

Trabajo Final Inteligencia Artificial I - Año 2023

Visión Artificial y Reconocimiento de Voz

Ingeniería en Mecatrónica

Alumno: Juan Manuel BORQUEZ PEREZ

Legajo: 13567



UNCUYO
UNIVERSIDAD
NACIONAL DE CUYO



FACULTAD
DE INGENIERÍA

► 1983/2023
40 AÑOS DE DEMOCRACIA

Contents

1 Resumen	2
2 Introducción	2
2.1 Visión Artificial	2
2.2 Reconocimiento de Voz	3
3 Especificación del Agente	4
3.1 Descripción y tipo de Agente	4
3.2 Tabla REAS	6
3.3 Descripción del Entorno	6
4 Diseño del Agente	7
4.1 Reconocimiento de voz	7
4.1.1 Recorte de Audios	7
4.1.2 Extracción de Características	9
4.1.3 Reducción de componentes	16
4.2 Reconocimiento de Imágenes	18
4.2.1 Separación del fondo	18
4.2.2 Extracción de Características	20
4.2.3 Clasificación con K-means	21
4.2.4 Clasificación	22
5 Ejemplo de Aplicación	22
5.1 Dataset	22
5.2 Implementación	22
5.2.1 Entrenamiento	23
5.2.2 Validación	23
5.2.3 Solución	24
5.2.4 Ejemplo	24
6 Resultados	27
6.1 Reconocimiento de voz	27
6.2 Reconocimiento de Imagen	28
7 Conclusiones	28

1 Resumen

En este informe se presenta el desarrollo de una solución al problema propuesto por la cátedra. Se tiene una máquina expendedora de 4 tipos de fruta: manzana, naranja, banana y pera. La máquina cuenta con una cámara para tomar fotos de las frutas en los estantes y un micrófono para solicitar frutas por voz. El software de la máquina identifica las frutas en las imágenes y sus nombres cuando son mencionadas por el usuario. La clasificación de la voz se realiza mediante un algoritmo **KNN** con $k=3$, y la clasificación de las imágenes se lleva a cabo con un algoritmo KNN con $k=1$ comparando cada imagen con los centroides obtenidos del entrenamiento de un segmentador basado en **KMeans**. Se construyó un dataset disponible en línea con audios de varias personas e imágenes recopiladas en línea o tomadas por alumnos. Los resultados obtenidos fueron suficientemente satisfactorios; en concreto, la validación del modelo de reconocimiento de voz se realizó con 24 archivos de distintas personas sin falla. El reconocimiento de frutas en imágenes, aunque no completamente probado, es vulnerable ante frutas descoloridas, siendo el color la característica principal para la separación. El desarrollo se encuentra principalmente documentado en notebooks de Jupyter.

This report presents the development of a solution to the problem proposed by the department. There is a vending machine with 4 types of fruit: apple, orange, banana, and pear. The machine is equipped with a camera to take photos of the fruits on the shelves and a microphone to request fruits by voice. The machine's software identifies the fruits in the images and their names when mentioned by the user. Voice classification is done using a KNN algorithm with $k=3$, and image classification is performed with a KNN algorithm with $k=1$ comparing each image with centroids obtained from training a KMeans-based segmenter. A dataset, available online, was built with audio from various people and images collected online or taken by students. The results obtained were sufficiently satisfactory; specifically, voice recognition model validation was performed with 24 files from different people without failure. Fruit recognition in images, although not fully tested, is vulnerable to discolored fruits, with color being the main feature for separation. The development is primarily documented in Jupyter notebooks.

2 Introducción

2.1 Visión Artificial

La visión artificial es la tecnología que le permite a los equipos industriales percibir las características del entorno a través de imágenes de forma automática. A diferencia de un simple procesamiento de imágenes, en el que el resultado de una imagen de entrada es otra imagen de salida modificada, la visión artificial implica la extracción de características relevantes de las imágenes que permitan identificar los elementos de interés que contienen. Las imágenes se pueden obtener con distintos tipos de sensores y es así que se tienen imágenes como las que se pueden obtener con una cámara tradicional sensible a la radiación en el rango del espectro visible o imágenes termográficas obtenidas con sensores sensibles a la radiación infrarroja del espectro por dar ejemplos.

La visión artificial clásica es un campo que se comenzó a desarrollar mucho antes del

desarrollo de las aplicaciones más avanzadas como el Machine Learning y sin embargo, a través de simples operaciones con características de las imágenes permitió identificar diferentes entidades en principio bien definidas como códigos de barras, bordes, objetos, colores, etc.

Las aplicaciones de la visión artificial son variadas e incluyen la detección de defectos en partes de máquinas, medición de partes, identificación y rastreo de objetos, identificación de textos, etc. Los principales elementos involucrados en la obtención de imágenes para la visión artificial son: una fuente de luz, un escenario específico y controlado para capturar una toma, aumentos y un sensor para capturar la imagen, en general una cámara de algún tipo.

En este trabajo se implementa la visión artificial en el sentido clásico para extraer características de imágenes de 4 tipos de frutas: peras, bananas, manzanas y naranjas con el objeto de hacer una segmentación del conjunto de imágenes en grupos según el tipo de fruta.

Inicialmente se planteó la solución al problema tratando de que sea lo suficientemente robusta como para poder identificar las frutas en cualquier tipo de fondo, en ese sentido se exploraron diversas características, máscaras y estrategias. Sin embargo, el problema de lograr la robustez no se pudo resolver de forma satisfactoria en todos los casos y por falta de tiempo se decidió tomar mayor control del escenario optándose finalmente por el uso de fondo blanco en todos los casos.

Para entrenar el segmentador fue necesario disponer de un dataset de imágenes de entrenamiento. De este dataset, algunas imágenes se recopilaron de páginas en internet mientras que la mayoría se obtuvieron tomando fotos a frutas con la cámara de un celular. En las imágenes capturadas no se tuvo demasiado recaudo en cuanto a la escena más que la utilización de luz natural y el posicionamiento de la fruta en algún fondo blanco.

Se exploraron diversas características de las imágenes como los bordes, texturas, color, etc., que fueron relevantes tanto para la separación de las frutas del fondo como para lograr la posterior segmentación del conjunto de imágenes en grupos de frutas.

2.2 Reconocimiento de Voz

El reconocimiento de voz es la capacidad de un sistema de software para transformar el discurso de una persona en su representación en texto, permitiendo la comunicación entre un humano y una computadora a través del habla. Este tipo de sistemas integran diferentes tipos de información contenida en la señal de audio, como la gramática, la sintaxis, la estructura y la composición del audio, incluso en presencia de ambigüedades, incertidumbres y perturbaciones como el ruido, con el objetivo de obtener una interpretación aceptable del mensaje que se desea transmitir. Estos sistemas se utilizan en aplicaciones como el dictado automático, el control por comandos de voz, traductores, reconocimiento de canciones, entre otras.

Este tipo de sistemas pueden utilizar aprendizaje deductivo o sistemas expertos, que son entrenados con los conocimientos de un conjunto de campos involucrados en el habla, tales como la lingüística, la fonética, la acústica, etc. También pueden ser sistemas que hagan uso de aprendizaje inductivo, en el cual el sistema tiene la capacidad de adquirir los conocimientos necesarios de manera automática. Dentro de esta última categoría se encuentran la mayoría de las técnicas utilizadas: Hidden Markov Models, N-Grams y

Redes Neuronales.

En el trabajo que se presenta aquí, el reconocimiento del discurso se limita a la identificación de los nombres de las frutas mencionadas. Tanto si se trata de una solución con aprendizaje automático como la solución que se presenta en este caso, en la cual no se utiliza tal técnica, es necesario llevar a cabo la extracción de las características que representan la información relevante contenida en la señal. La parte más complicada de esta solución radica precisamente en el procesamiento de las señales de audio para lograr pasar por alto perturbaciones como el silencio o el ruido, y la posterior extracción de características que permitan diferenciar audios de distintas frutas. Después de extraer un conjunto de características que permitan separar adecuadamente el conjunto de audios, la clasificación de un nuevo dato a través del algoritmo k-NN es algo trivial.

3 Especificación del Agente

3.1 Descripción y tipo de Agente

El agente se ha interpretado de la siguiente manera. El mismo consiste en una máquina expendedora de frutas. La máquina dispone de 4 estanterías, en cada una de las cuales se encuentra uno de los tipos de fruta considerados. Cuando un usuario desea obtener una fruta de la máquina expendedora, presiona un botón para hablar en el micrófono de la máquina y decir el nombre de la fruta deseada el que el agente puede identificar a través de su programa. Luego, el agente determina si la fruta se encuentra en alguna de las estanterías y, si es así, identifica en cuál de todos. Entonces, a través de un actuador empuja la fruta del estante para expenderla al usuario. Para la determinación de la existencia y ubicación de la fruta solicitada, previo al pedido por voz del usuario, el agente toma imágenes con su cámara de las frutas en los estantes y las clasifica.

Se considera que se trata de un **agente que aprende** debido a que los algoritmos que utiliza para la clasificación de las frutas en imágenes y por voz están comprendidos dentro de ese tipo de agentes ([**RusselNorvig2009**]). El aprendizaje como tal se evidencia sobre todo en el algoritmo K-means, ya que durante el entrenamiento, el agente se vuelve capaz de encontrar similitudes y diferencias entre los grupos de imágenes. Por otro lado, en la clasificación de frutas por voz, no existe una etapa de entrenamiento como tal, y el agente requiere toda la base de datos de audio para hacer una predicción en base a una nueva orden (aprendizaje basado en memoria [6]). En ambos casos, se puede decir que el agente tiene la capacidad de mejorar su habilidad para clasificar imágenes y audio mediante la incorporación de más datos a la base de datos de imágenes y audios utilizados para el entrenamiento, otra razón por la cual se considera como un agente que aprende. En esta implementación, sin embargo, no se contempla la posibilidad de que audios de nuevas órdenes o las nuevas imágenes tomadas de las frutas en la estantería sean incorporadas a la base de entrenamiento para reentrenar al segmentador K-means o para ampliar los datos del clasificador k-NN para audio. Terminada la validación del clasificador de audios y entrenado el segmentador de imágenes, el comportamiento del agente es como el de un **agente basado en modelos** dado que son los modelos entrenados que permiten identificar los nombres y los objetos. En suma, existe durante la propia implementación del agente un proceso de ajuste no automático sino asistido por el diseñador del sistema y en base a una comparación entre el rendimiento que tiene el agente en un instante

determinado y el rendimiento esperado por el cual se considera que el agente también aprende.

3.2 Tabla REAS

Rendimiento	Entorno	Actuadores	Sensores
<ul style="list-style-type: none"> Exactitud en el reconocimiento de las frutas medida por el número de aciertos respecto del total de ordenes del usuario. Rapidez en la respuesta del agente medida como el tiempo entre que el usuario lleva a cabo una orden y recibe la fruta requerida. Tratamiento cuidadoso de las frutas. 	<ul style="list-style-type: none"> El gabinete de la máquina con los estantes, el estado de los mismos y la iluminación. El entorno en donde la máquina se ubica, su ruido ambiental y la iluminación. Los usuarios de la máquina y las propias frutas. 	<ul style="list-style-type: none"> Elementos de manipulación de la cámara para desplazarla y tomar fotos en los estantes. Elementos para manipulación de las frutas, para colocarlas en los estantes y dispensarlas. Indicadores para decir al usuario si la fruta no se encuentra. 	<ul style="list-style-type: none"> Micrófono para recibir la orden. Cámara para capturar imágenes. Botón que presiona el usuario para hacer la orden.

