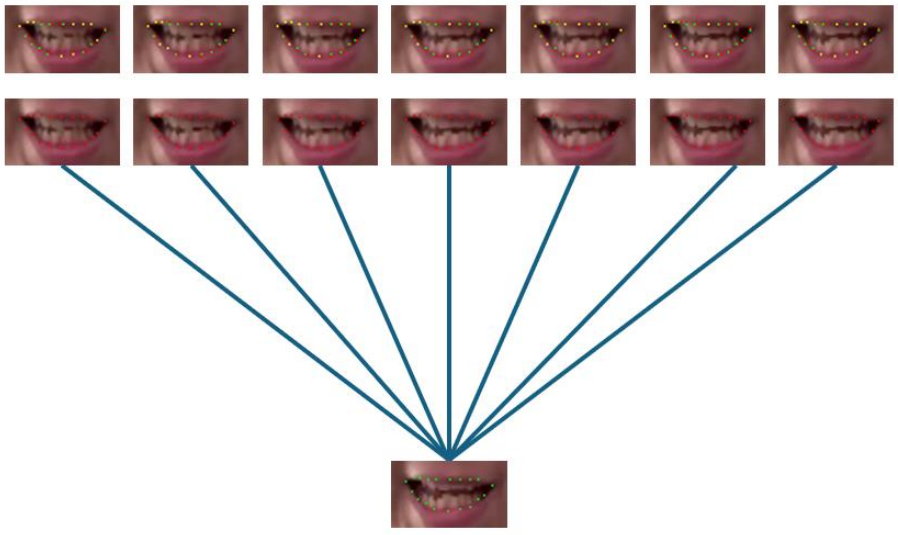

Figura 21: 1a parte del alineamiento del DTW con descriptores de Fourier

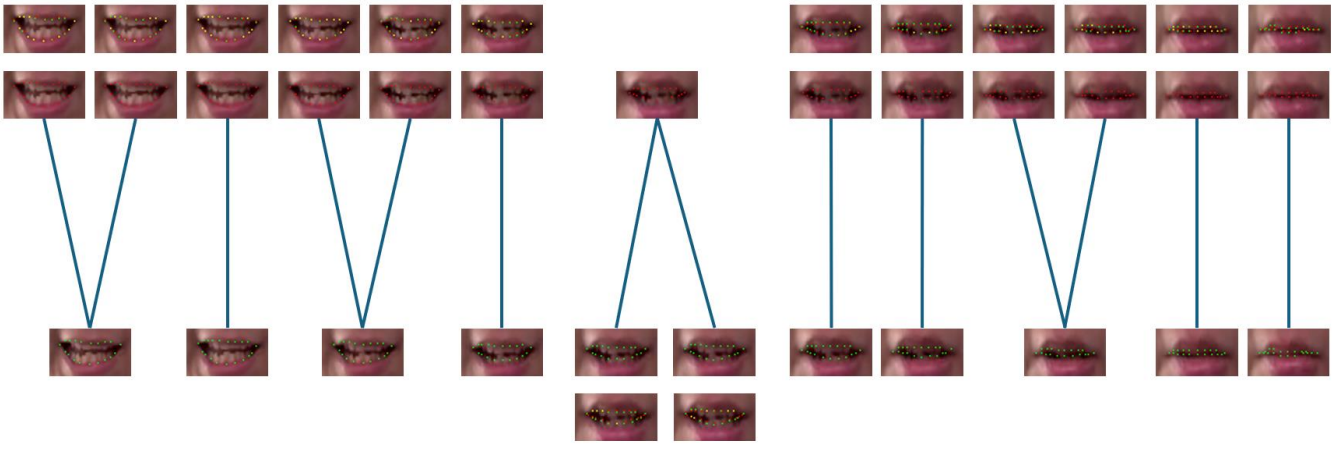

Figura 22: 2a parte del alineamiento del DTW con descriptores de Fourier