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

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 Introduccion

2.1 Visión Artificial

La visión artificial es la tecnología que le permite a las 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 imagén de entrada es otra imágen de salida modificada, la visión artificial implica la extracción de características relevantes de las imágenes que permitan identificar los elementos de interés del entorno. 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 camara tradicional sensible a la radiación en el rango del espectro visible o imágenes termográficas obtenidas con sensores sensibles a la radiación infraroja 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 elementos en principio bien definidos como códigos de barras, bordes, objetos, colores, etc.

Las aplicaciones de la visión artifical 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 o un elemento para lograr dicho escenario como un gripper, aumentos y un sensor para capturar la imágen, en general una camará 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 para tratar de separar a las frutas del fondo. 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 optandose 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 de las imágenes se obtuvieron de recopilación de imágenes en línea 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 preprocesamiento 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 programa del agente le permite identificar el nombre de la fruta mencionada. Luego, el agente determina si la fruta se encuentra en alguna de las estanterías y, si es así, identifica en cuál de las estanterías se encuentra la fruta. 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 requerimiento del usuario, el agente toma imágenes a través de una cámara de las frutas en los estantes y las clasifica. La actualización de esta información se realiza siempre que la disposición de frutas en los estantes es modificada.

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 ([**?**]). 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 [**?**]). 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, 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 reflexivo simple**.

3.2 Tabla REAS

Rendimiento	Entorno	Actuadores	Sensores
• 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 en que el usuario lleva a cabo una orden y recibe la fruta requerida. • Tratamiento cuidadoso de las frutas.	• 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.	• 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. • Sistema de iluminación para preparar la escena al tomar las imágenes.	• 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

- **Parcialmente observable**: Aunque se cuente con una escena controlada, es decir, la iluminación dentro de la cabina es suficiente, el color del fondo es el adecuado, la cámara funciona correctamente, se hace uso de un micrófono con poco nivel de ruido que funciona adecuadamente y el ambiente no es ruidoso, la probabilidad de que se pueda identificar con exactitud tanto a las frutas en los estantes como la orden del usuario no es del 100%. Las características relevantes del entorno son justamente qué frutas se encuentran en los estantes y qué orden dio el usuario. El no poder acceder con exactitud a las características relevantes del entorno es equivalente a un agente trabajando con un sensor impreciso y es esta una de las condiciones en las que se puede considerar al entorno como parcialmente observable [**?**].
- **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.

- **Determinista**: Dada la percepción que tiene el agente del estado actual del entorno y las acciones que toma en consecuencia, el siguiente estado del sistema estará determinado; será una fruta menos en el estante en el que se identificó que se encontraba la fruta solicitada. Si existe un mecanismo de reposición automático, otra fruta ocupará su lugar, la que en principio no se puede anticipar de qué tipo será. Sin embargo, esto no implica algún efecto en el rendimiento del agente de manera directa y, por lo tanto, no se considera fuente de indeterminismo. Por otro lado, no se puede predecir la orden que el usuario realice en un instante posterior, pero esto no debe ser considerado como fuente de indeterminismo tampoco [**?**].

- **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 en las que en cada alberga una fruta de 4 tipos diferentes, la cantidad de posibles combinaciones será de 256. En total habrá solamente 1024 posibles estados. Luego se considera que se trata de un entorno discreto.

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 [**?**] 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 [**?**], acompañada de un repositorio en GitHub [**?**] 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 espectrograma 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 superponene la señal original y el flujo espectral normalizados en el rango de -1 a 1.

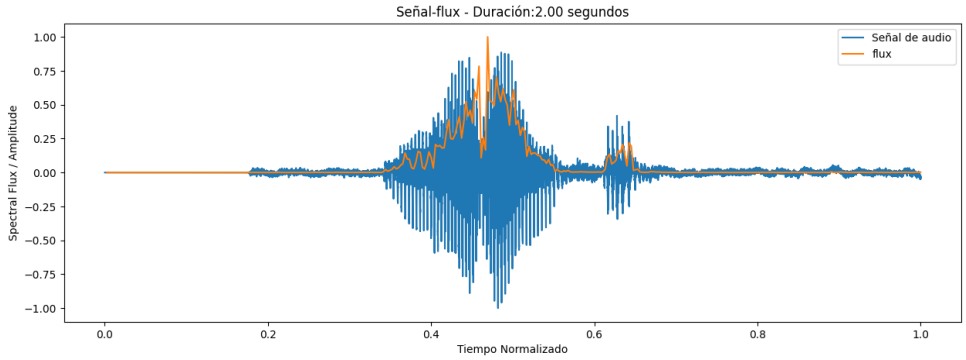

Figure 1: Flujo Espectral sobre la señal de Audio

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.