Table 1: Tabla REAS

3.3 Descripción del Entorno

- Completamente observable:** Si la escena es controlada, es decir, el nivel de iluminación dentro de la cabina es suficiente, el color del fondo es el adecuado, la cámara funciona correctamente, el micrófono funciona correctamente y el ambiente no tiene demasiado nivel de ruido - como se supone - entonces los sensores permiten acceso a toda la información relevante del entorno para la toma de una decisión por parte del agente.
- Multi Agente:** Se considera que se trata de un entorno multiagente dado que la pronunciación de las frutas de una u otra forma puede tener efecto en que el agente entregue la fruta solicitada, otra diferente o ninguna, lo cual afecta el rendimiento del agente. Dicho de otra manera, el estado que percibe el agente está afectado por el comportamiento del usuario considerado en sí como un agente.
- Determinístico:** No existe fuente de elatoriedad en la operación del agente como para que la respuesta no se pueda conocer con total certidumbre. Para una misma

conjunto de imágenes de frutas y para un mismo audio de entrada, la respuesta, sea la deseada o no, siempre es la misma en distintas ejecuciones.

- **Episódico:** La clasificación del audio y de las imágenes se hace en episodios aislados. La clasificación que se haga de una próxima orden o de las imágenes de las frutas en las estanterías no depende de las clasificaciones hechas anteriormente.
- **Estático:** El agente no tiene que hacer un seguimiento del entorno mientras hace la clasificación del audio y de las imágenes dado que el mismo no cambia cuando esta haciendo una determinación; las estanterías no cambiarán hasta que se expenda una fruta y el usuario no podrá dar una orden hasta que la actual esté completa.
- **Discreto:** Como se dijo, el estado viene dado por las frutas en las estanterías y la orden del usuario. Asumiendo que el usuario, entendido como un agente, solamente solicitará frutas válidas por el micrófono, la cantidad de posibles órdenes en un instante determinado son solamente 4 (pera, banana, manzana o naranja). De la misma manera, en 4 estanterías, cada una de las cuales alberga una fruta de 4 tipos diferentes, la cantidad de posibles combinaciones será de 256. En total habrá solamente 1024 posibles, es decir, la cantidad de estados es contable y finita.

4 Diseño del Agente

Para el diseño de ambos sistemas se realizaron variadas pruebas que son muy extensas como para documentar en este informe, por lo que se decidió presentar aquí solamente una descripción del diseño final de los sistemas con algunas descripciones de la evolución y justificaciones de diseño. Sin embargo, está disponible en línea en un repositorio de GitHub [3] toda la investigación realizada junto con los datasets que se utilizaron.

4.1 Reconocimiento de voz

La principal fuente de información que se utilizó fue una lista de videos [5], acompañada de un repositorio en GitHub [4] centrado en la extracción de características para el reconocimiento de voz y música.

4.1.1 Recorte de Audios

Uno de los principales problemas que se tuvo que resolver fue el de recortar los audios para preservar únicamente la parte hablada de los mismos. Inicialmente, esto se llevó a cabo con funciones de librerías cargadas, como `librosa.trim`, para la que hay que definir un umbral por debajo del cual algo es considerado como silencio. Este tipo de solución no pareció ser tan robusta, sobre todo cuando los audios presentaban cierto nivel de ruido tanto al inicio como al final del audio, dado que superaban el nivel considerado como silencio. Luego de la exploración de diversas alternativas propias, se concluyó con una solución suficientemente robusta para el recorte de los audios.

Se pudo observar que el **flujo espectral** era una excelente característica para identificar las partes habladas de un audio de las partes no habladas, siendo poco sensible a los ruidos. El flujo espectral es una característica del audio muy útil en la identificación

de eventos de sonido. Se calcula a partir de un spectrograma de magnitudes (energía) calculando la diferencia entre frames sucesivos, esto se eleva al cuadrado para eliminar el efecto de pequeñas variaciones y se suma a lo largo de todos los intervalos de frecuencia para obtener un valor para cada frame. En la Figura 1 se muestra un ejemplo de cómo el flujo espectral indica el comienzo y finalización de una parte hablada. En la misma, para un audio de ejemplo en el que se menciona la fruta naranja se superponen la señal original y el flujo espectral normalizados en el rango de -1 a 1.

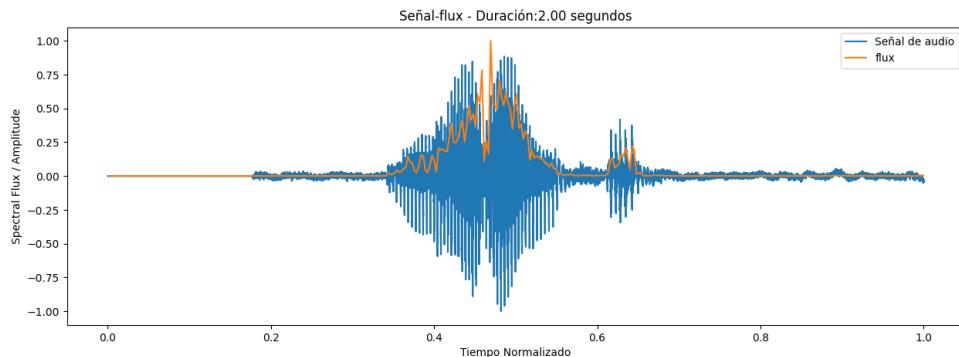


Figure 1: Flujo Espectral sobre la señal de Audio

Para producir el recorte del audio con esta propiedad basta con definir un umbral y proceder. Sin embargo, este corte es sensible a ciertas perturbaciones en el audio que se presentan como picos iniciales y finales en la señal. En ese caso, hay que definir un umbral suficientemente grande de modo de pasar por alto esos picos. Al hacer eso, el audio queda recortado de más, eliminando partes del audio habladas en los extremos, como sucedería en el ejemplo que se muestra en la Figura 2.

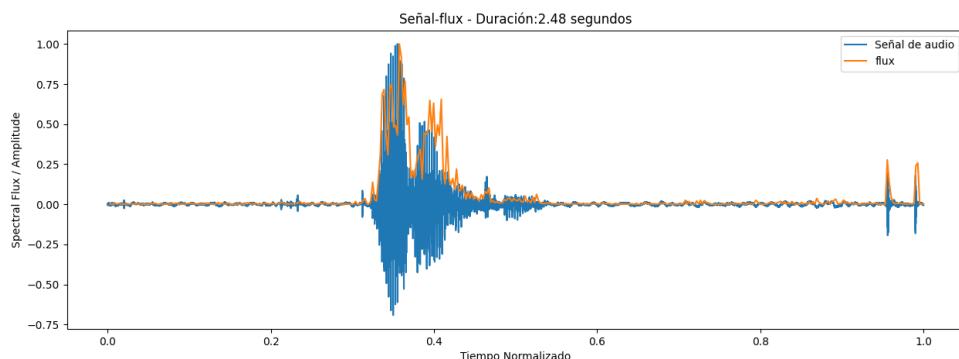


Figure 2: Flujo Espectral - Audios con Picos

Como estrategia para resolver este problema, se propone definir un umbral mínimo y un umbral máximo. El umbral máximo debe ser tal que permita pasar por alto los picos, y luego el umbral mínimo sirve como ajuste fino del corte. De esa manera, se buscaría en la señal de flujo espectral el primer instante a la izquierda y a la derecha del audio en donde se supere el umbral máximo, y desde ese punto y buscando hacia la izquierda en el extremo izquierdo o hacia la derecha en el extremo derecho encontrar el primer instante de tiempo en el que la señal de flujo espectral se encuentre por debajo del umbral mínimo.

El problema que se presenta en este caso es que existen audios en los que el flujo espectral es prácticamente nulo aún en partes habladas, como en el ejemplo de la Figura 3. Esto hace que el umbral mínimo deba ser prácticamente igual a cero.

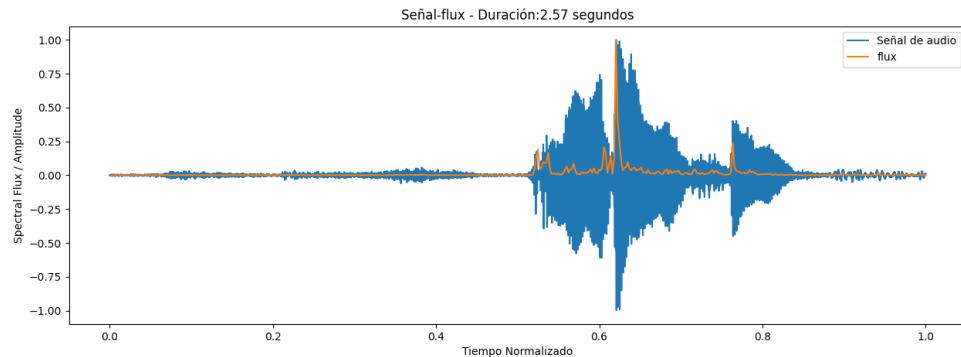


Figure 3: Flujo Espectral - Umbral mínimo

Para resolver este problema, se introduce una segunda característica el valor **RMS** de la señal que sirve como envolvente de la señal original. Ahora, la estrategia es la misma, pero el umbral mínimo se define a partir de una fracción del valor RMS y no a partir de una fracción del valor del flujo espectral. En la Figura 4 se muestra un ejemplo de este corte. En esa figura, la línea horizontal de color rojo indica el umbral de corte grueso por flujo espectral, mientras que la línea horizontal de color azul indica el umbral de corte fino por RMS. Las líneas punteadas verticales indican los puntos de corte, las rojas indican los puntos determinados por el corte grueso, mientras que las azules indican los puntos finales de corte fino por RMS. Como se puede observar, ahora es posible superar los picos finales que se presentan en la señal de audio.

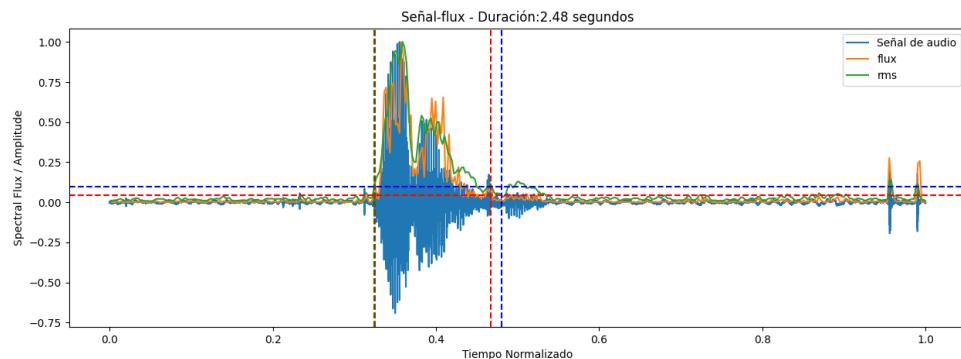


Figure 4: Flujo Espectral - RMS - Ejemplo de Corte

4.1.2 Extracción de Características

En las figuras de esta sección, en el eje X se representa la razón entre el valor RMS de las señales de audio y el valor máximo de la señal, mientras que en el eje Y se representa la característica de que se trate. Los colores de los puntos se corresponden con los colores de las frutas que representan.

La extracción de características comenzó con pruebas con los coeficientes de Mel, **MFCC** (Mel Frequency Cepstral Coefficients por sus siglas en inglés), dado que la investigación arrojó que los mismos son características ampliamente utilizadas en el reconocimiento de voz. Estos tienen la capacidad de describir los fonemas (unidades de sonido de un idioma) y toman en cuenta la percepción del oído humano al utilizar la escala logarítmica de Mel para la representación de características en función de la frecuencia.

En primer lugar, se probó utilizando la media de cada coeficiente de Mel a lo largo de la duración del audio, conservando aquellas componentes que producían la mayor contribución a la separación o que tenían la menor variación dentro de cada grupo. Varias otras pruebas se realizaron con los coeficientes de Mel, por nombrar otra, se probó la utilización de los valores de los mismos a lo largo de todo el audio dispuestos en un solo vector largo, para lo que se tuvo primero que normalizar los audios en amplitud y en duración sin lograr tampoco una separación y agrupamiento satisfactorio.

En el camino, se descubrió una técnica denominada Análisis de Componentes Principales, **PCA** (Principal Component Analysis, por sus siglas en inglés), que permite la reducción de un conjunto de k observaciones en un espacio m -dimensional a un conjunto de k observaciones en un espacio n -dimensional con $n < m$, conservando la mayor cantidad posible de variación a través de los datos, pero de modo tal que las componentes del nuevo espacio son linealmente independientes entre sí.

Al no obtener los resultados esperados haciendo uso solo de los MFCCs, es que se decidió hacer pruebas con otras medidas agregadas del audio, entre ellas **BER** (Band Energy Ratio, por sus siglas en inglés), **ZCR** (Zero Crossing Rate, por sus siglas en inglés), la envolvente del audio, etc.

A continuación, se detallan aquellas características que finalmente se utilizaron.

- **BER:** Esta medida proporciona información sobre cómo está distribuida la energía en distintas partes del espectro de frecuencia. En esta solución se calcula como la fracción de la energía comprendida por debajo de cierta frecuencia de corte.
 - **Máximo:** Se utiliza el máximo del BER para una frecuencia de corte de 600 Hz. En la Figura 5 se muestra cómo se puede lograr una separación de las peras respecto de los demás.
 - **Mínimo:** Se utiliza el mínimo del BER a las frecuencias de corte de 1900, 5000 y 9000 Hz, en las Figuras 6, 7 y 8, respectivamente. Como se puede ver, la primera permite la separación de las peras respecto de los demás, la segunda logra una separación de las bananas respecto de las manzanas y la última una separación de las manzanas respecto de los demás, observándose cierta estratificación de los grupos en el medio.
 - **Desviación estándar:** Para el BER normalizado y considerado respecto de la media. Se tomó con frecuencias de corte a 8000 Hz (Figura 9) y 1000 Hz (Figura 10). Se puede observar cómo en el primer caso se logra una separación de las manzanas respecto de los otros grupos mientras que en el segundo caso se logra una separación de las peras respecto de los otros grupos.
- **Zero Crossing Rate (ZCR):** Esta medida cuenta la cantidad de veces que una señal cruza el eje horizontal (cero) en un intervalo de tiempo dado. El ZCR expresa

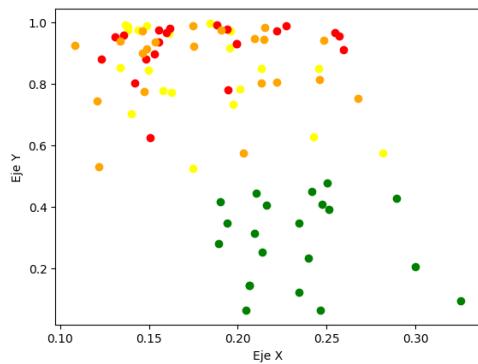


Figure 5: Máximo BER

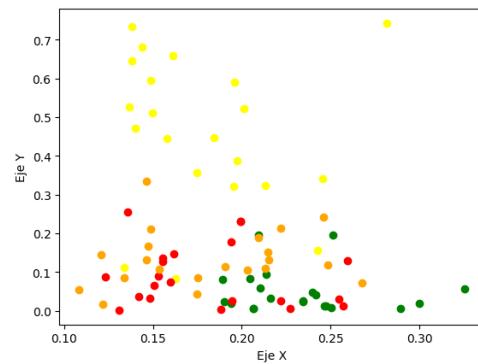


Figure 6: Mínimo BER - 1900 Hz

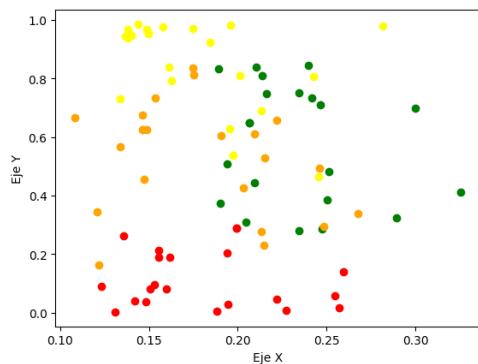


Figure 7: Mínimo BER - 5000 Hz

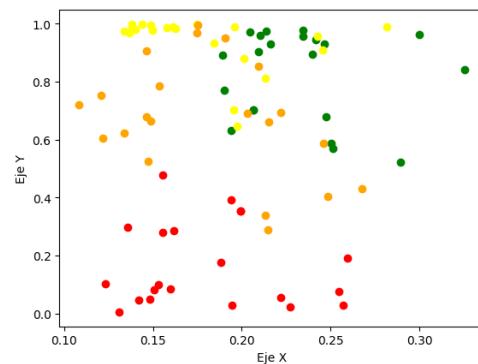


Figure 8: Mínimo BER - 9000 Hz

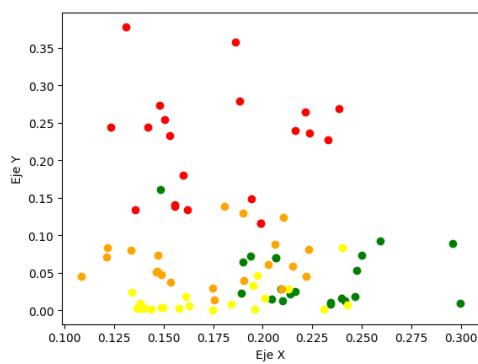


Figure 9: Std BER - 8000 Hz

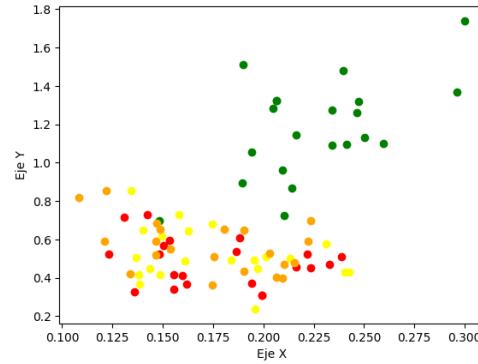


Figure 10: Std BER - 1000 Hz

una tasa, representando la frecuencia con la que la señal cambia de polaridad. Un ZCR alto indica que la señal cambia de polaridad con frecuencia. Por otro lado, un ZCR bajo indica que la señal mantiene la misma polaridad durante un período de tiempo prolongado, lo que podría ser característico de señales más suaves.

- **Media:** Se obtiene respecto del valor máximo luego de un filtro pasa banda con corte en 1000 y 5000 Hz. Como se puede ver en la Figure 11 esto logra la separación de las manzanas respecto de los demás grupos. Eso se debe a una variación de esta propiedad que no presentan el resto de los grupos cuando se pronuncia la letra 'z'.

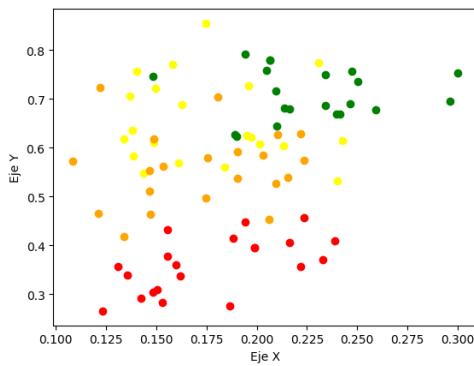


Figure 11: Zero Crossing Rate - Media

- **Máximo:** Se obtiene luego de un filtro pasa banda con cortes en 10 y 1000 Hz. En la Figura 12 se puede observar como nuevamente las manzanas se separan del resto de los grupos quedando las peras y las manzanas en grupos separados.

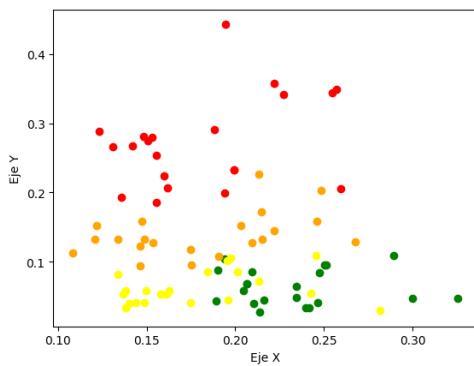


Figure 12: Zero Crossing Rate - Máximo

- **Desviación Estándar:** Se obtiene respecto de la media luego de un filtro pasa banda con cortes en 20 y 10000 Hz. En la figura 13 se puede ver cómo las manzanas se separan del resto y en el centro se pueden observar un grupo casi solo de naranjas.
- **Media a 3/14:** Luego de un filtro pasabanda con cortes en 1000 y 5000 Hz se calcula la media del audio normalizado en ese punto en la duración del audio

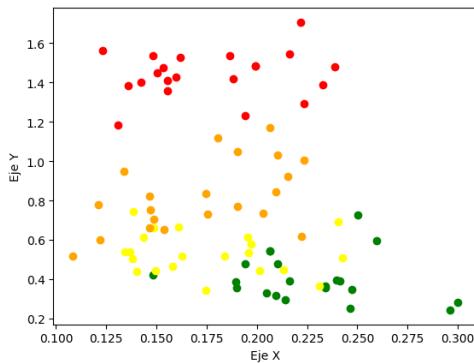


Figure 13: Zero Crossing Rate - Desviación Estándar

a lo largo de 10 frames, 5 a cada lado. Nuevamente se observa una separación de las manzanas (Figure 14)

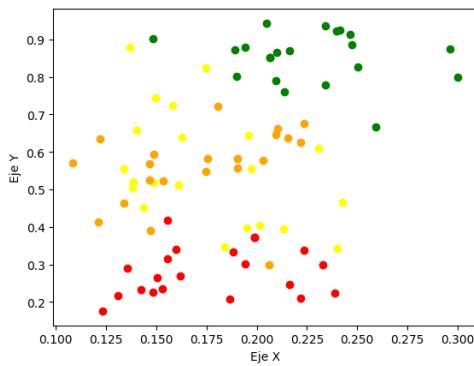


Figure 14: Zero Crossing Rate - Media a 3/14

- **Máximo a 3/4:** Luego de un filtro pasabanda con cortes en 10 y 10000 Hz se calcula el máximo en ese punto en la duración del audio a lo largo de 20 frames, 10 a cada lado buscando resaltar las diferencias entre las naranjas y las demás frutas cuando se pronuncia la letra 'j' (Figura 14).