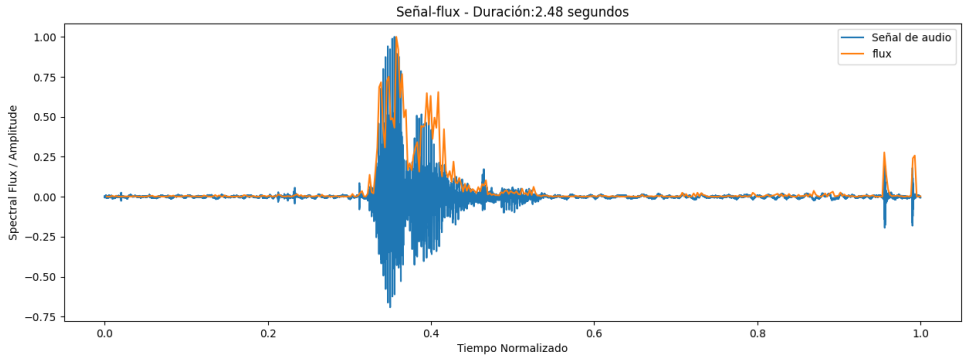

Figure 2: Flujo Espectral - Audios con Picos

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 mínmo 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, 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 máximo. 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 míximo deba ser prácticamente igual a cero.

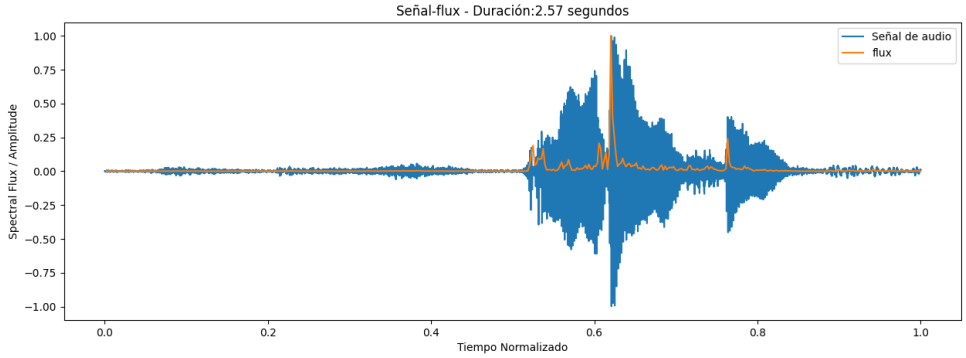

Figure 3: Flujo Espectral - Umbral mínimo

Para resolver este problema, se introduce una segunda característica que sirve como envolvente de la señal original, que es el valor **RMS** de la señal. Este se calcula en frames a lo largo de la duración de la señal. 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.

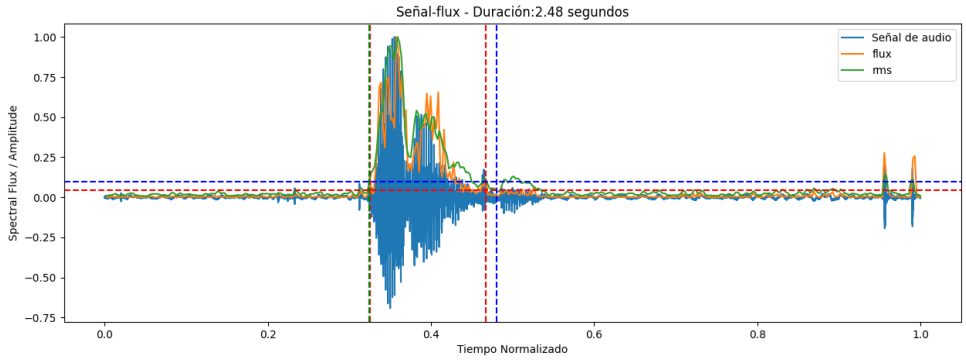

Figure 4: Flujo Espectral - RMS - Ejemplo de Corte

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 del BER del audio. Se tomó con frecuencias de corte a 8000 Hz (Figura