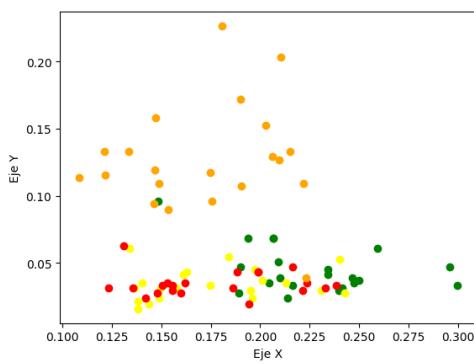


Figure 15: Zero Crossing Rate - Máximo a 3/4

- **Spectral Roll Off:** El *Spectral Roll-off* es una medida que indica la frecuencia por debajo de la cual se encuentra un cierto porcentaje de la energía total del espectro. Un valor bajo indica una concentración en frecuencias bajas, mientras que un valor alto implica una distribución hacia frecuencias más altas.
 - **Media:** Se obtiene respecto del máximo luego de un filtro pasa banda con cortes en 100 y 8500 considerando un porcentaje del 28% de la energía del espectro. Se separan las manzanas y las peras también las naranjas de las bananas y de las peras aunque queda un solapamiento entre bananas y naranjas (Figura 16).

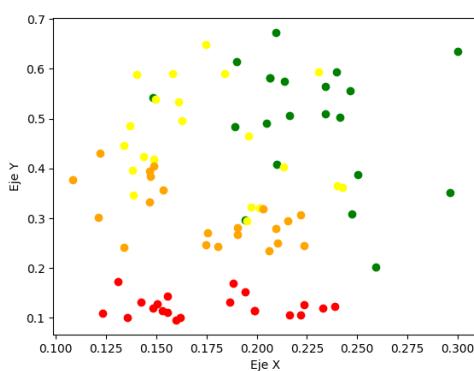


Figure 16: Spectral Roll Off - Media

- **Máximo:** Se obtiene luego de un filtro pasa banda con cortes en 100 y 8500 considerando un porcentaje del 55% de la energía del espectro. Se observa una separación de las manzanas (Figura 17).

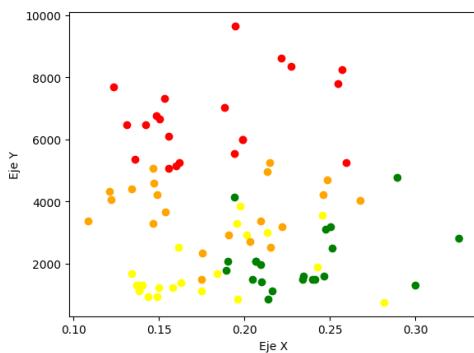


Figure 17: Spectral Roll Off - Máximo

- **Desviación Estándar:** Se obtiene respecto de la media luego de un filtro pasa banda con cortes en 50 y 8500 considerando un porcentaje del 28% de la energía del espectro. Nuevamente se observa una separación de las manzanas (Figura 18).
- **MFCCs:** Como ya se mencionó, es de las principales características utilizadas para el reconocimiento de voz dado que tiene la capacidad de identificar fonemas y demás.

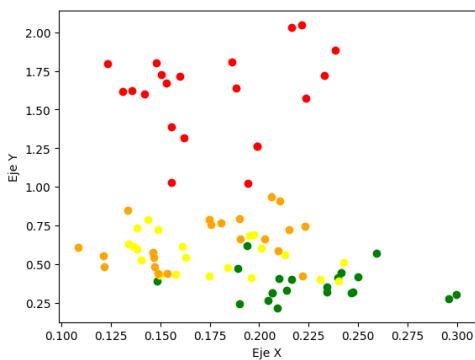


Figure 18: Spectral Roll Off - Desviación Estándar

- **Máximo del coeficiente 3:** Se obtiene luego de un filtro pasa banda con cortes en 500 y 5000. Se separan las manzanas (Figura 19).

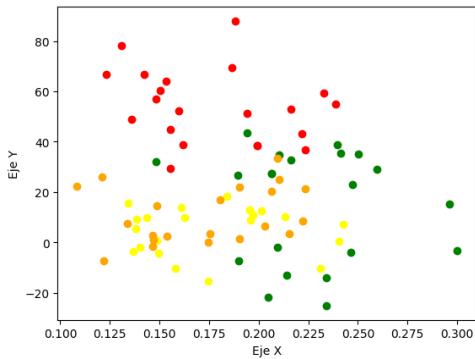


Figure 19: MFCCs - Máximo del coeficiente 3

- **Desviación estándar coeficiente 1 a 4/5:** Se obtiene respecto de la media luego de un filtro pasa banda con cortes en 10 y 8000 en 20 frames alrededor de este punto en la duración del audio (10 frames de cada lado). Aunque existe cierta dispersión, se observa una separación de las naranjas (Figura 20).

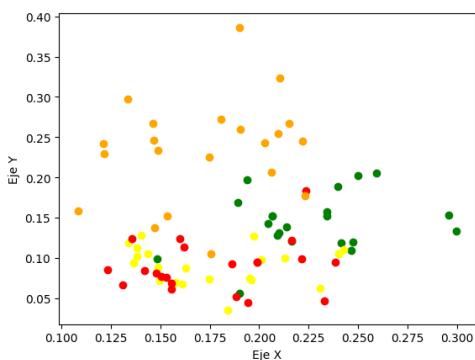


Figure 20: MFCCs - Desviación estándar coeficiente 1 a 4/5

- **Envolvente:** El cálculo del valor RMS en cada frame en que se divide un audio funciona como una buena envolvente de amplitud de la señal. De esta envolvente se toman 30 componentes equi-espaciadas de las cuales se conservan la componente 11 y la 12 (Figura 21). Dentro de todo se logra una separación de las manzanas respecto al resto

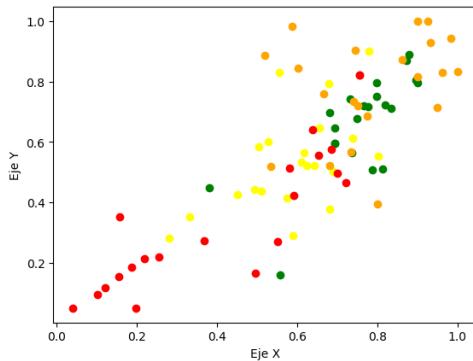


Figure 21: Envolvente componentes 11 y 12

Como se puede notar, algunas de las características extraídas son demasiado específicas. Sin embargo, este conjunto de características fue seleccionado después de muchas pruebas, al considerarse que lograban los mejores resultados en términos de agrupamiento y separación.

4.1.3 Reducción de componentes

La cantidad de componentes del vector de características de cada audio es de 17. Se pretende hacer la visualización en un gráfico de los agrupamientos que se consiguen, y para eso, luego de la extracción de las características de los audios se utiliza PCA para reducir el conjunto de componentes a 3. En la Figuras 22 y 23 se observa el agrupamiento conseguido.

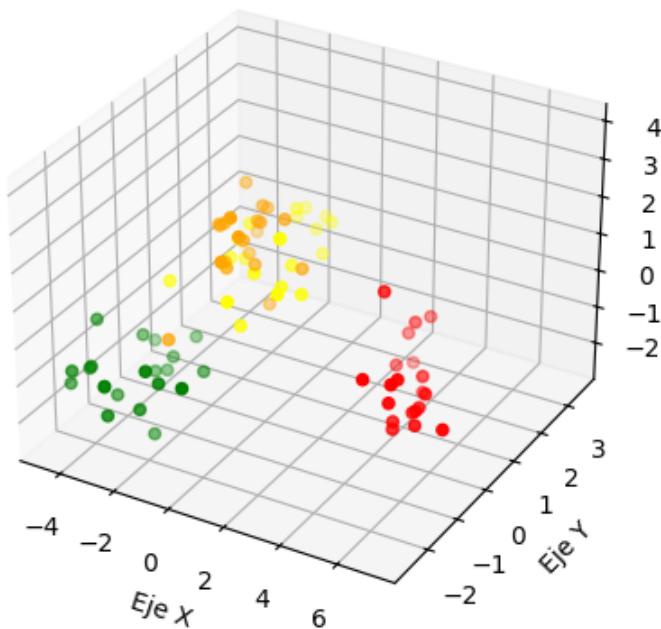


Figure 22: Agrupamiento post PCA - 1

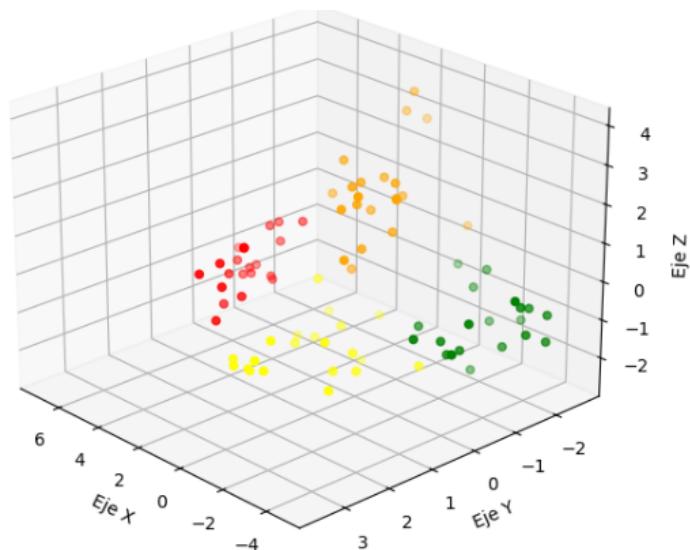


Figure 23: Agrupamiento post PCA - 2

4.2 Reconocimiento de Imágenes

4.2.1 Separación del fondo

Para lograr la separación de la fruta del fondo se exploraron diversas alternativas. A continuación, se describe brevemente el proceso indicando los resultados o los pasos de las partes que quedaron en la solución final.

Principalmente, el problema se abordó tratando de que las imágenes pudieran ser tomadas en diferentes tipos de fondos y no en uno solo normalizado, para que, por ejemplo, sea posible sacar la foto de la fruta sobre una mesa de cualquier tipo. Siguiendo esta línea, lo primero que se exploró fue la detección de contornos utilizando filtros de Sobel (figure 24), sin embargo, esto también detectaba las diferentes texturas que podía llegar a tener el fondo, con lo cual no era una solución viable, ya que no se encontraba forma de separar el contorno de la fruta del resto de los contornos. Entonces, se exploró el uso de máscaras

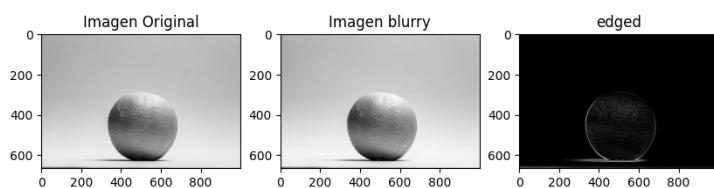


Figure 24: Ejemplo de aplicación de filtros de Sobel

de color que pudieran detectar cualquiera de los colores de las frutas que se consideran (naranja, verde, amarillo y rojo). Estas máscaras se diseñan en el espacio de colores HSV (Hue, Saturation, Value por sus siglas en inglés), que es una escala cómoda para el manejo de colores, dado que distribuye distintos colores en un cilindro en el que el color varía según el ángulo, la iluminación en la altura y la saturación o pureza del color en el radio, permitiendo pasar de forma continua por todo el rango de lo que se puede llegar a considerar rojo, por poner un ejemplo. El inconveniente se presentaba en estos casos cuando el fondo tenía colores parecidos a los mencionados, dado que quedaban incluidos dentro de lo que se consideraba como fruta.

Por este motivo, se comenzó un trabajo de pruebas de distintas máscaras para evaluar el comportamiento de las mismas. Las máscaras se construían considerando canales de distintas representaciones de la imagen. Por ejemplo, se consideraba el canal H del espacio HSV y se aplicaba binarización (pasar a blanco y negro) haciendo uso de OTSU [1]. Esto se practicó con los canales L, A, B del espacio LAB [7], con los canales H, S y V, con los canales R, G B y con la imagen en escala de grises. En la figura 25 se muestra un ejemplo de esto. En la misma, la máscara la máscara señalada como kmeans es una máscara obtenida de aplicar el algoritmo kmeans con dos clusters y en el que los datos a segmentar son los vectores que tienen los valores HSV, LAB, RGB y de escala de grises y hay tantos datos como pixeles de la imagen. Esta máscara como se puede ver funciona bastante bien para la separación de la fruta aunque se puede comprobar que no funciona de forma perfecta en todos los casos. La máscara indicada como "color" es la máscara obtenida por la detección de colores antes mencionada.

Lo que se puede observar que sucede es que los fondos de las máscaras, en algunas de las ocasiones, son negros y en otros son blancos. En lugar de tomar las máscaras en sí, lo que se buscó es, con Sobel [2], los contornos y luego se volvió a binarizar con OTSU para

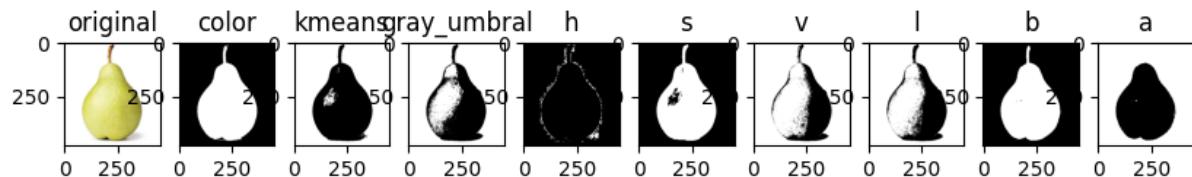


Figure 25: Máscaras probadas

resaltar solamente los contornos, como se muestra en la imagen 26 lo que permite pasar por alto la situación de qué color queda en el fondo. En esa imágen "Whole" representa la combinación de todos los contornos de las máscaras. Esta figura muestra que son varias las máscaras que pueden funcionar adecuadamente para separar a la fruta del fondo. Sin embargo esto es para frutas en las que el fondo de la imagen original es blanco o de otro color pero de tipo uniforme y sin texturas. En fondos mas complejos como por ejemplo, en una mesa con cierta textura, en cambio, en general no hay siquiera una máscara que permita separar correctamente a la fruta del fondo

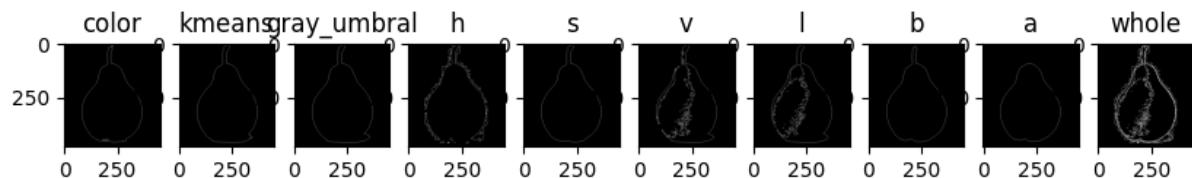


Figure 26: Contornos de las máscaras

Por esos motivos, finalmente se decidió utilizar solamente fondos blancos en las imágenes y se simplificó la separación de la fruta respecto del fondo.

El procesamiento final que se realiza en todas las imágenes para obtener las máscaras consiste en obtener los canales B del RGB, el canal S del espacio HSV y los canales A y B del espacio LAB, ya que se consideró que con estos se conseguían las máscaras más inertes ante sombras o variaciones de iluminación del entorno. Con esos datos por píxel, se procede a realizar la segmentación en dos clusters utilizando Kmeans de la biblioteca ‘sklearn.cluster’. Lo que sucede con Kmeans es que, en ocasiones, el fondo de la máscara obtenida por segmentación puede ser blanco o negro; sin embargo, lo que se pretende es un fondo negro en todos los casos, con la parte de la imagen donde se encuentra la fruta en blanco. Esto se resolvió evaluando los colores de los píxeles (blanco o negro) de la máscara en las esquinas y en las aristas, en matrices de píxeles con tamaños iguales a la 20-ava parte de la menor dimensión de la imagen. Luego, contando la cantidad de píxeles del total que se encuentran en blanco se puede determinar si el fondo de la máscara es blanco o negro. Si la cantidad es mayor al 75%, como quedó fijado, se considera que es fondo blanco y, por lo tanto, la máscara se invierte.

Con la máscara así definida, todavía se presentan algunos defectos en los bordes de la fruta como varias zonas pequeñas aisladas blancas, principalmente debido a los reflejos de la iluminación y la presencia de sombras. Para resolver esto, se aplican operaciones de erosión para limpiar la imagen y luego una operación de dilatación para expandir la máscara en la zona de la fruta y unir todos los bordes irregulares. De esta máscara

resultante, se obtiene el contorno. Este contorno luego se aproxima con líneas poligonales y se dibuja sobre una plantilla de fondo negro con relleno, de modo que todo en el interior del contorno es blanco y el exterior es negro. Esta es la máscara final que se aplicará posteriormente a la imagen para extraer las características y se guarda en un archivo.

En las figuras 27 y 28 se presentan como ejemplo una imagen y su máscara respectivamente.



Figure 27: Imagen Ejemplo



Figure 28: Máscara Ejemplo

4.2.2 Extracción de Características

Las características utilizadas para la separación fueron el color de la fruta y los momentos de Hu 3 y 4.

- **Color:** Para obtener la característica de color, primero se definen rangos de color en el espacio HSV que correspondan a los colores verde, naranja, rojo y amarillo. Luego, se aplica la máscara obtenida durante el procesamiento a la imagen original, obteniendo la fruta en el fondo negro. Se determina en qué rango de color cae cada píxel y se cuenta la cantidad de píxeles que caen en cada rango. El color de la fruta en la imagen se determina como el rango de color más frecuente en la imagen.

Un problema que se tuvo en esta separación es que en ocasiones las manzanas eran clasificadas en el grupo de las naranjas por un límite no muy bien definido entre esos rangos de color. Para resolver ese problema, se aplicó una segunda condición aprovechando el hecho de que las otras frutas presentan en general muy poco nivel de rojo comparado con el color principal en sus cáscaras. De esta manera, para que una manzana no sea clasificada como naranja, además de determinar el primer color frecuente, se encuentra el segundo color frecuente. Si el primer color frecuente es naranja y el segundo color frecuente es el rojo, y además la cantidad de rojo es mayor a cierto porcentaje del color primero (se adoptó un 35%), entonces se acepta que en realidad el color de la fruta es rojo. De esta manera, solo las manzanas que erróneamente son clasificadas como de color naranja se clasifican de color rojo y ninguna naranja es considerada de color rojo.

- **Momentos de Hu:** Con los datos hasta ese momento, el color bastaba para la separación de los grupos. Sin embargo, se pretendió incorporar otras características que pudieran separar a los conjuntos cuando, por ejemplo, las frutas se presentaran descoloridas. Los momentos de Hu también se calculan aplicando primero la máscara a la imagen, se recorta la imagen al rectángulo que enmarca al contorno de la fruta, y luego se obtienen con la función `huMoments` de OpenCV.

4.2.3 Clasificación con K-means

En cuanto al algoritmo K-means implementado, este utiliza la distancia euclídea como métrica. La inicialización de los centroides, aunque aleatoria, es "más inteligente", ya que hace uso de un algoritmo que aumenta las probabilidades de que los centroides estén suficientemente dispersos en el conjunto de datos, evitando grupos vacíos, por ejemplo. Cuando se utiliza esta forma de inicializar los centroides, el algoritmo se denomina K-means++ [8]. Básicamente, entre los datos, se eligen centroides de manera iterativa de modo tal que sea más probable elegir un dato como centroide cuanto más alejado esté de los centroides ya elegidos. Las imágenes 29, 30, 31 y 32 muestran la segmentación de las imágenes luego del entrenamiento.