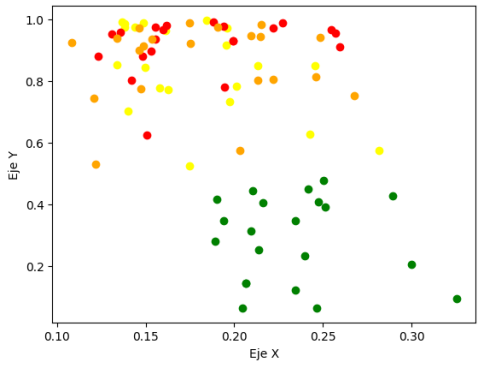

Figure 5: Máximo BER

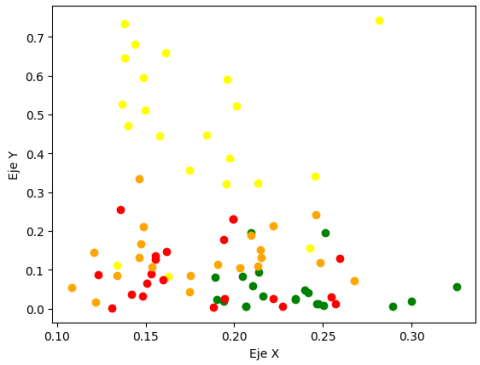

Figure 6: Mínimo BER - 1900 Hz

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 se expresa 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, lo que podría ser característico de señales con contenido de alta frecuencia o ruido. 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'.
  - **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.

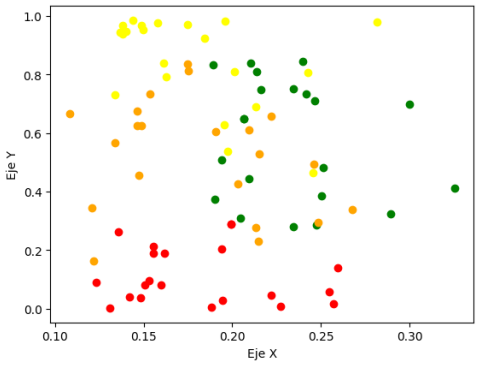

Figure 7: Mínimo BER - 5000 Hz

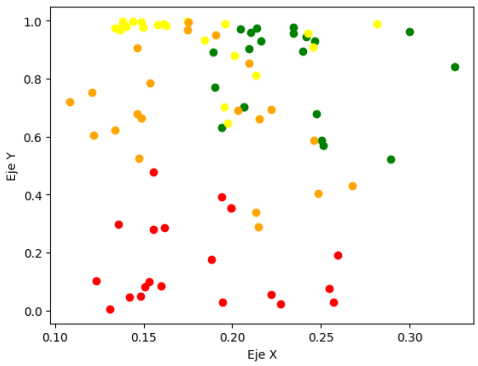

Figure 8: Mínimo BER - 9000 Hz

- **Desviación Estándar:** Se obtiene respecto de la media luego de un filtro pasa banda con cortes en 20 y 10000 Hz.
- **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 a lo largo de 10 frames, 5 a cada lado. Nuevamente se observa una separación de las manzanas (Figure **??**)
- **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'. Se observa una separación de las naranjas (Figure **??**).