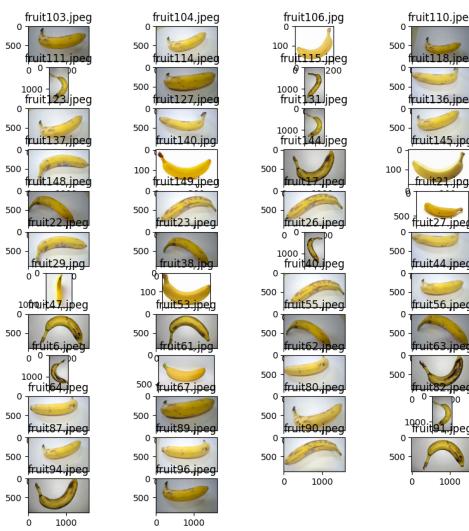


Figure 29: Cluster 0

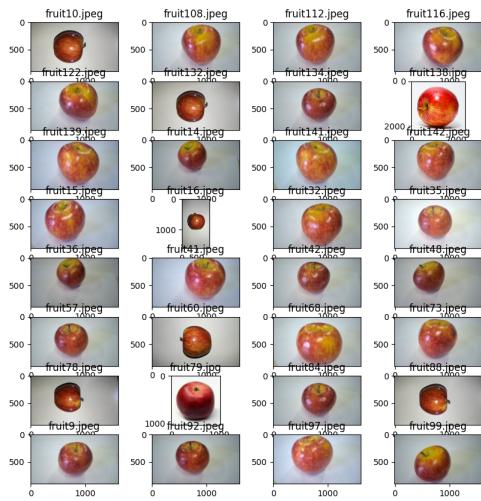


Figure 30: Cluster 1

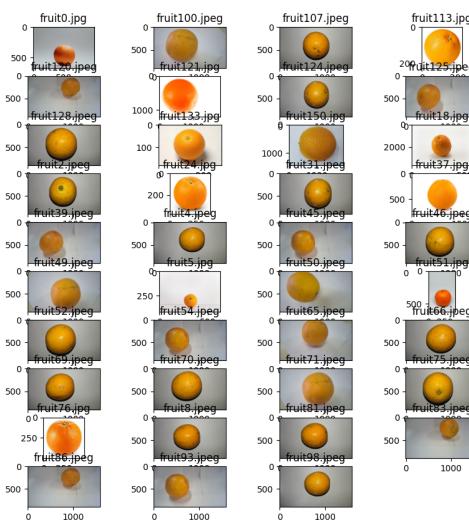


Figure 31: Cluster 2

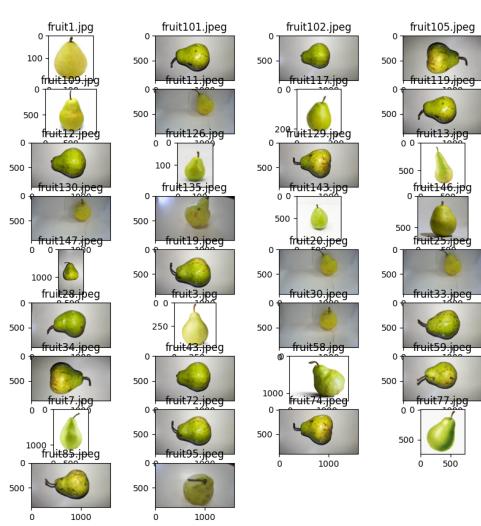


Figure 32: Cluster 3

4.2.4 Clasificación

Luego del entrenamiento, como lo que se pretende hacer es identificar nuevas imágenes en base al modelo entrenado, es necesario poder determinar a cuál de los clústeres pertenece. Esto se logra con el algoritmo de clasificación Knn, donde se utiliza $K = 1$. Para ello, se toman las características o coordenadas de los centroides obtenidos en la base de datos del algoritmo durante el entrenamiento del segmentador Kmeans. Nuevamente, la métrica utilizada es la distancia euclíadiana. La etiquetación de los centroides se realiza de forma manual luego de evaluar la separación lograda con Kmeans. El algoritmo utilizado para Knn es el mismo utilizado para la clasificación de audios.

5 Ejemplo de Aplicación

5.1 Dataset

Por un lado, el dataset se encuentra en './dataset'. Dentro de la misma se encuentran las carpetas "audio" e "imagen" para los archivos de audio e imagen, respectivamente. En cada una de ellas se encuentran las carpetas "training", "validation" y "test", que contienen los datos de entrenamiento, los que se utilizan para la validación de los modelos y test que contiene los archivos de prueba del modelo. En el caso de las imágenes, esta última contiene las carpetas "shelf1", "shelf2", "shelf3" y "shelf4", para contener cada una, una foto de la fruta en el estante correspondiente. A su vez, dentro de las carpetas del último nivel se encuentran dos subcarpetas "original" y "processed", la primera contiene los archivos originales y la segunda alberga los archivos procesados (audios recortados o máscaras de imágenes).

5.2 Implementación

Las implementaciones de las distintas partes del agente se encuentran en la carpeta './implementation'. Dentro de la misma, la implementación del reconocimiento de voz se halla en la carpeta "audio/knn", y la de imagen en la carpeta "imagen/kmeans". En estas carpetas, el archivo "training.py" realiza el entrenamiento de los modelos correspondientes. En el caso del audio, extrae las características de los audios de entrenamiento, aplica la reducción de componentes y guarda los datos en el archivo "model.pkl", que además contendrá el objeto para aplicar las mismas operaciones de transformación de coordenadas sobre nuevos audios. Para la imagen, extrae las características de las imágenes de entrenamiento y entrena al clasificador. Los datos se guardan en el archivo "**training_data.pkl**", que contendrá las características de las imágenes, los centroides obtenidos etiquetados y las imágenes de entrenamiento etiquetadas según el cluster de pertenencia.

Los archivos "validation.ipynb" prueban los modelos contra los datos contenidos en las correspondientes carpetas de validación. En el caso de imagen, un archivo adicional "processing.ipynb" permite obtener las máscaras de las imágenes del grupo de entrenamiento.

El archivo "testing.ipynb" para audio permite probar el clasificador contra audios grabados durante la ejecución del código y contra los audios cargados en la carpeta de testing.

5.2.1 Entrenamiento

Audio: Para entrenar el modelo de audio, se debe ejecutar el archivo de entrenamiento apretando "Run All" en el notebook (figura 33). El modelo se guardará y se mostrará en la última celda de código el agrupamiento obtenido (figura 34).

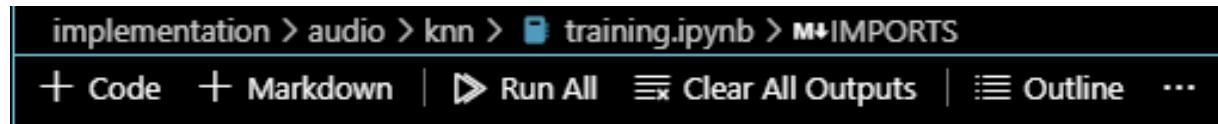


Figure 33: Ejecución

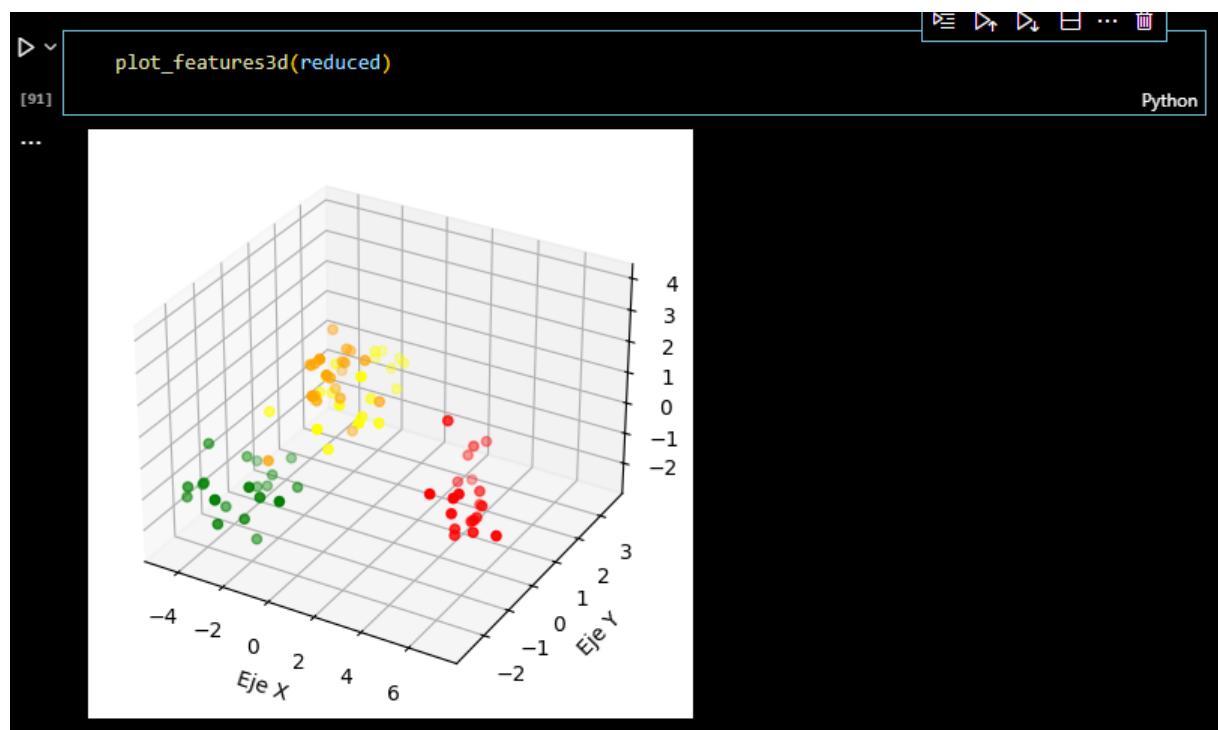


Figure 34: Agrupamiento obtenido

Imagen: Para el entrenamiento del modelo de imagen, se debe ejecutar primero el archivo de procesamiento "processing.ipynb" para obtener las máscaras de las imágenes. Luego, se debe ejecutar todo el código del archivo de entrenamiento. El modelo entrenado se guardará, se mostrarán los archivos en los clusters obtenidos y se mostrará la distribución de puntos en el espacio (figura 35).

5.2.2 Validación

Para la validación de cada modelo, hay que ejecutar el archivo de validación correspondiente. El mismo trabaja con los archivos en la carpeta de validación.

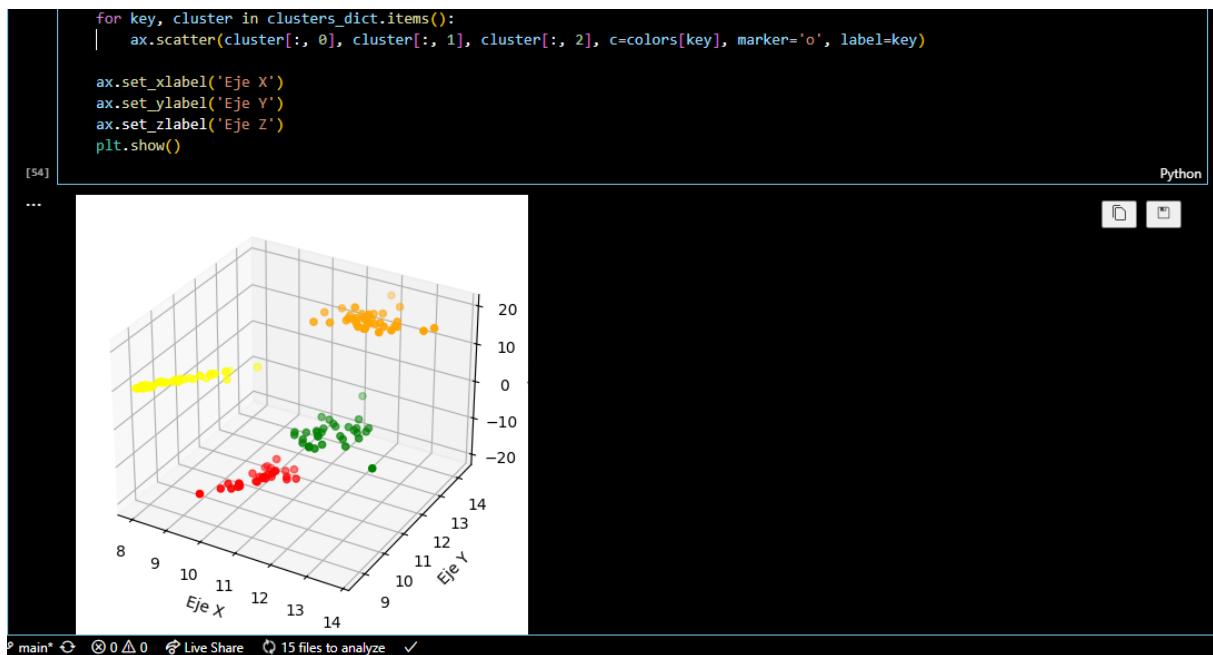


Figure 35: Entrenamiento reconocimiento de Imágen

Reconocimiento de Audio: Se muestra la ubicación de los puntos que representan los archivos de validación junto con los puntos del conjunto de entrenamiento en un gráfico (figura 36). Se indica para cada audio si se acertó la clasificación o no y se muestra el porcentaje de acierto (figura 37).

Reconocimiento de Imagen: Se clasifican las imágenes en la carpeta de validación con el modelo precalculado y se muestran las etiquetas predichas (figura 38).

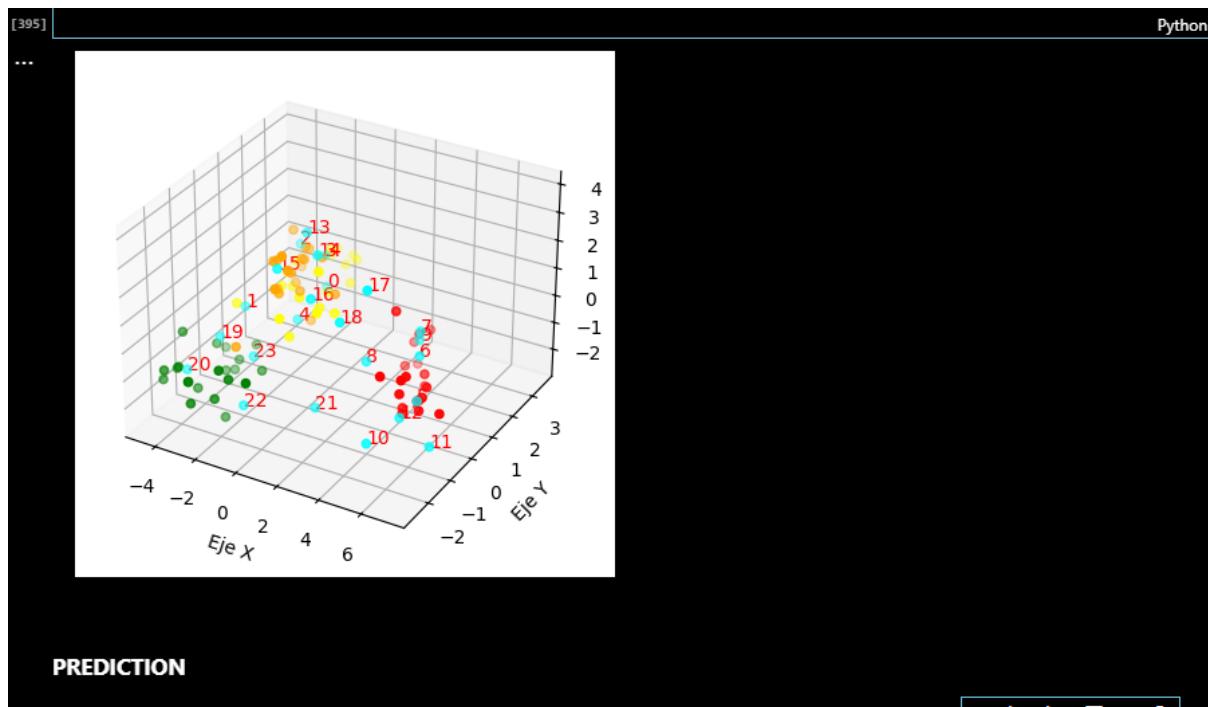
5.2.3 Solución

La solución final se encuentra en `'./implementation/solution/solution.py'`.

5.2.4 Ejemplo

Una vez entrenados los modelos según las indicaciones, sigue los siguientes pasos:

- Toma imágenes de las frutas en los estantes y colócalas en:
`'./dataset/image/test/shelf/original'` según el estante correspondiente, colocando solo una en cada estante.
- Ejecuta la solución:
 - El programa indicará la etapa en que se encuentra (figure 39).
 - Luego, se solicitará al usuario una acción (figura 40).
 - Si el usuario presiona 'ENTER', comenzará la grabación y se indicará que se encuentra en proceso de grabación durante 2.5 segundos (figura 41). Al finalizar



```

+-----+
|   audios | banana1.wav | banana2.wav | banana3.wav | banana4.wav | banana5.wav | manzana1.wav | manzana2.wav |
+-----+
| prediction |   banana   |   banana   |   banana   |   banana   |   banana   | manzana     | manzana     |
| results    |   acierto  |
+-----+
Se acertaron: 24/24
El porcentaje de aciertos es: 100.0

```

Figure 36: Presentación gráfica en la validación

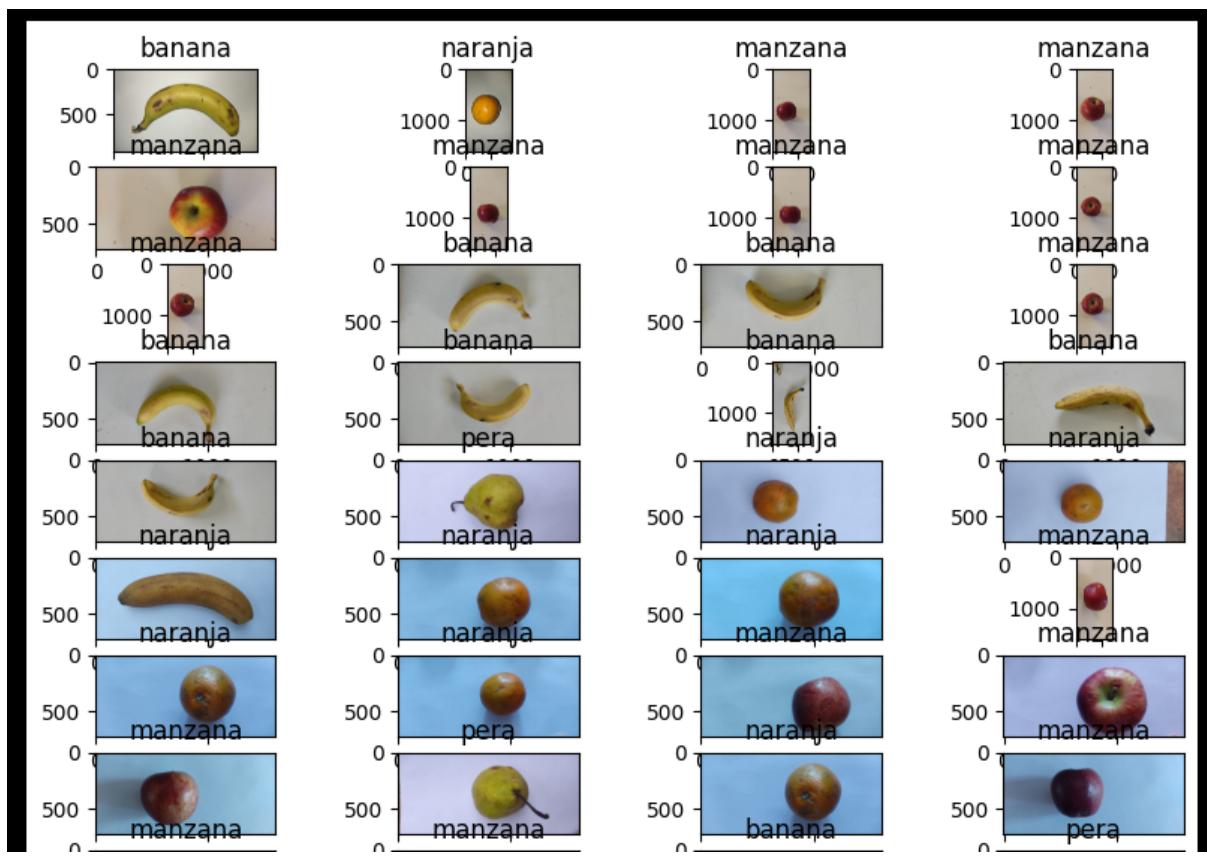


Figure 38: Máscaras probadas



Figure 39: Etapas del programa

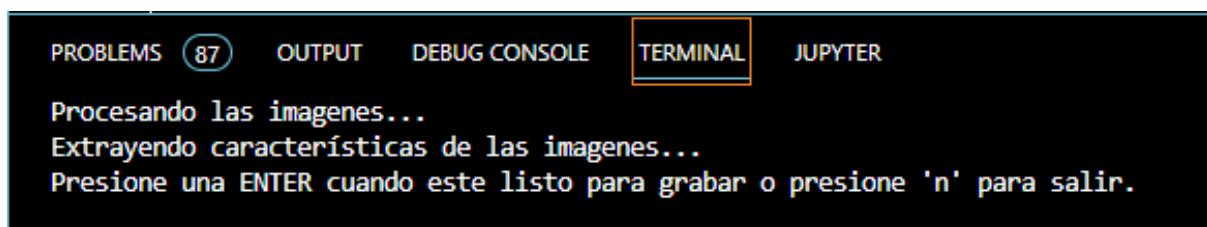


Figure 40: Prompt de usuario

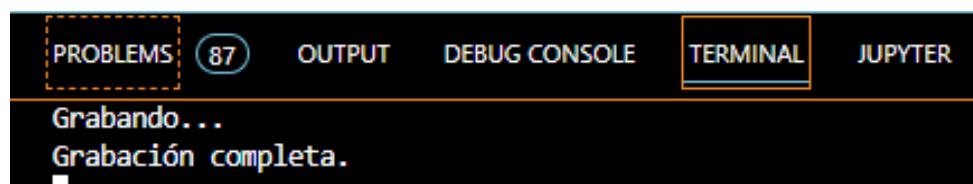
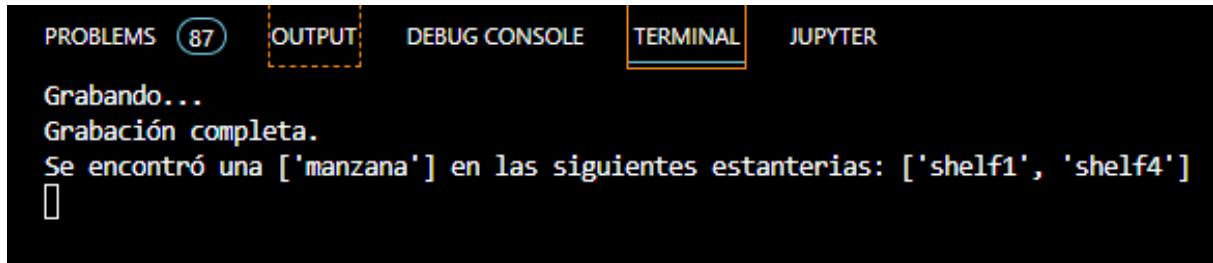


Figure 41: Grabación de audio.

la grabación, el archivo se guardará en la carpeta de archivos de prueba de audio y también el archivo procesado en las subcarpetas correspondientes.