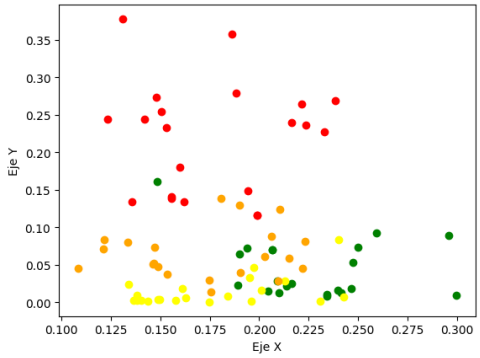

Figure 9: Desviación Estándar BER - 8000 Hz

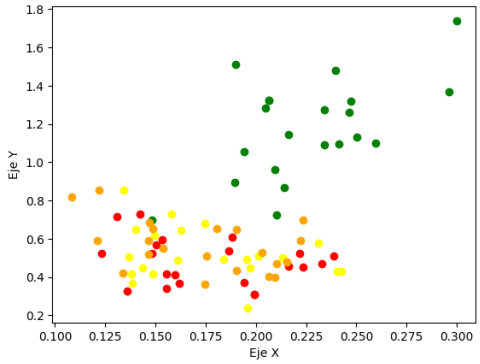

Figure 10: Desviación Estándar BER - 1000 Hz

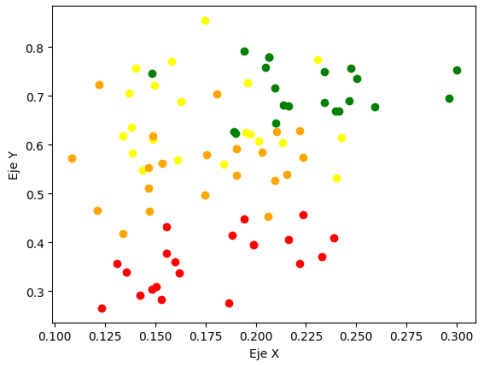

Figure 11: Zero Crossing Rate - Media

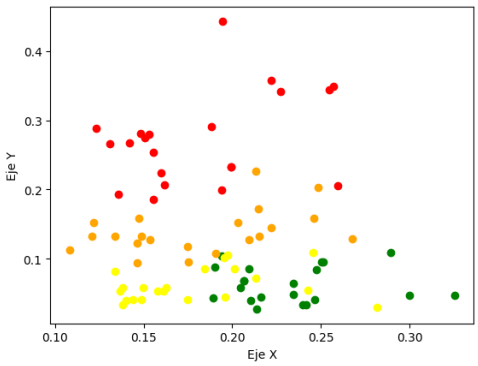

Figure 12: Zero Crossing Rate - Máximo

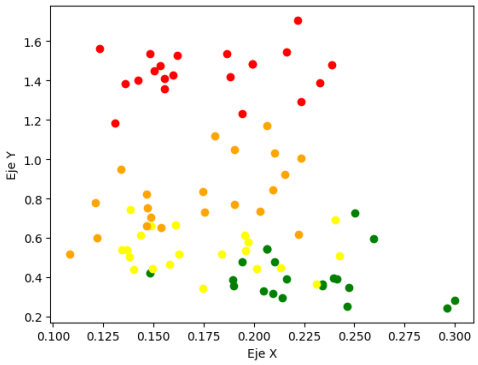

Figure 13: Zero Crossing Rate - Desviación Estándar

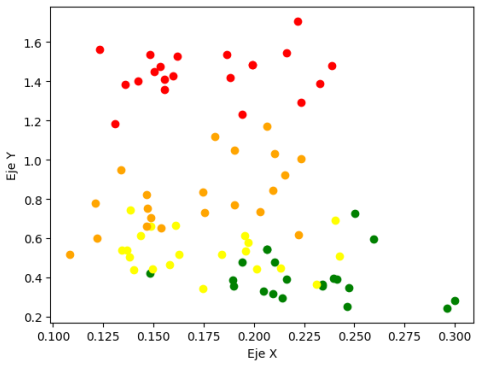

Figure 14: Zero Crossing Rate - Desviación Estándar

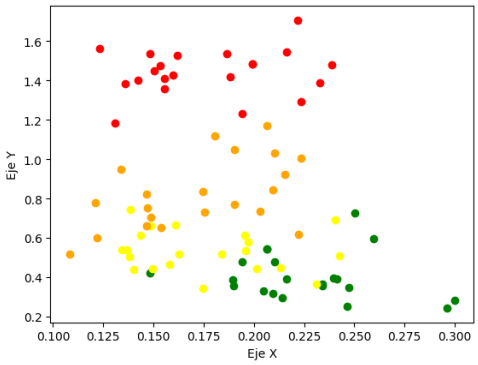

Figure 15: Zero Crossing Rate - Desviación Estándar