- Al finalizar el procesamiento, se mostrará un mensaje indicando si se encontró la fruta y en qué estante o estantes (figura 42) y se desplegará una figura con la disposición de frutas en los estantes y un recuadro verde alrededor de la fruta identificada, si es que se encuentra (figura 43).



```

PROBLEMS 87 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
Grabando...
Grabación completa.
Se encontró una ['manzana'] en las siguientes estanterías: ['shelf1', 'shelf4']

```

Figure 42: Resultado ejecucion

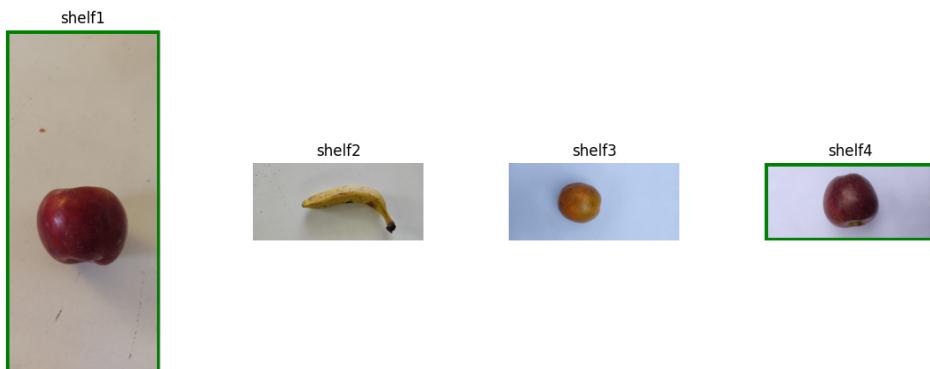


Figure 43: Presentación de los estantes

- Al cerrar la figura, se requerirá nuevamente la entrada del usuario hasta que decida terminar la ejecución.

6 Resultados

La evaluación de los resultados de la solución se hizo a través de la validación del modelo.

6.1 Reconocimiento de voz

Se incorporaron a la base de validación un total de 24 audios de todos los tipos de frutas etiquetados previamente y obtenidos de distintas personas. El modelo de reconocimiento de audio identificó correctamente el 100% de los audios, como queda puesto de manifiesto en el archivo de validación (figura 37). Otras pruebas se realizaron en diversas condiciones

y sin embargo se puede saber que el sistema no es perfecto, existen ciertas formas de pronunciar los nombres de las frutas que harán que no falle la predicción y tiene dificultades en la determinación de ciertas frutas como las naranjas a las que en general confunde con las bananas, cosa esperable por cuanto en el agrupamiento obtenido(figureas 22 Y 23) las bananas y las naranjas no se encuentran tan separadas.

6.2 Reconocimiento de Imagén

A la base de validación se incorporaron 47 imágenes de todos los tipos de frutas, todas en fondo blanco, en distintas posiciones, con iluminación natural en todos los casos. Se procuró utilizar frutas con el color mas vivo que haya sido posible. Los resultados se presentan en la imagen 44. Se puede obtener de esa imagen qué solamente identificó erróneamente a dos de las imágenes, ambas de una banana a las que identificó como naranjas. El porcentaje de acierto sería del 95%. De este experimento se puede determinar el peso que tiene la característica de color en la determinación del tipo de fruta y la poca contribución a la separación que enrealidad se logra con el uso de los momentos de Hu. Se puede notar que esas bananas mal identificadas tenían cierta cantidad de naranja en su cáscara y es muy probablemente esto lo que hizo que falle la identificación.

7 Conclusiones

En el desarrollo de este trabajo se exploraron diversas alternativas en la caracterización de datos de audio e imagen para tareas de clasificación y agrupamiento. Las principales dificultades surgieron al identificar las características más representativas del conjunto de datos, más que en la implementación de los algoritmos de clasificación (Knn) y segmentación (Kmeans). La investigación y las pruebas para encontrar un conjunto de características fueron tareas tediosas y arduas.

El conjunto de características extraídas, al menos para la caracterización del audio, terminó siendo en algunos casos demasiado específico, evaluando propiedades en ciertas partes a lo largo de la duración de un audio. En cuanto a las imágenes, la tarea que demandó más tiempo de prueba y trabajo fue la de separar la fruta del fondo para conservar solo sus características. Se encontró que era difícil lograr una buena separación cuando el fondo no era uniforme, y esta dificultad no se pudo resolver completamente, incluso probando diversas alternativas, incluida la posibilidad de utilizar algoritmos genéticos.

Sin embargo, bajo condiciones normalizadas tanto en el audio (pronunciación adecuada de los nombres y bajo nivel de ruido) como en la imagen (iluminación adecuada y frutas coloridas), se observó un reconocimiento generalmente satisfactorio. Se destaca la vulnerabilidad del sistema a variaciones en estas condiciones.

Por otro lado, la ejecución de los programas de entrenamiento requiere un tiempo considerable debido al procesamiento de datos, especialmente en el reconocimiento de imágenes. En cambio, la ejecución del programa principal es bastante rápida y sin mucha demora.

Tomando en consideración el tiempo invertido en el ajuste del sistema y la vulnerabilidad a variaciones en las entradas respecto del ideal, sería deseable contar con un sistema dotado de capacidades de aprendizaje automático. Este tipo de sistema tendría la capacidad de ajustarse de forma autónoma y asignar los pesos necesarios según convenga a las

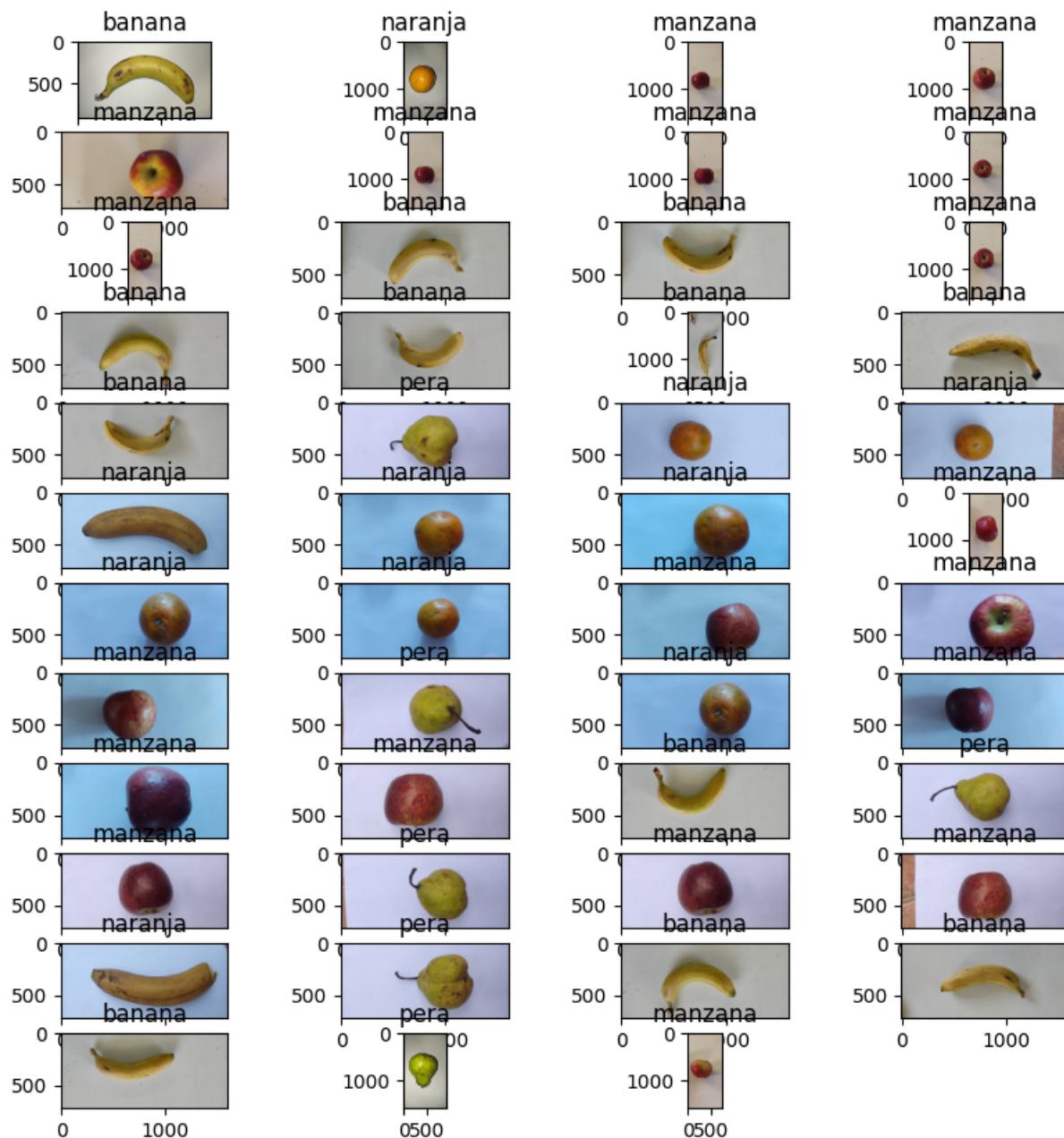


Figure 44: Resultados de la validación del Reconocedor de imágenes

distintas características extraídas de los datos.

Considero que explorar una solución de este tipo podría mejorar significativamente el rendimiento del agente en comparación con intentar mejorar el programa actual. Desarrollar una solución basada en aprendizaje automático podría proporcionar al sistema una mayor capacidad de generalización y adaptación a diversas condiciones, superando las limitaciones actuales, como la restricción a reconocer solo ciertas palabras y la falta de robustez frente a perturbaciones o variaciones, como el caso en el que una persona diga algo diferente a una de las frutas del problema actual y el sistema actual lo clasificaría de todas formas como una fruta.

Referencias

- [1] Digital Sreeni. *Auto segmentation using multi-otsu*. 2020. URL: https://www.youtube.com/watch?v=YdhhiXDQD14&list=PLZs0BAyNTZwZrTsptDOKRdSsV05mU5Qf0&index=20&ab_channel=DigitalSreeni.
- [2] Digital Sreeni. *Edge Filters For Image Processing*. 2020. URL: https://www.youtube.com/watch?v=Oy4duAOGdWQ&list=PLZs0BAyNTZwZrTsptDOKRdSsV05mU5Qf0&index=9&ab_channel=DigitalSreeni.
- [3] Juan2000-coder. *Fruit Recognition GitHub Repository*. 2023. URL: https://github.com/Juan2000-coder/fruit_recognition.
- [4] musikalkemist. *AudioSignalProcessingForML*. 2020. URL: <https://github.com/musikalkemist/AudioSignalProcessingForML.git>.
- [5] Nombre del Canal o Usuario de YouTube. *Valerio Velardo - The Sound of AI*. 2020. URL: <https://youtube.com/playlist?list=PL-wATfeyAMNqIee7ch3q1bh4QJFAaeNv0&si=hv23f1Z0qVSXiwrh>.
- [6] Stuart Russell and Peter Norvig. “V:Learning”. In: *AI A Modern Approach*. Ed. by Michael Hirsch. Pearson, 2009, pp. 736–739.
- [7] Wikipedia. *Espacio de Color Lab*. URL: https://es.wikipedia.org/wiki/Espacio_de_color_Lab.
- [8] Wikipedia. *K-means++*. URL: <https://es.wikipedia.org/wiki/K-means%2B%2B>.